EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

PHD THESIS

In Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy from Sorbonne University
Specialization: Data Science

Towards Automatic Understanding of Narrative Audiovisual Content

Alison REBOUD

Defended on 14/12/2022 before a committee composed of:

Reviewer	**Farah BENAMARA**, Université Paul Sabatier, Toulouse, France
Reviewer	**Vasileios MEZARIS**, Centre for Research and Technology Hellas, Thessaloniki, Greece
Examiner	**Claire-Hélène DEMARTY**, InterDigital Inc., Rennes, France
Examiner	**Jean Luc DUGELAY**, EURECOM, Sophia Antipolis, France
Thesis Director	**Ulrich FINGER**, EURECOM, Sophia Antipolis, France
Thesis Co-Director	**Raphäel TRONCY**, EURECOM, Sophia Antipolis, France

Dedicated to my family

Acknowledgements

Abstract

From movies and series produced by the entertainment industry and uploaded on streaming platforms, to social media where users display the story(ies) of their life trough videos, modern storytelling is digital and video-based. Understanding the stories contained in videos remains a challenge for automatic system. Having multimodality as a transversal theme, this research thesis breaks down the "understanding" task into different challenges which cover different aspects of the concept.

1. **Predicting memorability of multimedia.** With a growing sea of videos, interest in both being able to identify and to create memorable content is rising. Memorability is especially interesting because contrary to other associated concepts such as 'interestingness', it can be objectively measured through recognition tests. We explore the task of automatic video memorability prediction in the first chapter, using multimodal models with visual, textual and audio cues for different types of videos with a particular focus on user-created content.

2. **Summarizing multimedia.** After extracting memorable moments, we investigate how to extract moments which are important for the story in TV series. Because of a high annotation cost for this task, we decided to capitalise on the richness of the textual component generally accompanying this type of content to develop unsupervised approaches.

3. **Story modeling in multimedia.** Finally, the last chapter takes a step fruther towards story understanding. It does so by (i) Proposing PROZE, a new explainable approach for zero-shot text categorization which proves promising for the task of narrative aspects classification. (ii) Uncovering how Language Models can be used to generate important questions about TV series plots, that an automatically build summary should aim to answer.

Abrégé

Qu'il s'agisse de films et séries produits par l'industrie du divertissement et distribués sur des plateformes de streaming, ou de médias sociaux où les utilisateurs affichent les histoires de leur vie avec la fonctionnalité 'story', la narration moderne est numérique et basée sur la vidéo. Comprendre les histoires contenues dans les vidéos reste un défi pour les systèmes automatiques. Avec la multimodalité comme thème transversal, cette thèse décompose la tâche de "compréhension" en différents défis qui couvrent différents aspects du concept.

1. **Prédire le degré de mémorabilité d'un contenu multimédia.** Face à la multiplication des vidéos, la capacité à identifier et à créer un contenu mémorable suscite un intérêt croissant. La mémorabilité est une idée particulièrement intéressante car, contrairement à d'autres concepts associés tels que l'"intérêt", elle peut être mesurée objectivement par des tests de reconnaissance. Nous explorons la tâche de prédiction automatique de la mémorabilité des vidéos dans le premier chapitre, en utilisant des modèles multimodaux avec des indices visuels, textuels et audio pour différents types de vidéos.

2. **Résumer du contenu multimédia.** Après avoir extrait les moments mémorables, nous étudions comment extraire les moments qui sont importants pour l'histoire de séries télévisées. En raison du coût élevé de l'annotation pour cette tâche, nous avons décidé de capitaliser sur la richesse de la composante textuelle qui accompagne généralement ce type de contenu pour développer des approches non supervisées.

3. **Modélisation de la narration dans des contenus multimédia** Enfin, le dernier chapitre fait un pas de plus vers la compréhension narrative. Pour ce faire, il (i) propose PROZE, une nouvelle approche explicable, pour la catégorisation de textes, qui s'avère prometteuse pour la tâche de classification des aspects narratifs. (ii) découvre comment les modèles de langage peuvent être utilisés pour générer des questions importantes sur les intrigues des séries télévisées, auxquelles un résumé construit automatiquement devrait pouvoir répondre.

Contents

List of Figures

List of Tables

List of Abbreviations

CNN Convolutional Neural Networks.

NLP Natural Language Processing.

RNN Recurrent Neural Networks.

Chapter 1

Introduction

1.1 Context

According to the Narrative Paradigm by Fisher [82], telling stories is a natural human trait. From the adventures of Ulysses to online blogs, it is an activity with a long tradition, which takes the form of its epoch. If the Covid crisis has reminded us of our attachment to cinemas, cafes, bars and other places where adventures and anecdotes are traditionally shared, it has also accelerated the explosion in digital multimedia content consumption. This trend holds for narrative media produced by the entertainment industry, but also for user created content on social platforms, where one is offered the possibility to turn one's life into stories (*story* literally being the name of a feature on Instagram, Snapchat and Facebook). In the movie sector, Disney passed the 100 million global subscribers less than two years after the launch of its platform [1]. Despite a slowdown in subscribers as the pandemic boom wears off, Netflix did not lose the 36 million new sign-ups it gained from lockdown, passing 200 million customers worldwide [2]. This company, together with its competitors (HBO, Amazon Prime ...) were able to surf on the craze for TV series, a long despised format [36], which has now completely gained recognition, as the creation of the Cannes International Series Festival in 2018 [3] demonstrates. The weight of Netflix on the entertainment business, is such that is now also a major actor in movie production, releasing critically acclaimed films directly on the internet [4]. With regards to user-created videos, it is estimated that about one in six people in the United States uses TikTok weekly [5], a platform to share short videos, launched 5 years ago. To cope with the success of these video-based social media, Meta created the reels functionality, their own version of TikTok videos. On their side, Youtube, the Alphabet-owned video giant, with an

[1] The Guardian - Disney forecast to steal Netflix's crown as world's biggest streaming firm
[2] Netflix records dramatic slowdown in subscribers as pandemic boom wears off
[3] Website of the Cannes International Series Festival
[4] Netflix snags 7 awards, nearly doubling its all-time Oscars tally
[5] Massive TikTok Growth: Up 75% This Year, Now 33X More Users Than Nearest Direct Competitor

estimated two billion monthly users [6]developed Shorts, its TikTok-like short video platform [7]. Finally, as showcased by the 'Storytelling Goes Here' campaign, Meta now ambitions to further develop their video sharing capabilities, by facilitating the uploading of long-form videos via new clipping options. [8]

All these figures, confirm that people watch and produce stories, using online tools and that short highlights, are particularly popular when it comes to videos on social media. In brief, modern storytelling is digital and video based. Such an evolution is particularly interesting for AI researchers: while offering meaningful real world use cases - the sector is in need for various tools to help navigate through an ever growing sea of videos-, the richness and complexity of multimedia content poses some challenging research questions.

The spectrum of tasks related to multimedia understanding is wide and encompasses different levels of complexities. If location, faces or image classification were approached by automated systems quite successfully [136, 241], higher level tasks such as story plot understanding or video summarization, which require models to get closer to human understanding, remain challenging. For example, to create TV series summaries, an automatic system would need to capture scenes that are important for the narrative. One can imagine how a traditional model generating a trailer based on low-level features, will fail to capture the semantics of the video and how there will be a gap between the models and the way humans process content. Which scenes do humans consider to be essential parts in narrated media? Which types of videos do people remember? Can it be predicted automatically? These are a few of the high level questions related to multimedia understanding, that we are interested in answering in this thesis. Following this path will involve dealing with the specific challenges posed by multimedia content, namely its multimodal nature and its diversity.

Besides the visual stream, videos often contain sound, speech and are sometimes accompanied by metadata such as transcripts, titles or descriptions. Ideally, an *ai* model would be able to combine information from these different representations, in the most effective way. The topic of multimodality as well as the relation between what is said and what is done (what is going on visually) in a video, will be a transversal theme in this thesis.

For some tasks, such as audiovisual summarization (that should here be considered as the task of binary classifying scenes as interesting or not), a high diversity in video domains can be a challenge. *Interestingness* is indeed quite a fuzzy concept which is often domain-dependant. For example, interesting moments for the narratives of a TV series that spans over episodes, will differ from interesting moments of a football match. Similarly, there is an heterogeneity in the amount and nature of metadata linked to a video: if a movie or TV series is often

[6]TikTok overtakes YouTube for average watch time in US and UK
[7]YouTube Outlines Key Areas of Growth, Including the Rise of Shorts and its Expanding Creator Economy
[8]Meta Launches Facebook Reels to All Users, Expanding its Short-Form Video Push

accompanied by a profusion of additional text such as fandom synopsis, reviews or wiki articles, it is not necessarily the case for user created videos. In this thesis, we decided to work with TV series and movies as well as with user-generated short videos.

1.2 The MeMAD Project

This work has been done in the context of the EU funded H2020 research project MeMAD. The acronym stands for Methods for Managing Audiovisual Data and its aim has been to "develop methods for an efficient re-use and re-purpose of multilingual audiovisual content targeting to revolutionize video management and digital storytelling in broadcasting and media production" [9].

The creation of MeMAD has been motivated by the increase of audiovisual big data and the resulting needs to deal with it and use it more efficiently by the entertainment industries including television, cinema and streaming services. More concretely, automatic language-based methods for managing, accessing and publishing video content to facilitate its re-use were some of the main research directions and goals of MeMAD.

In addition to project goals, MeMAD formulated four use cases:

- Content delivery services for the re-use by end-users/clients through media indexing and video description
- Creation, use, re-use and re-purposing of new footage and archived content in digital media production through media indexing and video description
- Improving user experience with media enrichment by linking to external resources
- Automated subtitling/captioning and audio description for general purpose use and for the deaf, hard-of-hearing, blind, and partially-sighted audiences

The project partners of MeMad were four research institutes, Aalto University and University of Helsinki from Finland, University of Surrey from the United Kingdom and EURECOM from France, four companies, YLE from Finland, Limecraft from Belgium and Lingsoft plus Lingsoft Language Services from Finland, as well as the French Institut National de l'Audiovisuel or National Institute of the Audiovisual. Our contribution to the project falls into (i) the "Automatic Multimodal Content Analysis" package which developed tools for for multimodal analysis, description and indexing of video content (ii) the "Media Enrichment and Hyperlinking" package which is centered around the use of natural language processing and semantic technologies to predict which TV moments will lead to viewers' interest and how such moments should be enriched.

[9] https://memad.eu/

1.3 Research Questions

How to identify memorable moments in media content?

20th of April 2022, just before the TV debate [10] between Emmanuel Macron and Marine Le Pen (the two second-round candidates for the 2022 French presidential elections), the political journalist Maxence Lambrecq explains at the radio [11], that in preparation for the debate, much of Emmanuel Macron' political advisors attention was spent on 'managing his smile', finding a facial expression which would evoke friendliness rather than arrogance. He justifies this attention to physicality and gestures by stating that images are more easily remembered than speech. We here see that identifying the type of cues that people remember is a of foremost importance for whoever wants to tell and control a story. This traditionally includes actors from different fields such as advertisement, education, politics... Now that there has been a democratisation of video content creation, we can easily imagine, how social media users would also benefit from being able to automatically predict how memorable their video will be, before they post it online. Similarly, for these online platforms, being able to display the most memorable videos, would improve their user experience. Memorability is here defined as the quality or state of being easy to remember. It is a notion which is especially interesting to data science, because contrary to other associated concepts such as 'interestingness', it can be objectively measured through recognition tests. Following the success of image memorability prediction, the task of video memorability prediction was then formalised only a few months before the beginning of this thesis, in 2018, with the first edition of the *MediaEval Memorability Challenge*. In the chapter 3 of the thesis, we will explore what makes a video memorable, which modalities (textual, audio, visual) are relevant and will assess the generalisation capabilities of our approaches to other datasets.

How to summarize stories in media content?

After exploring memorability, in Chapter 4 we cover another dimension of interestingness: we aim at extracting the parts of a video which are *essential to the story*. In this context, interesting moments are the ones that are decisive for a narrative and summarization becomes the task of automatically selecting scenes which are important elements of the narrative structure of the video. After working with user-created content in the previous chapter, we here use videos created by the entertainment industry, focusing on summarizing stories from TV series episodes. As Bost [36] pointed out, modern TV series offer a realistic use-case for narrative summarization, because contrary to classical TV series composed of self-contained episodes,

[10] The debate is available at https://www.france.tv/actualites-et-societe/politique/3264511-le-debat-de-l-entre-deux-tours.html

[11] "Edition spéciale : Débat de l'entre-deux tours" on https://www.franceinter.fr/emissions/le-telephone-sonne/le-telephone-sonne-du-mercredi-20-avril-2022

their plots spans over numerous episodes. Series are generally divided into a set of episodes called seasons, which are releases annually or semi-annually. As a consequence, when a new season comes out, the viewers are often disconnected from the plot. Bost [36] found that 60% of the people they polled felt the need to be reminded of the major narrative elements of the previous seasons before watching the new one. This use-case is therefore an example of the type of issues that the development of automatic tools for TV series plots summarization, can solve. In this thesis, we use both self and not self-contained episodes. As a sub-topic, we specifically interrogate the use of visio-linguistic models and the potential of unsupervised approaches for this task.

How to automatically extract story elements in media content ?

After trying to isolate the most important moments of the narrative of TV series episodes, in Chapter 5 we further explore the general understanding of stories in TV series from automated systems. Because media content summarization, rather than a stand-alone task, is related to a wide range of other tasks such as the extraction of 'content-related' features, developing story-related video analytics tools directly complements the goal of the previous chapter. In this Chapter, using screenplays, we specifically ask how the tasks of question-answer generation and text classification allow the extraction of specific story elements. One aspect of story understanding is indeed being able to ask and answer high-level meaningful questions about the plot. Such questions could include topics such as the relationship between characters or the motive of an action *ie: why was someone killed?*. We explore if and how Language Models are able to generate such questions. We sometimes want to extract texts related to a topic that we chose in advance. This is what the task of *text classification* does. For instance, we have a crime series episode and we want to tag scenes which are about the crime scene. In this direction, Li et al. speak in favour of the development of story-based classification of movie scenes [165], arguing that the research carried by story theorists in identifying frequently re-occurring themes or sequences of events, common to most well-written stories [215, 273] is a good starting point for such as task. In this context, we interrogate how to extract fine-grained elements of stories (such as the victim, the cause of death or the perpetrator for the particular case of crime series). In this section, we consider that this can be done through domain-adaptation of zero-shot classification models. In particular, we interrogate the possibility of a system that would leverage both on the power of Language models and on the explainability of common sense database.

1.4 Contributions

The work conducted during this thesis has led to the following contributions:

- Contributing to advancing that state of the art in Media Memorability Prediction by participating to the *MediaEval Memorability Challenge* [57, 60, 85, 131] in 2019, 2020, and 2021. During this thesis, we delved into different facets of memorability prediction including multimodality (*how to best combine features from different modalities*), choices of visual, textual and audio features as well as the impact of perplexity as a proxy for novelty. We showed that short-term memorability can be best predicted - we achieved a 0.658 Spearman score on the Memento10K dataset- with multimodal models and that memory decay remains a challenging task. In this thesis, we also dedicated some attention to investigate the robustness of our approaches by testing them on a total of 5 datasets spanning a large variety of genres, including movies or vines. In particular, besides the 3 benchmark datasets, we used two different MeMAD datasets containing TV programs from two content providers: Yle (*Yleisradio Oy*, Finland's national public broadcasting company) and INA (*Institut National de l'Audiovisuel*, a repository of all French radio and television audiovisual archives). The code is published at https://github.com/MeMAD-project/media-memorability
- Developing PROZE, a model for Explainable and Prompt-guided Zero-Shot Text Classification that leverages knowledge from two sources: prompting pre-trained language models, as well as querying ConceptNet, a common-sense knowledge base which can be used to add a layer of explainability to the results. We evaluate our approach empirically and we show how this combination not only performs on par with state-of-the-art zero shot classification on several domains, but also offers explainable predictions that can be visualized. A demonstrator is available at http://proze.tools.eurecom.fr/
- Proposing two unsupervised approaches for TV series summarization. The first one is a fan-driven and character-centered approach which ranked first at the 2020 TRECVID [16] Video Summarization Task. After selecting the shots of interest via a face recognition step, a similarity score is computed between sentences from fan-made content (BBC EastEnders episode synopses from its Fandom Wiki[12]) and transcripts. The second approach leverages on the creation of large language models which enabled Zero-Shot text classification to perform effectively in some conditions. We explore if and how such models can be used for TV series summarization by conducting experiments with varying text inputs. Our main hypothesis being that interesting moments in narratives are related to the presence of interesting events, we choose candidate labels to be events representative of two genres: crime and soap opera and obtain competitive results. The code is published at https://github.com/alisonreboud/screenplay_summarization and https://github.com/MeMAD-project/trecvid-vsum.
- Studying the use of visio-linguistic models and pretraining choices for supervised TV series summarization. Visio-linguistic models have proven to be successful for several downstream tasks using paired text and images. Being presented as task-agnostic,

[12] https://eastenders.fandom.com/wiki/EastEnders_Wiki

we explore if and how they can be used for TV series summarization by conducting experiments with varying text inputs (dialogue and scenic textual from screenplays) and models fine-tuned on different datasets. We observe that such generic models, despite not being specifically designed for narrative understanding, achieve results closed to the state of the art. Our results suggest also that non aligned data also benefit from this type of visio-linguistics architecture. We provide our implementation at https://github.com/alisonreboud/mmf

1.5 Thesis outline

The remainder of this thesis is organized in four chapters. We can recapitulate the contributions on this thesis as seen from the three lenses of multimedia understanding as stated above:

1. In chapter 2, we start by presenting the state of the art on multimedia understanding. We start by giving an overview of multimodal and NLP side, during the period of writing this thesis. It is a period that is defined by two things: the advent of big pretrained Language Models and the emergence of prompting. We then present the fields of video summarization and memorability prediction.

2. In the chapter 3, we delve into the task of automatic video memorability prediction, showing that the task benefits from the use of multimodal approaches. We test the generalisation capabilities of our models by using a total of 5 datasets in this section. Finally, we explore new avenues such as perplexity or explainability.

3. In chapter 4, we focus on summarizing TV series with a multimodal approach and two unsupervised text approaches.

4. Finally, we devote chapter 5 to extracting story elements by developing a domain-adaptable zero-shot text classification method. We also start to investigate the capabilities of automatic question generation for story understanding.

Chapter 2

State of the Art

As discussed in the introduction, this thesis leverages on the metadata accompanying multimedia content and investigates the multimodal nature of videos. Thus, in this chapter, we present the state of the nlp and multimodal fields, to define some key concepts used throughout the thesis. Most of the contributions in the thesis are dedicated to uncovering what makes a video memorable and how to summarize the stories they contain. These two questions, on their own, unveil some interesting aspects about modern video-based storytelling, but we will in a third section of the related work, demonstrate that they also complement well the research done on automatic video summarization.

2.1 Natural Language Processing

Multimedia content can be represented as text: speech as subtitles and visual information as visual captions (or stage directions in the case of movies). Metadata such as title, description or synopsis also often accompany multimedia content. The linguistic components of multimedia content have therefore been used for a number of tasks related to video understanding. Sentiment analysis, segmentation and genre classification are just a few examples of such tasks. Similarly, much of the work realized in this thesis relies on Natural Languag Processing (NLP) tools. We therefore consider presenting the evolution of the field to be a good starting point to this related work section.

According to Liu et al. [167], the evolution of the NLP field can be summarized as a chronological succession of four paradigms:

- Fully Supervised Learning (Non-Neural Network)
- Fully Supervised Learning (Neural Network)
- Pre-train, Fine-tune

- Pre-train, Prompt, Predict

For each of these paradigms, which we will define in the next paragraphs, the effectiveness of the system relies on a specific type of engineering, respectively: features, architecture, objective and prompt. After presentation the domain of NLP at large and discussing some topics which will be of interest in the thesis, we briefly present Text Classification and Question-Answer Generation, two tasks that we investigate for the larger task of Narrative Understanding in this thesis.

2.1.1 Fully Supervised Learning

First, NLP, just like other machine learning domains, has long relied on a fully supervised learning paradigm, where for each task a specific model was trained with the input and output provided in the studied dataset. The proposed approaches depended on feature engineering, where researchers used their domain expertise to identify relevant features such as part-of-speech, sentence length or word frequency. At this stage, the field of NLP was then fragmented, with every task having their best practices. Paving the way towards a more integrated NLP field, dense vectors called word embeddings were introduced.

Previously, transforming text to numbers with a traditional count-based Bag of Words approach, created long and sparse one-hot vectors which had the dimension of the vocabulary. Word vectors, on the contrary, offer a more efficient representation with a lower dimensionality. Vectors which are close to each other in the vector space embody words or phrases with a high semantic similarity.

These vectors extracted from pre-trained models such as Word2Vec [183] and Glove [209] where used for a restricted part of the whole model parameters as they were then fed to additional neural network models which started to be massively used. Word2Vec pre-trained word embeddings capture the latent syntactic and semantic similarities among words via the two shallow architectures: Continuous Bag-of-Words (CBOW) and Skip-Gram (SG). These models rely on a feed-forward neural network with one hidden layer. CBOW takes neighboring words as input produces to learn the focus word, while SG does the inverse. To tackle the out-of-vocabulary issue (when words that are not in the training set appear in the test set) subsequent works such as CharCNN [129] or FastText [34] introduced character-level word or sub-word representations. For each downstream task, the rest of the final model parameters then need to be learned. Liu et al. [167] explain that the beginning of the deep learning era was then materialized in NLP by a shift from feature to architecture engineering. As shown in 'A Survey of the Usages of Deep Learning for Natural Language Processing' [198], such architectures mainly include Convolutional Neural Networks (CNN) [122] and different versions of Recurrent Neural Networks (RNN) [166, 306] such as GRU [52] or LSTM [107].

Because of their ability to process sequence data, RNN were used in many NLP tasks. Indeed, since for many NLP tasks, the order of elements matter, keeping a memory of the preceding words is advantageous. For example, to predict the last word of the following sentence - She is a little **girl** - the model needs to “remember” that the context preceding the word **girl**. A model which remembers the world **She** will avoid predicting **boy** instead of **girl**. As such, RNNs process sentences in a sequential way: to encode the next work, the model needs the computer hidden states of the previous word, which acts as the neural networks memory. Different variants of RNNs have been proposed, trying to address some initial limitations of the architecture. RNN only remember information from a limited number of adjacent words: as the model moves on to latter words in the sentence, the link between the first and last words stops to be captured adequately. LSTM models then tried to mitigate this issue by incorporating specific units offering a deeper processing of the hidden states. Another limitation of RNN is that they read a sentence only in one direction. However, in some cases, the context following a missing word might help to infer that word. Let’s consider the phrase "Armstrong was an American _, among the most influential figures in jazz". Without the text following the missing word, it would be complicate to know if we’re talking about the American trumpeter, astronaut or cyclist. To capture the context preceeding and following the word, Bidirectional RNNs process the sequence in both directions: forward and backward.

While CNNs are very popular in computer vision, they have also been used successfully in several NLP tasks because of their ability to detect specific patterns. Each convolution layer of CNNs allows to extract local features which size depends of the kernel, while the pooling part allows for dimensionality reduction. For language, patterns can be n-grams (groups of adjacent words) such as "I love" or "very bad". While CNN identifies patterns in space, RNN identifies patters over time. CNNs present the advantage of being faster than RNNs but they loose the information of the order of words. They are therefore more adapted to tasks such as text classification-the task of assigning predefined classes to text documents- rather than say machine translation. In short, CNNs can learn whether a specific feature is to be found in the sentence (such as a negation) but does not tell where it appeared in the sentence. If we take the example of a movie review: "I found his new movie repulsive. There are a lot of very violent scenes which bring nothing to the plot. Not only is it an unpleasant movie but there is also nothing clever about it". If the task is to tag the review as being positive or negative, the important information is to be found in ’movie repulsive’,’violent scenes’,’bring nothing’, ’unpleasant movie’ and ’nothing clever’, which can be captured by a CNN. While a RNN might allow to understand that the violent scenes didn’t bring anything to the plot (in itself ’violent scenes’ might not necessarily indicate a negative review), this added complexity is not always needed. Kim was the first one to [127] use Word2vec word vectors in combination with a CNN made of a convolutional layer, a dense layer with dropout and softmax output. They improved the state of the art with on four out of seven different tasks cast as sentence classification, including sentiment analysis and question classification.

While RNN and pretrained word embeddings were popular on different tasks, no model architecture was a clear winner for the whole NLP field and each task was rather treated independently. As stated by Qiu et al. [217], all these sequential models still suffer from locality bias and do not capture the long-range interactions between words. With the arrival of the Transformer model, which offers a solution to this issue, we observe the formation of a consensus in terms of model architectures in the field of NLP. This in turn announces a new paradigm shift towards objective function engineering.

2.1.2 With Transformer: Pre-train and Fine-tune

The creation of the Transformer architecture follows the introduction of *contextual embeddings*. When previous embeddings failed to model polysemous words, because they produced static representations, this method produces multiple embeddings for each word, varying with the context. Contextual embeddings indeed model word (or sub-word) semantics by parsing the entire sentence via a sequence model. Building on other early works on pre-trained contextual encoders [63, 166, 181, 223], Peters et al [210] created ELMO, which processes sentences in a bidirectional way. ELMO is used to extract text features which are then fed to a LSTM, trained on the task at hand. Later in 2018, based on the Transformer architecture [283] and its attention mechanism, came BERT (standing for Bidirectional Encoder Representations), a model which pushed the state of the art in many NLP tasks.

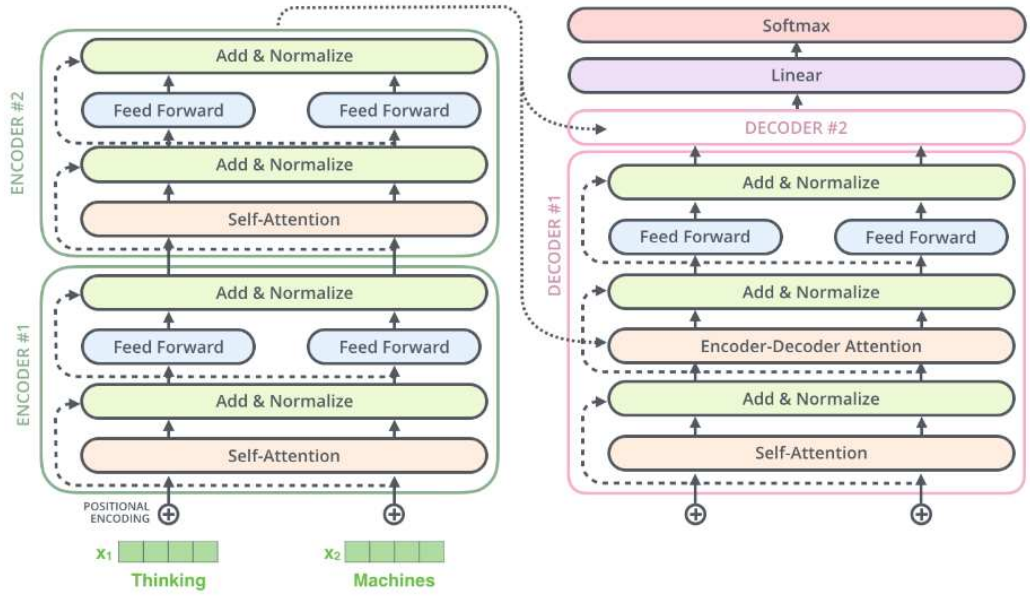


Figure 2.1: The Transformer model architecture [283][1]

The transformer architecture [283] is composed of an encoder component connected to a

[1] Illustration from https://jalammar.github.io/illustrated-transformer

decoder (see Figure 2.1). More precisely, the encoder and decoder components are made of a stack of 6 encoders or decoders respectively. Each encoder has two layers: a self-attention and a feed-forward neural network. The self-attention draws information from other words of the input sentence while encoding the particular word. Besides these two layers, the decoder also contain an encoder-decoder attention layer which tells the decoder on which parts of the input sentence it should focus. If we take the example in Figure 2.1, the words 'Thinking' and 'Machines' are both first turned into a vector before being fed to the encoder's layers. The self-attention layer transforms the initial embedding of each word into three vectors: a query, a key and a value vector ($\mathbf{q}$, $\mathbf{k}$ and $\mathbf{v}$)using for each a trainable matrix.

The self-attention mechanism is followed by a normalization layer. Finally, an attention vector is calculated (see equation 2.1)

$$\text{Attention}\left(\mathbf{q},\mathbf{k},\mathbf{v}\right) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}\right)\mathbf{v} \tag{2.1}$$

where d_k is the dimension of $\mathbf{k}$. This attention vector represents how much focus to place on other parts of the sequence.

The decoder architecture is similar to the one used by the encoder but an extra multi-head attention layer is used over the output of the latter. The last layer maps a float vector to a word.

The attention part offers to the possibility of modeling sentences of any length. Instead of the usual next word prediction, BERT was trained on *Cloze-style bi-directional language modeling* (predicting a randomly masked word) and on *next-sentence prediction*. Because the training tasks do not require annotated datasets, Language Models can be trained on large datasets in an unsupervised way and learn robust general-purpose features of the language it is modeling. In this new paradigm, the main body of the pre-trained Language Model is fine-tuned on the downstream task with a specific objective function. This paradigm is less prone to the creation of different architectures because much is now learned from the pre-trained models and experimenting with various architectures during pre-training is computationally expensive. Since BERT, a plethora of other transformer based models have been introduced : RoBERTa [170], ALBERT [140], GPT [218] or BART [148]. A comprehensive review of Pre-trained language models can be found in Qiu et al. [217] and Kalyan et al. [123]. Finally, during this thesis, what Liu et al. [167] call 'a second sea change' started to take place in 2021. After 'fully supervised training' and 'pre-train, fine-tune', the new procedure is to "pre-train, prompt, and predict".

2.1.3 Towards Pre-training, Prompting, Predicting

The new paradigm which involves a particular attention to prompt engineering, is presented in detail in Liu's et al. [167] survey on Prompting Methods in Natural Language Processing. In a nutshell, the main evolution is that instead of using objective functions to fit the pretrained models to the downstream task, the downstream task is now formulated in natural language to directly obtain an answer from the language model. For instance, if one was to recognise the emotion in the movie review "The story line was pretty awful and during the first part of the first short story i wondered what the hell i was watching", a following prompt could be "Overall, I think it was a _ movie". For translation, one could write "English: The story line was pretty awful. Spanish:_". With an appropriate prompt, a LM trained in an entirely unsupervised manner could in theory solve several tasks without the need for training [219, 220, 238].

2.1.4 Discussion on the general state of NLP

During this thesis, which took place between 2019 and 2022, the NLP field was then marked by 1) the Transformer-BERT revolution materialized by the publication of countless "BERT for X" type of papers (X being an NLP task) 2) the appearance of prompting methods.

Reacting to the advent of BERT and the Transformer architecture, Chernyavskiy et al. ask a legitimate question: is Transformers "The End of History" for NLP? (à la Fukuyama)[2] [51]? The answer is no. First, because BERT does not model everything. The study of what BERT learns and can represent (known as *BERTology*) indeed also pointed out its shortcomings. Ettinger et al. [76] for instance showed that BERT does not capture negation. Wallace et al [285] demonstrated that BERT is worse at numeracy than other models. Another important direction to improve language models is to inject external knowledge like KnowBERT [211] does with Wikipedia knowledge during pre-training. In this thesis, we focused on one particular limitation of BERT, which is its the lack of explainability of BERT and deep models in general, which is a required feature for many real world applications [86]. We also work on/and with models which relies on Common Sense Knowledge. Finally, at the end of this thesis, we also jumped on the prompting train, by exploring its potential for zero-shot classification, notably for TV series summarization and in-domain adaptation.

2.1.5 Text classification

Recently, approaches relying on large pretrained generic language models have proven very successful for a wide range of NLP downstream tasks. At the same time, due to an exponential growth in the number of complex documents, there has been a need for machine learning approaches that have the capacity to accurately classify text. Text classification can be de-

[2] https://en.wikipedia.org/wiki/The_End_of_History_and_the_Last_Man

fined as the task of assigning an appropriate category to a sentence or document, where the categories depend on the domain and topic, and benefited from those recent breakthroughs in NLP. Such systems can be deconstructed into four phases: Feature extraction, dimension reductions, classifier selection and evaluations [137].

Within multimedia understanding, text classification is consequently often used for applications such as genre classification, topic categorization or theme identification. First, an appropriate dataset needs to be selected or data collected, which is often expensive and most of the time requires human annotation. For the classifier selection phase, changes in the target labels or training corpus might require experimentation with different classifiers and repeating the search for one. Multimedia understanding models are often not very transparent and it is not always easy to recognize eventual problems with the classifier. Zero-shot classification is, however, particularly powerful because it can be used for labels partially or fully unseen during the model development, making it a ready to use tool without the need for task-specific datasets.

With the rising popularity of zero-shot classification methods, there are now more attempts to benchmark and evaluate them on text classification approaches. A major contribution to the field is Yin et al. [295] who tackled three main problems by proposing *Entail*, a zero-shot classification model based on using language models fine-tuned on the task of Natural Language Inference to classify documents: its restrictive focus on topic categorization, the treatment of labels as indices rather than as words with a meaning and an disparate evaluation with various datasets and evaluation setups. They propose a standardized evaluation on "conceptually different and diverse aspects": topics, emotions and situations. In particular, they showed that beyond the *restrictive* version of zero-shot classification (in which during a training phase, the classifier is allowed to see similar data with their labels), zero-shot classification also handles the *wild* version where the classifier does not see any examples of the labeled data.

2.1.6 Question-Answer Generation

While the task of Question Generation (QG) has not received as much attention as its sibling task of Question Answering (QA), it is a relevant task to text understanding. In particular, domain adaptation in QA often involves using the task of QG, in order to create domain specific datasets on which language models can be fine-tuned [72]. Most recent approaches rely on pre-trained transformers and often consider question generation and answer generation as dual tasks that can be combined in different ways during training [4, 43]. Another approach was to simplify QG by using a single transformer-based model for answer agnostic end-to-end question generation [171].

There have recently been remarkable efforts to make transformed-based question generation easily usable, such as the three models available at https://github.com/patil-suraj/question_generation which obtained competitive results on the SQuAD benchmark and represent different ways of treating the QG-QA paradigm. All these models are T5 based and fine-tuned on the Stanford Question Answering Dataset (SQuADv1) dataset [221]. SQuAD contains context paragraphs, each associated to sets of questions (100 000 in total) and the corresponding answer spans in these paragraphs. In Chapter 5, we use the base version of T5 as it consistently obtained better results on SQuAD than T5-small. The three models we will consider are:

- Single-task QA-QG model: Following [43], the text is first split into sentences. Then, the T5 model extracts elements that could qualify for answer like span (often NER for SQuAD) for each sentences and generates question-answer pairs. It therefore produces at least one question per sentence.
- Multi-task QA-QG: Following [4], this approach fine-tunes T5 in a multi-task way: it uses the task prefixes from T5 to extract an answer, generate a question, find the answer to the question and finally compare it to the results with the initial extracted answer.
- End-to-end QG: Following [171], the T5 model is trained to generate multiple questions simultaneously by providing the context paragraph. This model is answer agnostic and generates up to three questions per paragraph.

2.2 Multimodal Machine Learning

As in this thesis we work with audiovisual content and their metadata, we here present the state of research in multimodal machine learning. This field studies how to represent each modality - visual, audio and textual - and to combine them.

Approaching fundamentally multimodal problems like visual question answering, image retrieval or deep captioning traditionally involves using a combination of two models, each modeling their own modality. Since the deep learning era, Convolutional Neural Networks (CNN) were used for the visual modality, and Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) for the textual one. For instance, for image captioning, visual features are extracted from a CNN to then be used as input to a RNN which will create visual captions [108]. For tasks such as multimodal video summarization or multimodal sentiment analysis, which take multimodal inputs, features are usually extracted for each segment as well as modality and then injected in a model that learns to classify the given segments from the multimodal vectors. In practice, each modality was then treated independently [222].

Then, following upon the success of pretraining in the Computer Vision and NLP fields, Vision-Language (or visio-linguistic) Pre-Trained Models also emerged. Just like in NLP, the

Transformer architecture also became the backbone of such 'task agnostic' models. The paradigm consists in first encoding images and texts into dense representations, then designing an appropriate architecture to model the interaction between the two modalities and finally deciding on an appropriate pre-training task to obtain universal cross-modal representations. These models are then fine-tuned on the downstream task at hand. If their specific architectures differ, the general idea is to use self-attention or cross-attention layers to fuse both modalities. The cross-modal interaction is either modeled through a single stream (VisualBERT [153], VL-BERT [259], Oscar [155], VisDial-BERT [66]) and dual stream scheme(ViLBERT [172], Lxmert [270]). We here detail some of these models. A more comprehensive review of these visiolinguistic models, which pushed the state of the art in numerous tasks [39], can be found in Du et al. [74].

ViLBERT ViLBERT [172] is an extension of BERT which aims to learn the associations and links between visual and linguistic properties of a concept that could be a helpful feature for vision-and-language tasks. As shown in Figure 2.2, ViLBERT has a two-stream architecture modelling each modality (i.e., visual and textual) separately, and then fusing them through a set of attention-based interactions (co-attention). The keys and values of each modality are passed as input to the other modality's multi-headed attention block.

ViLBERT is pre-trained using the Conceptual Captions data set (3.3M image-caption pairs) [245] on two main tasks:

- Masked multi-modal learning: the model must reconstruct image region categories or words for masked inputs given the observed inputs.
- Multi-modal alignment prediction: the model must predict whether or not the caption describes the image content.

ViLBERT can be fine-tuned for many other tasks such as Visual Question Answering [7] and Caption-Based Image Retrieval [297]. This requires adding and training a task-specific classifier or regressor.

VisDial-BERT ViLBERT [172] has been adapted to Visual Dialog [187] by modifying the input representation to accept longer sequence (10-round long conversation). First, the model is pre-trained on English Wikipedia and BookCorpus with the masked language modeling and next sentence prediction. Next, it is trained on the Conceptual Captions and VQA with the masked image region. Finally, the model is fine-tuned on sparse annotation by getting an image, a caption, a dialog history, a question and a list of 100 possible answers. The goal is to output a sorting of the answers.

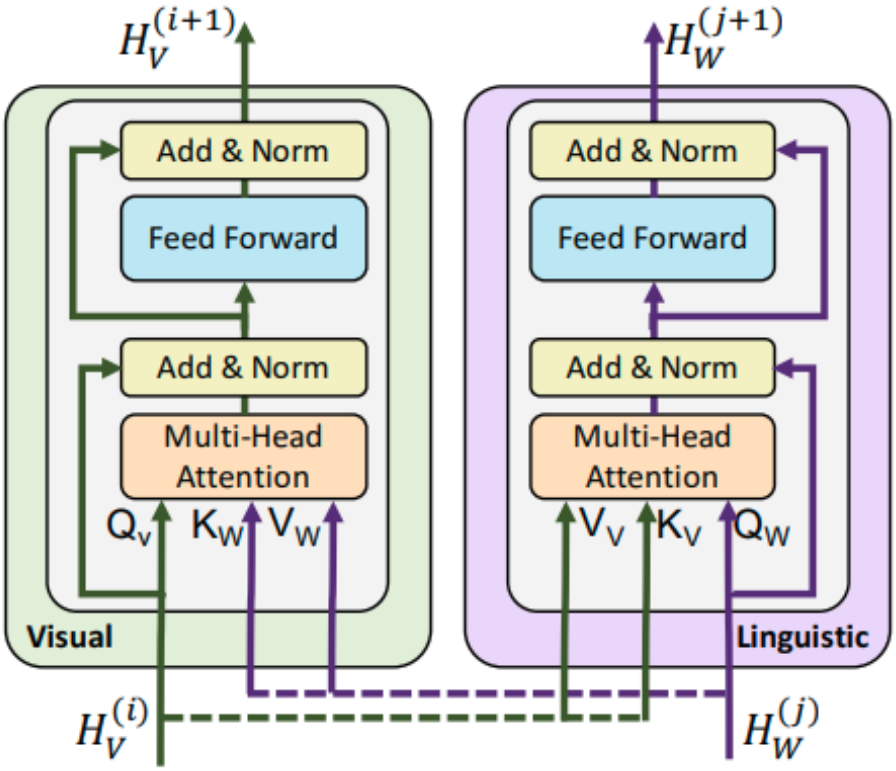


Figure 2.2: The co-attention mechanism of ViLBERT. [172]

VisualBERT VisualBERT [153] is a model inspired by BERT. It allows processing text and images jointly and using the self-attention mechanism to align elements of the input text and regions of the input image.

VisualBERT is pre-trained on COCO image caption dataset (100k images with 5 captions each). The training contains 3 phases:

- Task-Agnostic
  - Some elements of text input are masked and must be predicted.
  - Given an image and two captions, decide whether the second one describes the image.
- Task-Specific: Train the model using the data of the task with the masked language modeling.
- Fine-Tuning by introducing task-specific input, output and objective.

VL-BERT In VL-BERT [258], the visual feature embedding is newly introduced for capturing visual clues, while the other three embeddings follow the design of the original BERT paper. The visual geometry embedding is designed to inform VL-BERT the geometry location of each input visual element in the image. Each region of interest is then characterized by a 4-d vector denoting the coordinate of the top-left and bottom-right corner.

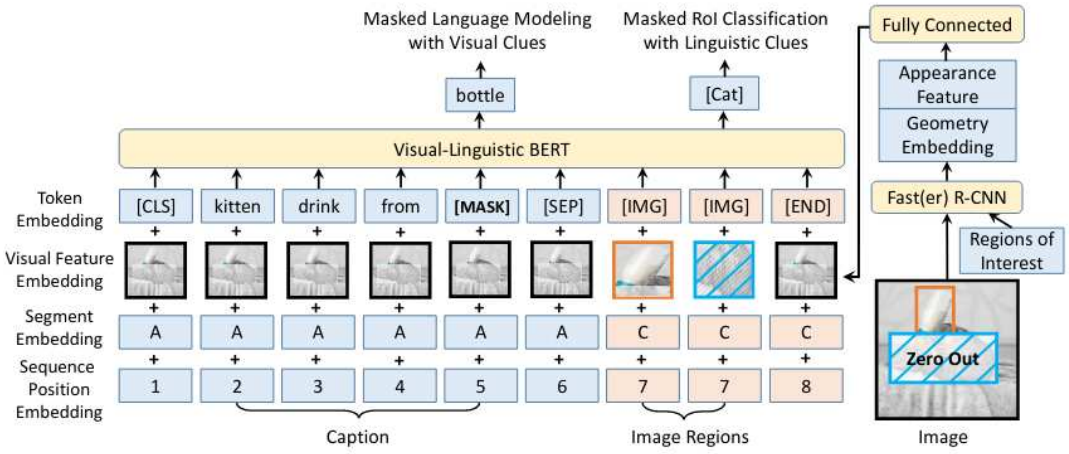


Figure 2.3: Architecture for pre-training VL-BERT. [258]

VL-BERT is pre-trained on both CC (captions are short) and BookCorpus+Wikipedia (text-only corpus to avoid over fitting on complex tasks).

- Task #1: Masked Language Modeling with Visual Clues
- Task #2: Masked RoI Classification with Linguistic Clues

Fine-tuning: the typical input formats Caption, Image and Question, Answer, Image.

Discussion In terms of research avenues to be explored, we could imagine models which also integrate the audio modality. In that direction, the very recent Data2vec [19] makes an attempt at unifying vision, speech and language. Then, because videos carry multi-modal information, being able to process a video stream instead of single images with such models would be useful. Ruan et al. [232] present the work done on Transformer based video-language pre-training. They also differentiate between single (VideoBERT [261], HERO [154], Decembert [272], VLM [292], VATT [3]and double stream architectures (CBT [260], Actbert [307] and Univl [175]).

Finally, noticing that these models are pretrained on datasets where the semantics from both image and text are aligned, we explore how to handle cases where this is not the case. The alignment vs. complementarity topic as well as related work on the particular aspect, is further explained in Section 4.2.0.1.

2.3 Video summarization

In this section, we review the literature for video summarization. For this task, videos are segmented (generally into shots or scenes) and each segment must be automatically labeled as being interesting or not. According to Bost [36], a summary must present three essential features.

- **Relativity** A summary must be related to its input source (one creates a summary *of* something). The summary can be a complete reformulation of the source (abstractive summarization) or merely a selection of some parts ordered in a certain way (generally chronologically for videos). Video summarization is often understood as an extractive task where keyframes or video skims are selected. For abstractive summarization, the video source might be turned into a textual summary.
- **Brevity** A summary must be shorter than its source. For videos, the maximum length or number of segments is generally a constraint.
- **Capturing 'important parts'**

As shown by the number of surveys published on video summarization [2, 12, 23, 24, 68, 109, 158, 184, 244, 257, 279], there are many ways to present the field. In this section, we will

briefly present its evolution - with a focus on the features used for it and on the movie/TV series domain - before exploring textual and multimodal approaches more closely, as well as outlining some specific sub-themes that we investigate in this thesis. More than an exhaustive bibliography, this section aims at presenting some gaps we identified in the field, which motivated our choice to focus on specific aspects of interestingness such as memorability and summarization of stories in videos.

2.3.1 From domain-specific to generic video summarization

At the early stage of video summarization research, most approaches focused on a certain genre, using saliency markers, generally obtained from low-level features. In the case of sport videos, for example, the notion of importance was not very ambiguous, as it could directly be derived from the rules of the sports. As such, for football [14] or baseball [46, 152], Hidden Markov Models (HMMs) were used to analyze the players and ball motions. Similarly, other in-domain knowledge was also exploited, such as the fact that salient events in sport videos are often surrounded by a peak in shot frequency (due to an editorial decision) and by loud sounds from commentators or the audience [98]. While already being less straightforward, the task of movie trailer generation, was also able to benefit from low and middle-level features. Trailers are summaries with an advertising purpose: they show video highlights which should motivate the user to see the full movie. Rather than fully unveiling the plot, this type of summary should then focus on salient scenes. As such, some methods also relied on editorial features such as sound energy [77, 114], a tool often used by movie creators to capture the interest of the viewer. Ma et al. [177], followed by Smeaton, et.al [252] then added some additional acoustic features such as the presence of speech and music which is assumed to correspond to salient sequences. Mel-frequency cepstral coefficients were then extracted to assign music, speech, or silence labels to audio segments. For the visual modality, low-level visual features such as image color/brightness, or motion intensity were used [114] and saliency maps for static shots drawn [77, 177]. Smeaton, et.al [252] relied on shot frequency. Finally, Evangelopoulos et al. [77] extended the concept of saliency to text, tagging the parts of speech of movie subtitles, following the assumption for example, that named entities are more salient than stop-words.

Some works also attempted to find a set of representative hand-crafted features for generic video summarization such such as color histogram [157], optical flow [156], and histograms of gradient [95], but it is really with deep-learning-based approaches that video summarization made tremendous progress.

2.3.2 Generic Deep Video Summarization: a vision problem?

With the deep learning era and the introduction of general-purpose video summarization datasets, a rich family of approaches, aiming to summarize any type of videos has then been proposed. These methods mainly concentrate on capturing visual interestingness. A possible reason for this focus might be that, with an increased diversity of videos in a dataset (e.g. some with speech some without), the lowest common denominator between these videos might simply be the fact that they all contain a stream of images. In any case, for SumMe [95] and TVSum [254] which are, by far, the most commonly used datasets for video summarization [12], the audio was muted during the ground truth annotation process. When working on these datasets - which contain a broad range of video domains [3]- video summarization becomes the task of detecting visual highlights.

According to a survey on general-purpose video summarization with deep neural networks [12], most techniques represent the visual content of the video frames by deep feature vectors extracted using neural networks pre-trained on large datasets for classification tasks. To obtain image features from videos, one would typically decide on a sampling rate and consider only the selected frames. Convolutional Neural Networks (CNNs) and Deep Convolutional Neural Networks (DCNNs) have been frequently used. Specific models include GoogleNet (Inception V1) [268], Inception V3 [269], AlexNet [138], as well as variations of ResNet [102] and of VGGnet [248]. Because image features, however, do not allow to recognize actions and movements, video features, which rely on 3D convolutions and allow to capture the temporal and spatial information of videos, are now also used. Models like C3D [274] and I3D [41], respectively trained on UCF-101 [255] + Sports 1 [124] and on on the Kinetics Human Action Video dataset [125], are two widely used models which allow to transition towards video summarization with action features [84, 121, 207]. The deep features are then fed to a summarizer neural network, trained with the objective of minimizing a loss function. Many supervised, unsupervised or semi-supervised architectures relying only on visual features have been developed and a comprehensive review of these can be found in Apostolidis et al. [12].

As explained in the introduction, instead of also proposing new visual approaches, we rather wish, in this thesis, to contribute to the less studied field of deep multimodal video summarization and to leverage metadata which often accompanies audio-visual content. As we also outlined previously, we are especially interested in summarizing and understanding stories in audio-visual content: stories can generally not be reduced to the visual modality (but can be transcribed to text via screenplays for example). For all these reasons, we do not describe all visual architectures developed for video summarization in further detail, but rather focus in the next sections on presenting multimodal and textual methods.

[3] holidays, events, sports, news, how-to's, user-generated-content and documentaries

2.3.3 Leveraging textual data for deep video summarization

In the survey on generic video summarization [12], 6 of the 40+ methods reviewed are multimodal. It does not necessarily mean that importance is estimated according to both the visual and audio modality of the video, but that they use multimodal representations to model interestingness i.e. besides visual information, they also use textual video metadata. The authors argue that deep learning based multimodal approaches are generally supervised and involve the use of textual video metadata (video title and description). They identify a group of techniques around semantic/category-driven summarization, which aims to increase the similarity between the semantics of the summary and of the associated metadata, action or video category [146, 253, 305]. Zhou et al. [305] learns video-level categories [4] and encourages summaries to contain category-related information by a reward mechanism. Another contribution in category-driven video summarization is Lei et al. [146], who perform action classification and with a reinforcement learning model, select the key-frames which are the most related to the action category. Otani et al. [197] project video-descriptions-pairs to a common semantic space, from which they then select video segments that correspond to cluster centers. Close to that work is Yuan et al. [299] who select meaningful segments by minimizing their distances to various video-level textual side information, after having projected both visual and textual information into a latent subspace. Wei et al. [287] generates descriptions from visual content and selects the segments matching the best with the human descriptions of video summaries. Realizing that most existing approaches neglect the audio information, Zhao et al. [304] developed an Audio Visual Recurrent Network (AVRN) to fuse audiovisual features. Finally, a very recent line of work let the user customize its summary by writing a text-based query [110, 111, 112, 120, 191, 206, 282, 291]

While the field of generic video summarization has been very prolific in the last years, methods for domain-specific video summarization continued to be proposed, as shown by Sreeja et al.'s [257]survey which presents genre-specific frameworks for video summarization. The domains for which deep-learning multimodal frameworks were proposed include soccer games [235], egocentric videos [281] as well as movies and series [29, 31, 202]. Ben et al. [29] proposed a deep multimodal framework for video segment interestingness prediction based on the genre and affective impact of movie content. For the same task, Berson et al. [31]merged textual (Word2Vec embeddings), audio (MFCC) and visual (Resnet) features. Papalampidi et al. [202] handled the task of summarizing narrative in movies by constructing a multimodal (text, audio, visual) similarity graph between movie scenes. Because movies and TV series tell a complex story, over long videos, where what is said matters, they are often accompanied by subtitles and therefore by text. Therefore the domain of entertainment is a good case for

[4] Changing Vehicle Tire, Getting Vehicle Unstuck, Groom Animal, Making Sandwich, Park-our, Parade, Flash Mob Gathering, BeeKeeping, Bike Tricks,Dog Show, Base Jump, Bike Polo, Eiffel Tower, Excavator River Crossing, Kids Playing in Leaves, MLB, NFL, Notre Dame Cathedral, Statue of Liberty, and Surfing

approaches involving text. Some methods that we present now are even purely text-based.

Papalampidi et al. [201] took upon the challenge of formalizing narrative structure. Based on expert knowledge on narratives, they consider that movie scripts contain five turning points (Opportunity, Change of Plans, Point of no Return, Major Setback and Climax) and show that it is feasible to automatically identify them from screenplays.[5]. The authors also release the so-called TRIPOD dataset [6] contains movie screenplays and Turning Points annotations. Papalampidi et al. [200] demonstrated that these turning points can also be used as a latent representation when gold standard TV series summaries are available. Hesham et al. [106] proposed Smart-Trailer (S-Trailer), a framework to create movie trailers based on subtitles only, after classifying movies into genres.

2.3.4 Discussion

2.3.4.1 How can one define 'interestingness'?

As we have seen, a summary needs to relate to its source and capture its most interesting parts while remaining brief. Although being somehow intuitive for humans, the notion of interestingness is difficult to formalize. First because it is subjective. Second because, as we have seen in this section, it is domain-dependent. For soccer games, goals are interesting and can be apprehended via motion tracking or sound reactions from the crowd. For an action movie, loud music could be associated with interesting moments. However, we already notice that even for the specific domain of entertainment videos, there are different ways to interpret video summarization and interestingness. While some works aim to capture highlights/salient moments for trailers [29, 31, 106, 252], other works consider that interesting scenes are the ones allowing to describe the narrative of a movie/TV series [37, 200, 201, 202]. Not only does interestingness depend on the domain but also on the application and the specific guidelines given during the annotation process. One way to clarify interestingness is then to define specifically the domain and the intent. This is what we do in Chapter 4, where we focus on the less studied task of summarizing stories from TV series episodes.

Now, we have also seen that research has answered the explosion in videos uploaded on the web, by developing frameworks for generic video summarization datasets. For such diverse corpus, where the specificity of each video domain is ignored, one can not explicitly know was is meant by interesting. To decrease the level of opacity around the way a user or annotator interprets the concept of interestingness, we see two solutions.

- First the field of query-based summarization where the user has to explicitly state what

[5] The authors report a 17.33% Partial Agreement score on the percentage of turning points where there is an overlap of at least one scene between the prediction and the ground truth

[6] https://github.com/ppapalampidi/TRIPOD

his interest is.

- Second, instead of asking an annotator to identify interesting moments - and therefore requesting that he internally comes with his own interpretation of interestingness - another possibility is to directly measure his cognitive reactions when being presented with video segments. Such a direction can encompass brain waves measures (EEG), eye tracking movements (ET), but also recognition tests which reveal what videos people remember.

In Chapter3, we follow this last direction and seek to automatically predict how memorable a video is.

2.3.4.2 What you hear, what you see

This literature overview also allows us to understand why Apostolidis et al. [11], suggests as a future direction for video summarization, the 'development of multimodal summarization approaches that estimate importance according to both the visual and audio modality of the video'. The correspondence between these modalities is something that we pay a particular attention to throughout this thesis: in Chapter 3, to predict memorability and in Chapter 4 in the context of multimodal TV series summarization.

2.3.4.3 The need for unsupervised approaches

Finally, creating ground-truth video summaries is a time consuming process [229] (especially if the source video is long). Apostolidis et al. [11] therefore argue that the field would benefit from more works aiming to unveil the potential of unsupervised approaches. They consider the investigation of methods, which introduce domain-specific rules to be especially relevant. As explained in the related work section on NLP, the last years saw a paradigm change with the extensive use of task-agnostic models which require no (zero shot) or limited (few-shot) training. Seizing this opportunity, in Chapter 4, we leverage on the successes of NLP to develop an unsupervised approach for TV series summarization.

2.4 Video memorability

In this section, we present the work done in the field on video memorability since 2018, the year in which the first large scale dataset annotated for video memorability was released. The creation of this field almost correlates with the beginning of this thesis. Until the end of 2020 (with the the release of the Memento 10K dataset which broaden the research in the field), video memorability was mainly approached in the context of the *MediaEval Memorability*

Challenge. Thanks to a fruitful discussion between the organizers and the participating teams, each year was not only marked by the development of new approaches by the participating teams, but also by the release of new or improved versions of video memorability datasets, offering new challenges. Our overview of the state of the art follows this dynamic between dataset creation and new approaches.

2.4.1 Context and Definition

The image memorability prediction domain [25, 79, 116, 126] developed rapidly in the last decade. Early works discovered that some semantic contents such as people, animals and objects were more memorable than some others like landscapes. With the explosion of digital video content, the task of memorability prediction expanded to videos [57, 97, 247]. However, in 2018, contrary to the field of image memorability, the field of video memorability lacked a definition and an established measurement protocol [56]. There was also no public large-scale dataset to train models on. The Memorability challenge intended to address these issues.

The Predicting Media Memorability Task asks its participants to automatically predict both a short and long term memorability for each video in the dataset. The ground truth scores were obtained from human annotators which passed two recognition tests, one a couple of minutes after the viewing session (short term score) and one after one to three days (long term score). The test consisted in pressing the space button when the person though they had been shown the video during the viewing session. With this protocol, target videos (shown more than once) and mixed with fillers videos (only shown once).

The long term measure is an addition that previous work on image memorability prediction did not cover [116, 126]. This choice was motivated by the fact that memories change in the long-term [182], and especially in the day after the memorization as shown by Ebbinghaus forgetting curve [188]. They therefore expect long-term memory scores to be more useful for most applications.

In terms of evaluation, the official metric used in the challenge is the Spearman's rank correlation between the predicted memorability scores and the ground-truth memorability scores computed over all test videos. This metrics allows for a normalization of the models' output and for an easier comparison. Pearson correlation and Mean squared error, are other non official metrics computed by the organizers.

A different take on the task of memorability prediction was proposed in 2020 by Newman et al. [193], who derived a theoretical formulation of memorability decay. They show in their study, that the hit rate decays linearly as a function of lag and that the decay rate is video-specific. Therefore, for each video, they compute a memorability score and a decay rate. The quality of the predicted curve is evaluated with a R2 score. In this thesis, we limit ourselves to

predicting short and long-term memorability scores.

2.4.2 Video memorability: Multimodality and High-level features are key

To the best of our knowledge, between 2018 and 2020, the VideoMem Dataset [57] was the only large dataset annotated for video memorability (10 000 soundless videos). This dataset, which we present in more details in Section 3.1 was created by the organizers of the *MediaEval Memorability Challenge* and used for its 2018 and 2019 editions. The winning teams of this challenge can therefore be considered state of the art for this period. One of the two winning teams [94] trained their models on visual and text features (extracted from the captions accompanying the videos). Their key findings are that models based on InceptionV3-Preds, LBP and ColorHistogram offer poor results and are outperformed by C3D-Preds and HMP based models. BoW features and high level representations learned by CNNs outperform all of the aforementioned features and their ensemble methods perform the best. Interestingly, their caption based model is linear and allows to identify some semantic patterns associated with video memorability.

Figure 2.4: Terms for the most positive coefficients. From Gupta et al. [94]

Figure 2.4 and 2.4 respectively show the most and least memorable terms according to their models. The other winning team proposed a visual model (CNN+ LSTM) [278] showing once

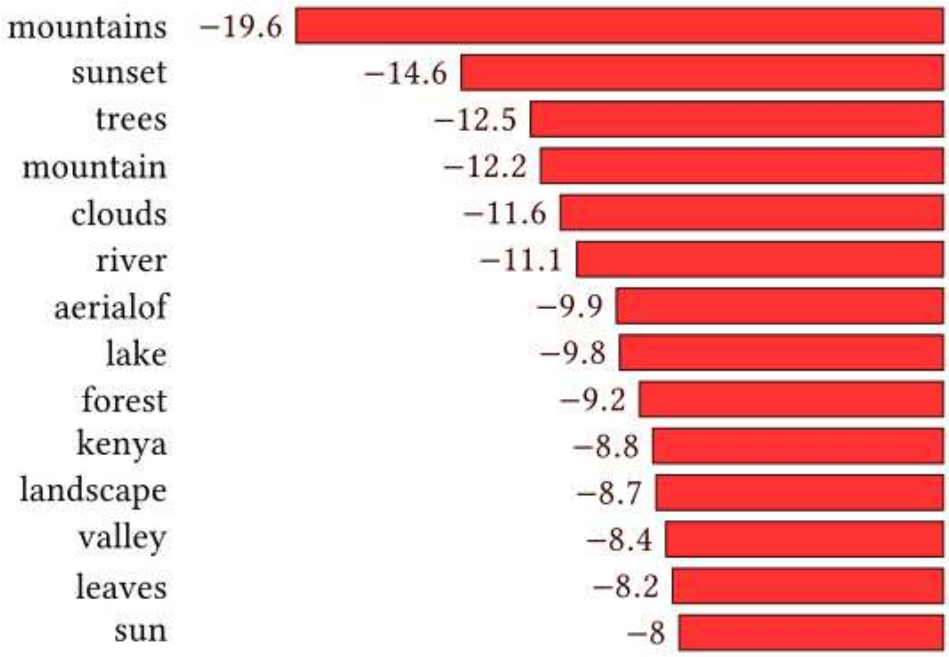


Figure 2.5: Terms for the negative coefficients. From Gupta et al. [94]

again, that high level representations extracted from deep convolutional models performed the best in terms of visual features. For the 2019 edition, we obtained the first place for long term, with our multimodal (text + visual) ensemble approach [225]. One of our contributions is to have produced additional automatically generated captions and to have extracted visual embeddings from an image to caption generation model. Our approach is further detailed in Chapter 3. The other winning team [17] also proposed ensemble methods based on CNN and text embeddings. They also experimented with other features such as emotions and C3D, obtaining a marginal gain.

While the introduction of the VideoMem Dataset [57] in the context of the memorability challenge, fostered the research in the field of video memorability, some of its limitations were pointed out. First, the audio was muted, so there was no sound or speech. Second, the videos were very short and static. These issues were addressed in the 2020 edition of the challenge with the release of the TRECVid 2019 Video-to-Text dataset dataset which contained more actions and included sound (despite generally not containing any speech).

While a majority of people first use visual cues to remember people, locations, ... [130], literature seem to have established that auditory memory is less powerful than visual memory and fades quicker [18, 32, 54]. However, just like humans tends to remember the same images, Ramsay et al. suggest that memorability can also be considered as a intrinsic property of

a sound [224]. Because it included sound, the 2020 edition of the challenge offered a great opportunity to further investigates the link between sound and memorability. As such, many teams proposed multimodal models integrating audio features. The MG-UCB team [304] used VGGish audio features [105] and we proposed to use Google AudioSet Ontology [87]. The DCU-Audio team [265], who obtained the best scores for the 2020 edition of the challenge, started investigating the power of audio features, with the best long-term memorability prediction obtained from an xResNet34 only trained on audio spectrograms. Following the release, in 2020, of the Memento10K dataset [193], another large dataset for video memorability which we present in section 3.3, Sweeney et al. [266] continued to investigate the role of audio features and proposed a multimodal late fusion system with audio gestalt threshold, with the interesting idea that not all audio information help predicting memorability. The hypothesis is that a gestalt threshold score can help differentiate between the cases were audio features are distracting and the cases were they are helpful. A gestalt can be understood as a sum of high-level conceptual audio features shown to be strongly correlated with audio memorability: imageability, human causal uncertainty (Hcu), arousal and familiarity. The authors derived proxy measures for these 4 features.

If the video obtains a gestalt over a 0.8 threshold, then in addition to frames and original captions features, audio features as well as audio augmented captions are also considered to predict the memorability score. As for the features used for the final memorability prediction, it includes image features which are extracted from a Resnet fine-tuned on LaMem [126](a dataset for image memorability) and text features from ASGD Weight-Dropped LSTM. If audio is considered, Mel-frequency cepstral coefficients or VGGish features are used and the captions are augmented with audio tags obtained via a PANNs [135] network. Thanks to the publication of Memento 10K, several other papers on video memorability were published between the 2020 and 2021 edition of the Memorability challenge. If all these works insist on the semantic features being high-quality predictors of memorability, the architectures and modalities they consider, differ. Newman et al. [193] jointly train their model on memorability prediction and on the captioning task, in order to ensure that their model extract semantic features from the dataset videos. Considering language as a 'concise, relatively cheap and comprehensible way to encapsulate semantics', Kleinlein et al. [133] focus on text.

They motivate their choice of a simpler model which extracts embeddings from a pretrained SBERT topic detection model and feed it to a linear regression, by some preliminary illustrations of the link between topics and memorability. Tables 2.6, 2.7 and 2.8 show the distribution of memorability scores within each topic detected from Sentence-BERT embeddings, respectively for VideoMem short-term, VideoMem long-term and Memento10K. For VideoMem they conclude that for short-term, topics seem to have different degrees of memorability each, whereas this does not seem to be the case for long-term memorability (they attribute the later to the relative lack of annotations per sample for the long-term problem). For Memento10K,

specific topics such as 'woman, girl, camera, food, spoon, baby and gun, man, shooting' appear to be memorable. On the contrary, topics related to nature and landscapes are less memorable. These observations are consistent with the existing literature, and notably with Figures 2.4 and 2.5.

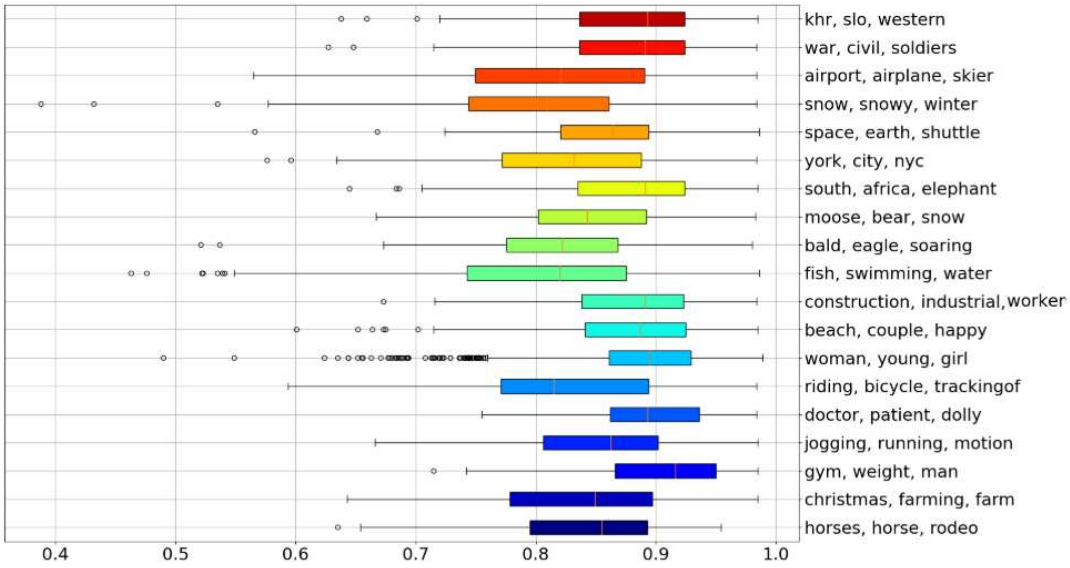


Figure 2.6: Distribution of short-term memorability scores within each detected topic in VideoMem. Kleinlein et al. [133]

Bainbridge [21] discusses our understanding of memorability for visual information in a more theoretical manner. In particular, he questions whether memorability should be considered as a single attribute, as a combination of attributes or as an arrangement of attributes. Based on literature he concludes that memorability is not synonymous with other low-level properties, nor does it serve as a proxy for an alternate high-level property. Taking the examples of faces and dance movements, where a combination of features only explained percents of variability, it posits that rather than a combination of features, an arrangement could be more appropriate. For example, he states that atypical or distinctive faces tend to be the most memorable. Similarly, images that are located in sparser areas of the attribute space (as defined by features extracted from convolutional neural networks) tend to be more memorable [174]. While this is a promising direction, it remains an open question what attributes constitute such a space, whether they contribute in equal weight, and whether there are separate influences from low-level visual features, versus high- level semantic information. In that sense, our contribution to the 2021 edition of the challenge (see Section 3), which, among others interrogates the role of perplexity as a potential proxy for sparsity when it comes to text.

With the Memorability-EEG Pilot Subtask at MediaEval'2021, Sweeney et al. [267] outlined a new direction for video memorability prediction. It aims to highlight the relevance of EEG

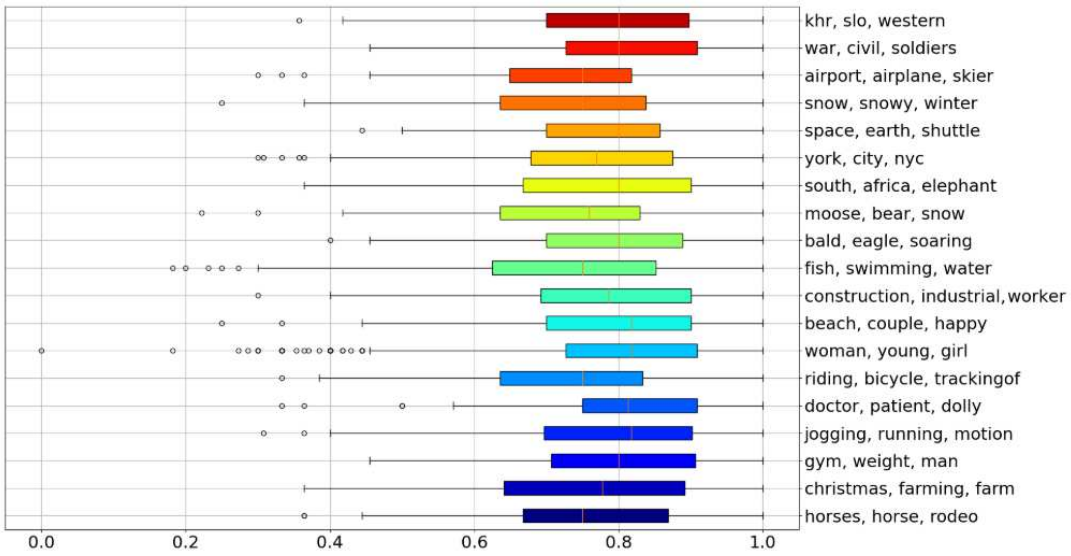


Figure 2.7: Distribution of long-term memorability scores within each detected topic in VideoMem. Kleinlein et al. [133]

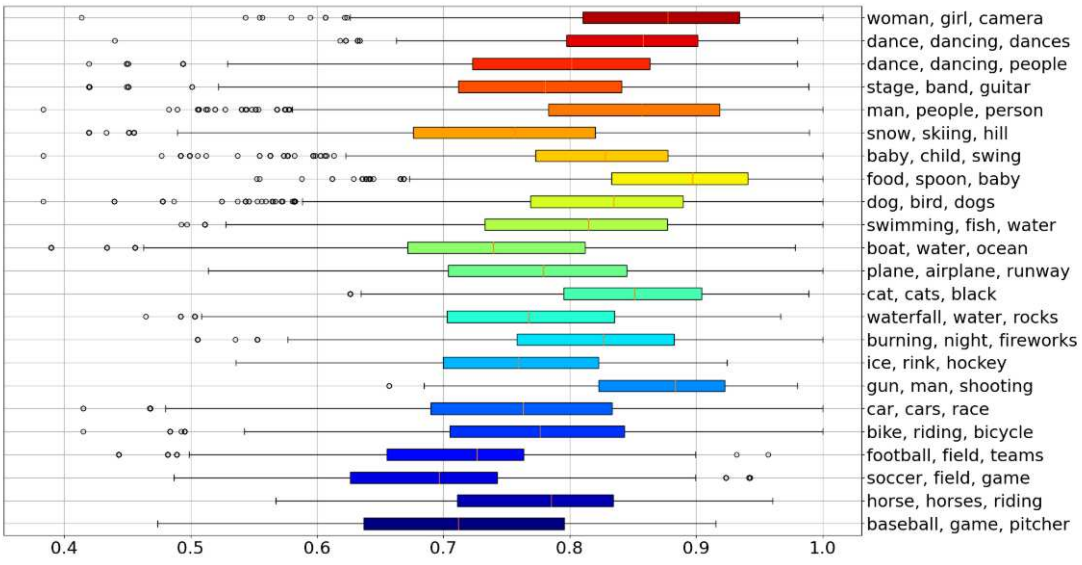


Figure 2.8: Distribution of memorability scores within each detected topic in Memento 10K. Kleinlein et al. [133]

data for the task, either alone or together with other sources of data. In particular, in order to encourage the creation of EEG-computer vision approaches, the pilot's ambition was to guide researchers who are not specialist of the field to manipulate such data.

Chapter 3

Predicting Memorability of Media Content

This chapter is dedicated to answering our first research question "How to identify memorable moments in media content?". As stated in the related work chapter 2, working with a objectively quantifiable proxy such as memorability for the task of summarization is interesting because it limits the subjectivity associated to the concept of "interesting moments". In an attempt to formalise visual interestingness, Constantin et al. [61] argue than rather being a standalone concept, it is closely linked to many aspects of perceptions such as emotions, aesthetics or memorability. Memorability, in particular, was described as ”an intrinsic property of images” [40, 115] because of its high inter-annotator agreement and has been used to create video summaries [81]. The two concepts are related but do not overlap: a video segment can be memorable without it being an essential part to include in a summary [61]. Outside of summarization, the analysis of video memorability is by itself relevant for many applications such as content retrieval, education, summarization, advertising, content filtering, and recommendation systems [60]. According to the *Benchmarking Initiative for Multimedia Evaluation* (MediaEval): "Efficient memorability prediction models will also push forward the semantic understanding of multimedia content".

In the last years, the *MediaEval Memorability Challenge* [57, 60, 85, 131], to which we participated in 2019, 2020, and 2021, has surely been the most active actor in fostering the research in automatic video memorability prediction. Over the three editions of the challenge, we have been able to explore different aspects of memorability prediction such as fusion methods of different modalities, features selection (visual, textual and audio) and the role of novelty. Besides participating to the challenges, we also investigated robustness by testing our approach on two different MeMAD datasets. These MeMAD videos correspond to broadcaster Radio and TV programs that come from two content providers: Yle (*Yleisradio Oy*, Finland’s national public broadcasting company) and INA (*Institut National de l’Audiovisuel*, a repository of all French radio and television audiovisual archives). In total, we obtained results for 5 different datasets across a variety of genres, from vines to movies. It is, here, worth mentioning that

because of license rights, different datasets were available at different points of time in the thesis. This, together with some limitations observed in the MediaEval 2018 dataset, explains why each approach is not tested on each dataset. Since we already presented the task and its related work in 2.4, we directly delve into the approaches we proposed during this thesis, presenting the different datasets along the way.

This section covers the following publications:

1. Reboud, A., Harrando, I., Laaksonen, J., Francis, D., Troncy, R., Mantecon, H.L. **Combining Textual and Visual Modeling for Predicting Media Memorability.**. In *10^{th} MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval'2019)*, 27-29 October 2019, Sophia Antipolis, France.

2. Reboud, A., Harrando, I., Laaksonen, J,. Troncy, R. **Predicting Media Memorability with Audio, Video, and Text representation**. In *11th MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval'2020)*, 11,14-15 December 2020, Online.

3. Reboud, A., Harrando, I., Laaksonen, J,. Troncy, R. **Exploring Multimodality, Perplexity and Explainability for Memorability Prediction**. In *12th MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval'2021)*, 13-15 December 2021, Online.

3.1 Combining Textual and Visual Modeling to Predict Media Memorability

The first edition of the Memorability prediction took place in 2018, prior to the beginning of this thesis. Some lessons learned from 2018 best approaches for both the long term [94] and short term tasks [278] is that high level representations extracted from deep convolutional models performed the best in terms of visual features. Furthermore, the best long term model [94] was a weighted average method including Bag-of-Words features extracted from the provided captions. Following these cues, we created a multimodal weighted average models with visual deep features and textual features extracted from both the provided video titles, as well as from automatically generated deep captions.

In this section, after presenting the VideoMem dataset, we describe the approach proposed by the MeMAD team for the MediaEval 2019. Our best approach is a weighted average method combining predictions made separately from visual and textual representations of videos. In particular, we augmented the provided textual descriptions with automatically generated deep captions. For long term memorability, we obtained better scores using the short term

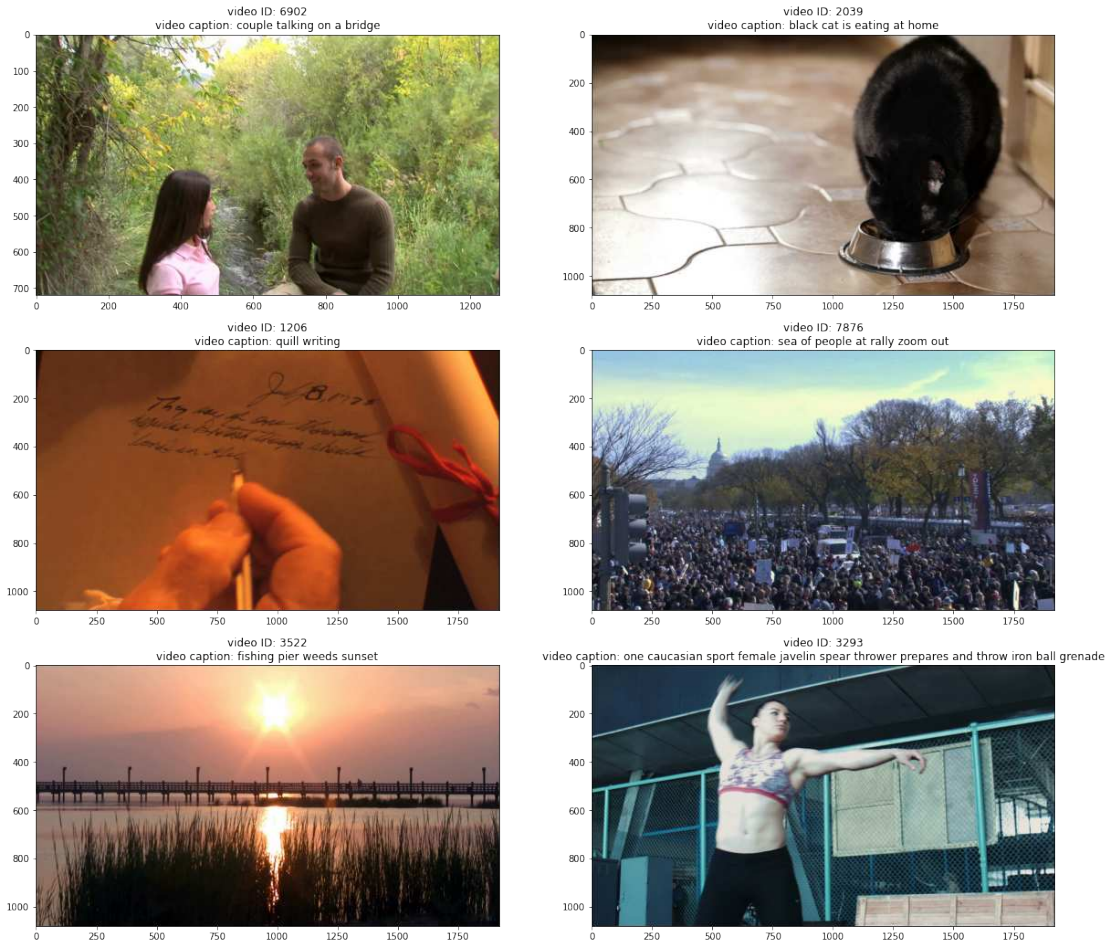


Figure 3.1: Extracted frames from random samples

predictions rather than the long term ones. Our best model achieves Spearman scores of 0.522 and 0.277, respectively, for the short and long term predictions tasks. In this section, we also interrogate the potential of visiolinguistics embeddings as an alternative to a weighted average.

The VideoMem Dataset

The dataset used in this section is from MediaEval 2018 and contains 10, 000 7-second videos, split in a a 8,000-samples development set and a 2,000-sample testing set. It contains the following fields:

- Video sources: videos are proposed in `.webm` format.
- Ground truth (only for the dev-set): video's name, short-term and long-term memorability scores, and number of annotations used to calculate scores.

- Pre-extracted frame-based visual features for the first, middle and last frames (e.g., Histogram of Oriented Gradients (HoG) [64], Color Histogram and ORB features, fc7 layer of InceptionV3 [269], Aesthetic Visual Features (AVF) [96]
- Pre-extracted video-level visual features, repretsent the motion in the sample. We provide the Histogram of Motion Patterns (HMP) [5] and the output of the final classification layer of the convolutional neural network C3D model [274]
- Short caption-like title or description text

Some examples are visible in Figure 3.1 . For the annotation of this particular dataset, the number of videos watched for the short term test 180 among which 40 target videos. During the long term annotation, 120 new fillers were added to the 40 target videos. The authors point out the difficultly of obtaining participants to annotate long-term scores, having for consequence a higher number of annotations for short than for long-term scores.

Combining modalities with a weighted average (MediaEval 2019)

Visual Approaches

With the aim of predicting the memorability scores, we created an Ensemble Approach (Figure 3.2) which combines visual and textual modalities with a weighted average of the scores obtained by each modality.

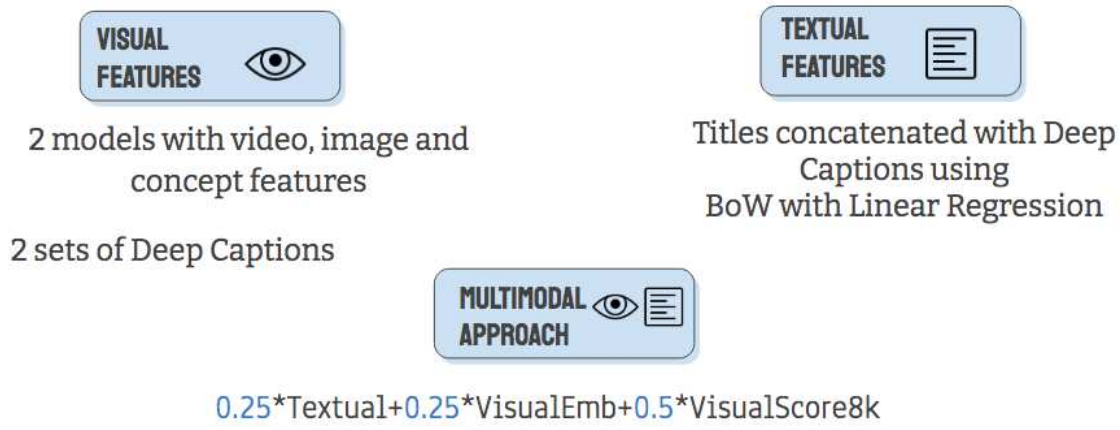


Figure 3.2: Ensemble Approach for Memorability Prediction

VisualScore. Our visual-only memorability prediction scores are based on using a feed-forward neural network with visual features in the input, one hidden layer of 430 units and one unit in the output layer. The best performance was obtained with 6938-dimensional features consisting of the concatenation of I3D [41] video features, ResNet-152 and ResNet-101 [103] image features and two versions of SUN-397 [290] concept features. The image and concept

features were extracted from the middle frames of the videos. The hidden layer uses ReLU activations and dropout during the training phase, while the output unit is sigmoidal. We trained separate models for the short and long term predictions with the Adam optimizer. The number of training epochs was selected with 10-fold cross-validation with 6000 training and 2000 testing samples.

CaptionsA. Our first captioning model uses the DeepCaption software[1] and is quite similar to the best-performing model of the PicSOM Group of Aalto University's submissions in TRECVID 2018 VTT task [251]. The model was trained with COCO [161] and TGIF [159] datasets using the concatenation of ResNet-152 and ResNet-101 [103] features as the image encoding. The embed size of the LSTM network [107] was 256 and its hidden state size 512. The training used cross-entropy loss.

CaptionsB. Our second model has been trained on the TGIF [159] and MSR-VTT [293] datasets. First, 30 frames have been extracted for each video of these datasets. Then, these frames have been processed by a ResNet-152 [103] that had been pretrained on ImageNet-1000: we keep local features after the last convolutional layer of the ResNet-152 to obtain features maps of dimensions 7x7x2048. At that point, videos have been converted into 30x7x7x2048-dimensional tensors. A model based on the L-STAP method [65] has been trained on MSR-VTT and TGIF: all videos from TGIF, and training and testing videos from MSR-VTT have been used for training, and validation has been performed throughout training with the usual validation set of MSR-VTT, containing 497 videos. Cross-entropy has been used as the training loss function. The L-STAP method has been used to pool frame-level local embeddings together to obtain 7x7x1024-dimensional tensors: each video is eventually represented by 7x7 local embeddings of dimension 1024. These have been used to generate captions as in [65].

VisualEmbeddings. The local embeddings used for CaptionsB have also been used to derive global video embeddings, by averaging the mentioned 7x7 local feature embeddings. These global video embeddings have then been fed to a model of two hidden layers, the first one and the second one having respectively 100 and 50 units, and ReLU activation function. The number of training epochs is 200 with an early stopping monitor.

Textual Approaches

Through initial experiments and from last year's results on this task, the descriptive titles provided with each video prove to be an important modality for predicting the memorability scores. In order to build on this observation, we generate captions for each video using the two visual models described above (**CaptionsA** and **CaptionsB**). While the generated captions are not always accurate, they seem to noticeably help the model disambiguate some titles

[1] https://github.com/aalto-cbir/DeepCaption

and use some of the vocabulary already seen on the training set (e.g. the title contains words such as *couple"* or *"cat"* while the generated caption would say *"a man and a woman"* or *"an animal"*, respectively, which are more common words in the training set and thus help the model generalize better on inference time). The models described in this section use a concatenation of the original provided title and the generated captions as their input.

Multiple techniques for generating a numerical score from this input sequence were considered (in ascending order of their performance on cross-validation).

Recurrent Neural Network. We use an LSTM [107] to go through the GloVe embeddings [209] of the input and predict the scores at the last token. This model performed consistently the worst, probably due to the length of the input sequence at times, and the empirical observation that word order doesn't seem to matter for this task.

Convolutional Neural Network. We use the same model as [128] except for a regression head instead of a classifier trained on top of the CNN, and GloVe embeddings as input. This model leaks less information thanks to max-pooling, and performs much better than its recurrent counterpart.

Self-attention. Similar to the previous methods, we feed our input text to a self-attentive bi-LSTM [162] to generate a sentence embedding that we use to predict the memorability scores. This model performs on par with the CNN method.

BERT. We used a pre-trained BERT model [70] to generate a sentence embedding for the input by max-pooling the last hidden states and reducing their dimension through PCA (from 768 to 250). This model performs better than the previous ones but it is more computationally demanding.

Bag of Words. We vectorize the input string by counting the number of instances of each token (and frequent n-grams) after removing the stop words and the least frequent tokens. The score is predicted by training a linear model on the counts vector. This simple model performs the best on our cross-validation, which can be justified by the lack of linguistic or grammatical structure in the titles and generated captions that would justify the use of a more sophisticated model.

For all the models considered, the addition of the generated captions improves the prediction score on the validation set considerably. It also should be noted that the use of short-term scores for long-term evaluation yields substantially better results throughout all of our experiments.

Method	Short-Term	Long-Term
Textual	0.441	0.239
VisualScore	0.495	0.268
WA1	0.512	_
WA2	**0.522**	**0.277**
WA3	0.520	0.275
WA3lt	_	0.260
Insight@DCU [17]	**0.528**	0.270
UPB-L2S [60]	0.477	0.232
RUC [286]	0.472	0.216
EssexHubTV [151]	0.467	0.203
TCNJ-CS [284]	0.455	0.218
HCMUS [277]	0.445	0.208
GIBIS [73]	0.438	0.199

Table 3.1: Our and other teams results on test set for short and long term memorability measured by Spearman score

Results and Analysis

During the evaluation process, we created four test folds of 2000 videos and therefore four models trained on 6000 videos. For the VisualScore approach, we decided to use predictions from a model trained on the entire set of 8000 videos (VisualScore8k), as well as the mean predictions from the combinations of the four models trained on 6000 videos (VisualScore6k). For the Long Term task, all models except from the WA3lt exclusively use short-term scores.

- WA1 = 0.5Textual+0.5VisualScore
- WA2 = 0.25Textual+0.25VisualEmb+0.5VisualScore8k
- WA3 = 0.25Textual+0.25VisualEmb+0.5VisualScore6k
- WA3lt = WA3 with long-term scores

Table 3.1 shows the results obtained by our different runs as well as by the other teams. When comparing our runs, we observe that the weighted average method which was trained on the whole training set and included our two visual approaches and our textual approach works the best for short term predictions. For long term prediction, one of the key observations to make is that WA3lt got the second worst results. This is consistent with our early observation that short-term scores for long-term evaluation yields substantially better results. In comparison with the other participating teams, we observe that we obtain the second best score for short-term memorability and the best long-term score. The other winning team, Insight@DCU [17], also proposed ensemble methods based on CNN and text embeddings. They also experimented with other features such as emotions and C3D, obtaining a marginal gain. In general, comparing all the approaches is difficult because of the number of parameters

differing between teams: the architecture, the modalities, the features used for one modality and so on. One thing we can say, is that the two teams who performed the worst proposed unimodal models. HCMUS [277] combines CNN features with a LSTM0 and GIBIS [73] uses I3D features with a regressor. This suggest that using text and features in combination with visual features is key. In terms of type of features, Resnet, I3D and CNN are more powerful than lower level features. Interestingly, RUC [286] and TCNJ-CS [284], on top of visual features (respectively CNN and Resnet) and text features, both used AMNet [79] an end-to-end architecture with a Soft Attention Mechanism and a LSTM, trained on a dataset for image memorability(the LaMem dataset), with only marginal gain.

Discussion

We describe a multimodal weighted average method outperforming the best results of the Predicting Media Memorability Task 2018. One of our key contribution is to have demonstrated that using automatically generated deep captions helped improving the predictions. We also conclude that, quite surprisingly, a simple n-gram frequency count was more efficient at modelling memorability than more sophisticated textual models on the text modality. Finally, the fact that long term memorability was better predicted using short term predictions indicates that the scores on long-term modality are more volatile, and that a deeper link between short and long term memorability may be at play.

Using Visio-Linguistic Models to Predict Media Memorability

In this second approach, we wanted to experiment with a visio-linguistic model described in Section 2.2. We use a pre-trained frozen version of ViLBERT to extract attention-pooled features for each modality (the output of the co-attention head represented in Figure 3.3). Those features (`pooled_output_t`, `pooled_output_v` and their fusion The output representations considered are textual H_W, visual H_V, their concatenation $[H_V, H_W]$ and their fusion (summation $H_V + H_W$ and multiplication $H_V * H_W$) are used then to separately train a regressor to predict memorability scores.

Visual Input. Since we are dealing with short video streams, choosing the middle frame of each input might allow us to reduce the computations needed to treat the whole video without losing important features. As we can notice in Figure 3.1, there is indeed a good match between the extracted middle frame and the descriptive caption, and therefore reasons to believe that we do not lose much information by focusing on the middle frame. Some early experiments, in which we randomly chose 5 frames from each video, extracted the visual features for each frame and then average them, showed that only little gain was achieved from adding this step in this particular case. This correlates with Leyva et al. [151] who observed that

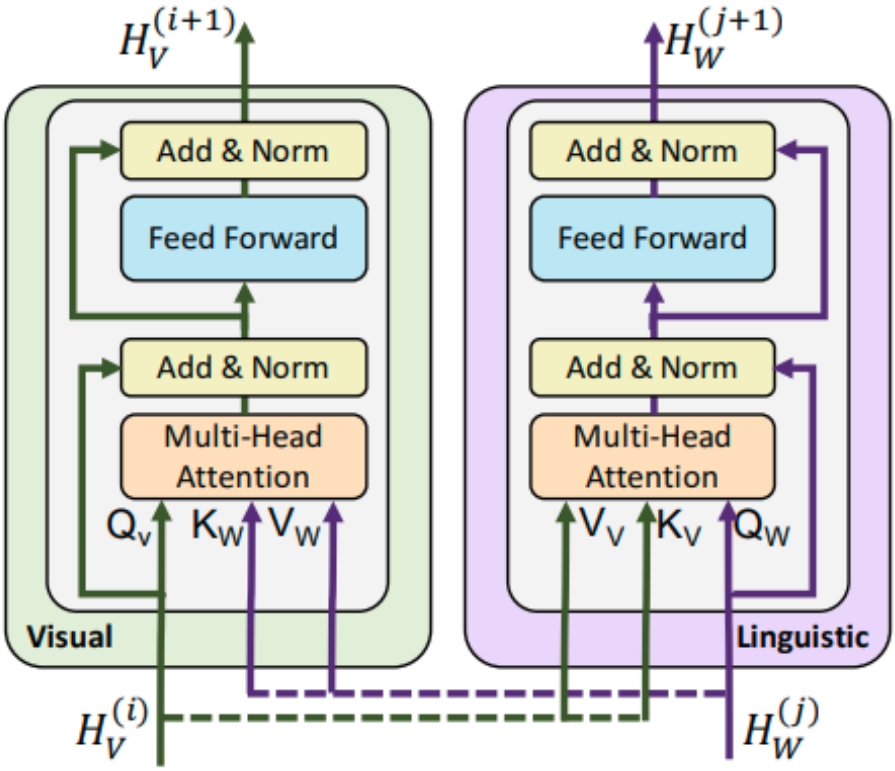


Figure 3.3: The co-attention mechanism of ViLBERT. [172]

in general, in this dataset most frames of a video are correlated and that the frame selection criterion is therefore not relevant.

For computational efficiency, we therefore stick to choosing the middle frame of each video. The next step is to extract visual features from these frames. 100 feature boundary boxes are extracted for each representative frame using maskrcnn-benchmark [179].

For the test-set results, the regressor was trained on the whole training-set and evaluated on the test-set. The different results of this approach are detailed in Table 3.2 for VQA and NLVR2 fine-tuned models for short and long-term. From this table we can conclude that summation and concatenation outperform only visual or multiplication. The model finetuned on VQA outperforms that one finetuned on NLVR2.

	H_V		$H_V * H_W$		$[H_V, H_W]$		$H_V + H_W$	
VQA fine-tuned model	0.453	0.238	0.446	0.231	0.455	0.247	**0.459**	**0.248**
NLVR2 fine-tuned model	0.417	0.208	0.427	0.215	0.416	0.209	**0.428**	**0.217**

Table 3.2: Transfer learning results

Team	Best STM Score	Best LTM Score
Insight@DCU [17]	**0.528**	0.270
MeMAD	0.522	**0.277**
UPB-L2S [60]	0.477	0.232
RUC [286]	0.472	0.216
EssexHubTV [151]	0.467	0.203
ViLBERT	0.459	0.248
TCNJ-CS [284]	0.455	0.218
HCMUS [277]	0.445	0.208
GIBIS [73]	0.438	0.199

Table 3.3: ViLBERT study and the MediaEval 2019 results.

Table 3.3 allows for a comparison with the scores obtained by the challenge teams. From this table, we can see that VilBERT is competitive but our multimodal weighted average method (MeMAD) provided better results for both subtask. In conclusion, we consider that the use of visio-linguistic features as a complement to other vision and language features, could be be a promising research avenue.

3.2 Towards real life videos: Predicting memorability of dynamic videos with sounds

Two main limitations of VideoMem, the memorability dataset we worked with so far, have been pointed out. First, the videos did not contain much action: they could more or less be summarized in a frame. Second, the videos were muted. Addressing both these limitations

Figure 3.4: A sample of frames of the videos in the TRECVid 2019 Video-to-Text dataset from [85]

would bring us closer to real life scenarios, where videos generally contain more than one action and include sound. The 2020 edition of the challenge was then marked by the use of a dataset more complex that in the previous editions. The new dataset contains short videos with more complexity (user-generated content from the Vine[2] platform rather than stock videos). This means increased difficulty in representation and the addition of the audio modality. The full description for this task is provided in [85].

In this section, we then present a method is inspired from 2019's best approaches but also acknowledges the specifics of the 2020's edition dataset. More specifically, because in comparison to last year's set of videos, the TRECVid videos contain more actions, our model uses video features and image features for multiple frames. In addition, because this year sound was included in the videos, our model includes audio features. Finally, as in Section 3.1 from our experiments with visio-linguistic features, we concluded that using such features in complement to our ensemble model, could be promising, we further test the relevance of visio-linguistic representation for the Media Memorability task. In a first part, we present the approach we developed in the context of MediaEval 2020 and the results obtained. In a second part of this section, we assess the robustness of this approach, testing it on two MeMAD datasets. The datasets are described in the relevant sections.

The TRECVid 2019 Video-to-Text Dataset

The dataset contains videos from the TRECVid 2019 Video-to-Text dataset [15]. The videos are user-generated vines (instead of stock videos as it was the case in the VideoMem dataset). According to the challenge organisers, in these videos there are 'much more action happening in them compared with those in the 2019'. Figure 3.4 shows a sample of frames of the dataset videos. We can recognise sport, landscape and funny images. In the version of the dataset used for the 2020 edition, 590 videos are used for the training set, 410 for the development and 500 for the test set. Each of them was annotated by more than 16 annotators with a short term memorability score. Similarly to the the VideoMem dataset, fewer annotations are available for the long-term scores. The game annotation protocol remains the same as the one described in Section 2.4 and in [55]. In particular, two versions of the memorability game were presented: one on Amazon Mechanical Turk (AMT) and another general purpose one for general with the following language options: English, Spanish and Turkish. In the game, participants are being shown respectively 180 and 120 videos for the short-term and long-term memorisation steps.

Audio, Video, and Text representations for Vines Memorability (MediaEval 2020)

In this section, we describe the multimodal approach proposed by the MeMAD team for the MediaEval 2020 "Predicting Media Memorability" task. Our best approach is a weighted average method combining predictions made separately from visual, audio, textual and visiolinguistic representations of videos. Our final model[3] is a multimodal weighted average with visual and audio deep features extracted from the videos, textual features from the provided captions and visiolinguistic features. It achieves Spearman scores of 0.101 and 0.078, respectively, for the short and long term predictions tasks.

Approach

We trained separate models for the short and long term predictions using originally a 6-fold cross-validation of the training set, which means that we typically had 492 samples for training and 98 samples for testing each model.

Audio-Visual Approach

Our audio-visual memorability prediction scores are based on using a feed-forward neural network with a concatenation of video and audio features in the input, one hidden layer of units and one unit in the output layer. The best performance was obtained with 2575-dimensional features consisting of the concatenation of 2048-dimensional I3D [41] video

[2] www.vine.co

[3] https://github.com/MeMAD-project/media-memorability

features and 527-dimensional audio features. Our audio features encode the occurrence probabilities of the 527 classes of the Google AudioSet Ontology [87] in each video clip. The hidden layer uses ReLU activations and dropout during the training phase, while the output unit is sigmoidal. The training of the network used the Adam optimizer. The features, the number of training epochs and the number of units in the hidden layer were selected with the 6-fold cross-validation. For short term memorability prediction, the optimal number of epochs was 750 and the optimal hidden layer size 80 units, whereas for the long term prediction these figures were 260 and 160, respectively.

We also experimented with other types of features and their combinations. These include the ResNet [104] features extracted just from the middle frames of the clips as this approach worked very well last year. The contents of this year's videos are, however, such that genuine video features I3D and C3D [275] work better than still image features. When I3D and AudioSet features are used, C3D features do not bring any additional advantage.

Textual Approach

Our textual approach leverages the video descriptions provided by the organizers. First, all the provided descriptions are concatenated by video identifier to get one string per video. To generate the textual representation of the video content, we used the following methods:

- Computing TF-IDF, removing rare (less than 4 occurrences) and stopwords and accounting for frequent 2-grams.
- Averaging GloVe embeddings for all non-stopwords words using the pre-trained 300d version [209].
- Averaging BERT [70] token representations (keeping all the words in the descriptions up to 250 words per sentence).
- Using Sentence-BERT [228] sentence representations. We use the distilled version that is fine-tuned for the STS Textual Similarity Benchmark[4].

For each representation, we experimented with multiple regression models and finetuned the hyper-parameters for each model using the 6-fold cross-validation on the training set. For our submission, we used the *Averaging GloVe embeddings* with a Support Machine Regressor with an RBF kernel and a regulation parameter $C = 1e-5$.

We also attempted enhancing the provided descriptions with additional captions automatically generated using the DeepCaption[5] software. We did not see an improvement in the results,

[4] https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-mean-tokens
[5] https://github.com/aalto-cbir/DeepCaption

which is probably due to the nature of the clips provided for this year's edition (as DeepCaption is trained on static stock images from MS COCO and TGIF datasets).

Visiolinguistic Approach

ViLBERT [172] is a task-agnostic extension of BERT that aims to learn the associations and links between visual and linguistic properties of a concept. It has a two-stream architecture, first modelling each modality (i.e. visual and textual) separately, and then fusing them through a set of attention-based interactions (co-attention). ViLBERT is pre-trained using the Conceptual Captions data set (3.3M image-caption pairs) [245] on masked multi modal learning and multi-modal alignment prediction. We used a frozen pre-trained model which was fine-tuned twice, first on the task of Video-Question Answering (VQA) [8] and then on the 2019 MediaEval Memorability task and dataset.

The 1024-dimensional features extracted for the two modalities can be combined in different ways. In our experiment, multiplying textual and visual feature vectors performed the best for short term memorability prediction but using the sole visual feature vectors worked better for long term memorability prediction. Averaging the features extracted from 6 frames performed better than only using only the middle frame. We experimented with the same set of regression models as for the textual approach. In our submission, we used a Support Machine Regressor with a regulation parameter $C = 1e-5$ and an RBF or Poly kernel respectively for short and long term scores prediction.

Results and Analysis

We have prepared 5 different runs following the task description defined as follows:

- run1 = Audio-Visual Score
- run2 = Visiolinguistic Score
- run3 = Textual Score
- run4 = 0.5 * run1 + 0.2 * run2 + 0.3 * run3
- run5 = run4 with LT scores for LT task

For the Long Term task, all models except *run5* use exclusively short-term scores. For runs 4 and 5, we normalize the scores obtained from runs 1, 2 and 3 before combining them.

Table 3.4 provides the Spearman score obtained for each run when performing a 6-folds cross-validation on the training set. We observe that our models use only the training set, as the annotations on the later-provided development set did not yield better results. We hypothesize that this is due to the fewer number of annotations per video available as many videos had a score for 1, for instance, which we do not observe on the training set.

Method	Short Term	Long Term
run1	0.2899	0.179
run2	0.214	0.1309
run3	0.2506	0.1372
run4	**0.3104**	**0.2038**
run5	0.067	0.1700

Table 3.4: Average Spearman score obtained on a 6-folds cross validation of the Training set.

Method	SpearmanST	SpearmanL
MeMAD1	0.099	0.077
MeMAD2	0.098	-0.017
MeMAD3	0.073	0.019
MeMAD4	**0.101**	**0.078**
MeMAD5	0.101	0.067
AvgTeams	0.058	0.036
DCU-Audio [265]	**0.137**	**0.113**
MG-UCB [304]	**0.136**	0.077
CUC-DMT [67]	0.06	0.049
KT-UPB	0.053	0.037
Essex-NLIP [118]	0.042	0.043
DCU@ML-Labs [144]	0.034	-0.01
GTHU-UPM [132]	0.016	-0.041
MMSys	0.007	0.048

Table 3.5: Results on the Test set for Short Term (ST) and Long Term (LT) memorability

We present in Table 3.5 the final results obtained on the test set using models trained on the full training set composed of 590 videos. We also report on the best run of every participating (two teams miss a reference because they did not submit a paper presenting their approaches). We observe that the weighted average method which uses short term scores works the best for both short and long term prediction, obtaining results which are approximately double the mean Spearman score obtained across the teams. Our best results (Spearman scores) on the test set are however significantly worse than the ones we obtained on average over the 6-folds of the training set suggesting that the test set is quite different from the training set. The results for Long Term prediction are always worse than the ones for Short Term prediction. Finally, both our scores and the mean score across team are below the ones obtained last year on the VideoMem Dataset.

The DCU-Audio team obtained the best runs for both tasks with different approaches for each subtask [265]. For the short-term, they best approach was a purely visual one, not trained on any data from the challenge. Rather they used a ResNet50 model pre-trained on ImageNet [138] and fine-tuned on the newly released Memento10k [193] dataset. The introduction of this new video memorability dataset is promising for the field and we will present, use it and investigate its potential for transfer learning in the next section. This dataset could however, not be used for transfer learning for long term scores as it does not include such scores. Instead, DCU-Audio proposed a purely audio approach which investigates the relevance of audio gestalt. Similarly to our approach, the MG-UCB team ranking second for short term memorability [304] experimented with a multimodal weighted average model including C3D [275] and ResNet152 [104] visual features, VGGish audio features [105] and GloVe [208] textual embeddings. While not necessarily visible in the final submission, the papers presenting the experiments ran by other teams who obtain final results below 0.1 short-term Spearman, give us some additional insights: CUC-DMT [67] proposed a method based on Bert and Multi-level features because on the training set, they outperform all the features provided by the organisers including C3D. DCU@ML-Labs [144] submission who is based exclusively on C3D features also performs worst than our different approaches. The results obtained by Essex-NLIP [118] confirm that colour histogram-based features such as RGBHist and HSVHist are not sufficient to capture video memorability. Finally, an important observation is that every team report significantly higher results on the dev set than on the test set.

Discussion

This work describes a multimodal weighted average method proposed for the 2020 Predicting Media Memorability task of MediaEval. One of the key contribution is to have shown that based on our experiments during the model construction or testing phase, adding more modalities is beneficial. Similarly to last year, short term scores predictions correlated better

with long term scores than the predictions made when training directly on long term scores. Finally considering the difference of results obtained between the training and test set, it would be interesting to investigate further the differences between these datasets in terms of content (video, audio and text) and annotation. We conclude that generalizing this type of task to different video genres and characteristics remain a scientific challenge.

On robustness and the influence of segmentation: Predicting Media Memorability in MeMAD corpora

Despite the limitations described above in terms of generalization of the method for predicting memorable moments, we have attempted to predict the memorability scores of segments from MeMAD videos using the ensemble approach we developed for MediaEval 2020. Our goal is to assess the robustness of our approach when being confronted with a very different dataset. We do not have ground truth annotations corresponding to short-term or long-term memorability scores that could be used for training or even for testing. We have envisioned to leverage on viewer data for programs available via IPTV, since fine-grained analytics is potentially available, considering that viewing peaks would match interesing moments. However, this data was mostly flat without highlighting clear peaks.

We selected two MeMAD datasets:

- Yle Urheiluruutu: 12 episodes of a sport magazine program[6]
- Surrey20: 20 movie excerpts with a narrative arc[7]

As we do not have any ground truth available for these datasets, we perform a post-hoc qualitative analysis, mainly focusing on the 6 segments predicted to be the most and the least memorable moments. We only considered short term memorability, since in both our 2019 and 2020 approaches, we fond out that predicted short term memorability was a better proxy for the true long term score than the predicted long term score.

Yle Urheiluruutu

Urheiluruutu is a Finnish sports magazine program by Yle highlighting the sports events and results of that day. Each episode is very short (3-5 minutes) but there are also longer programs weekly (up to 20 minutes) in which some sports phenomena are discussed in more detail. The Yle Urheiluruutu dataset is composed of 12 episodes published every day between January 6, 2021 and January 17, 2021. Each episode lasts from 4 to 20 minutes, the ones being published

[6] Production 3380 in Flow at https://platform.limecraft.com/memad/#productions/3380/material/
[7] Production 4236 in Flow at https://platform.limecraft.com/memad/#productions/4236/material/

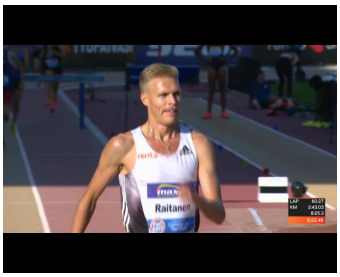
on Saturday and Sunday being at least twice longer than the ones published on week days.

We chose shots as our segment unit and we computed the memorability score per shot. As opposed to the MediaEval task setting, we did not have human written captions for every segment. For the textual part, we therefore solely rely on automatically generated deep captions using the PicSOM tool developed as part of WP2. Figure 3.5 and 3.6 show respectively the middle frame and the deep captions of some of the **most** and **least** memorable Urheiluruutu segments. We specifically chose shots from different videos (among the 12 programs) and different parts (of a program).

(a) a man in a hat is standing in the water

(b) a woman is jumping in the air on a court

(c) a man is playing a guitar while wearing a hat



(d) tennis player is holding a racket on a court

(e) a man holding a tennis racquet on a tennis court

(f) a woman is dancing and singing on stage

Figure 3.5: Middle frame and deep caption of some of the **most** memorable Urheiluruutu segments

We observe that the middle frame of the most memorable moments are much more diverse visually than the images from the least memorable moments, which are almost all ice hockey scenes. We also see a couple of faces in the most memorable segments and none in the least memorable ones. This is in line with the 2019 MediaEval dataset were a lot of video with faces were considered memorable.

In terms of automatically generated deep captions, we observed that the sports are often misidentified for the most memorable segments. In particular, 2 captions out of those 6 examples wrongly mention the sport 'tennis'. However for the least memorable segments, 'hockey' was once correctly identified. For the four other hockey pictures, the captions mention 'ski' which is incorrect but related to winter sport nevertheless. Based on these observation,

we performed a keyword search in the deep captions of the whole dataset. It showed that the words 'hockey' and 'ski' become more frequent as the memorability score of the captions drops. The keyword 'tennis', on the contrary, is more frequent in the top memorable captions.

(a) a group of people on skis in the snow

(b) a hockey player is unk a goal

(c) a man riding skis down a snow covered slope

(d) a man is running and then unk his arms

(e) a man is dancing and singing in a room

(f) a man riding skis down a snow covered slope

Figure 3.6: Middle frame and deep caption of some of the **least** memorable Urheiluruutu segments

Surrey20

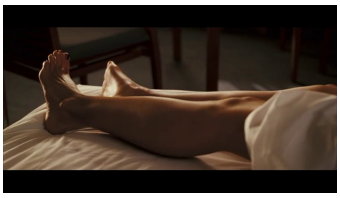
The Surrey dataset is composed of 20 movies excerpts and it has been thoroughly described in Deliverable D5.2. For this dataset, we have considered two different initial segmentations in order to predict the memorability score of each segment. First, we consider a simple shot segmentation as we did for the Yle Urheiluruutu sport magazines (Section3.2). Second, we consider the Story Grammar annotations which were human made as part of the Deliverable D5.2. We also rely on automatically generated deep captions. Figures 3.7 and 3.8 show respectively the middle frame and deep captions of some of the **most** and **least** memorable Surrey20 shots. For all Figures, each segment is from a different film excerpt.

Shot segmentation. From Figures 3.7 and 3.8, we observe that the generated deep captions seem to be more correct than the ones generated for the Urheiluruutu videos. However, it is difficult to observe any features from the captions or images that would be specific to the most or least memorable shots groups. For example in terms of captions, 'a man is sitting in a room

(a) a man is driving a car and looking at something

(b) a man is sitting in a room and talking

(c) a woman is walking down a hallway with a man



(d) a woman is laying on her stomach and smiling

(e) a woman is dancing in a room

(f) a person holding a camera in a bathroom

Figure 3.7: Middle frame and deep caption of some of the **most** memorable Surrey shots

(a) a man is walking through a door and then he stops

(b) a man is jumping on a chair and then falls

(c) a man is putting his hand on his face

(d) a man is typing on a computer keyboard

(e) a man is lying on a bed and talking

(f) a man in a suit and tie is standing in front of a window

Figure 3.8: Middle frame and deep caption of some of the **least** memorable Surrey shots

and talking' ((b) in Figure 3.7) and 'a man is lying on a bed and talking' ((e) in Figure 3.8) seem pretty close but their memorability ranking is not.

An interesting example, however, is the image (b) from Figure 3.7 and image (d) from Figure 3.8 depicting different moments of the same scene. Our approach predicted that the image with the face of the person would be memorable whereas the one without his face is one of the least memorable segment.

Overall, the results we observe might suggest that predicting what is memorable in a movie is a task that requires considering other aspects such as dialogue or information about the story. That is why, we did an experiment with the same dataset but with longer segments created by an annotator who segmented the video where the story grammar was changing.

Story Grammar segmentation. For this experiment, we made use of the dialogue that we concatenated with the automatically generated deep captions. Figures 3.9 and 3.10 show the middle frame, deep captions, subtitles and Story Grammar label of respectively some of the most and least memorable Surrey20 segments.

We observe that the memorable segments are completely different from the ones selected using a simpler shot segmentation. Among the most memorable segments, there is: one woman smiling, an injured man in the ground and someone with a birthday cake. These events are diverse but seem to be accurate candidates for memorable moments. We also can see that a large majority of the least memorable segments do not contain dialogues or only very short ones. This is interesting because our model was not trained on any dialogue data, but rather on captions. These results suggest that our model was able to somehow integrate the dialogue information. If we have a closer look at the subtitles of the most memorable segments, we find 'happy birthday', Wake up dad! Dad wake up!', or 'This is my luck. This is my luck!', which as far as a human can tell, seem to be potentially memorable moments.

Each of the segment is associated to a Story Segment which was not used as an input to our model. We can see that in the least memorable segments, four out of six are labeled as 'Setting' when none of the most memorable moments have this label. The most memorable moments, on the contrary, do not seem to be associated with one Story Segment in particular.

In conclusion, we have experimented with two different genres of videos (sport magazines and excerpt of movies), as well as two different types of segmentation (shot segmentation and human generated Story Grammar segments). It is not possible to compare these results with the ones obtained on MediaEval due to the lack of a ground truth. However, these experiments showed some interesting observations that would need to be further researched. First, our model considers some sports to be more memorable than others (tennis versus ski/hockey) which may be correlated to their frequency. Second, for movies, we demonstrate that using

(a) DeepCaption: A man in a car with a cell phone.

(b) Subtitle: Bruce: Yep, yep. Meeting started. Without me. This is my luck. This is my luck!

(c) Story Grammar: Consequence

(j) DeepCaption: a man is lying on the ground and <unk> his head

(k) Subtitle: Julio: Wake up dad! Dad wake up!

(l) Story Grammar: Consequence

(d) DeepCaption: a woman is walking down a path with a child

(e) Subtitle: No dialogue in this sequence

(f) Story Grammar: Internal Response

(m) DeepCaption: a man is sitting in a chair and smiling

(n) Subtitle: Andy: Happy birthday.

(o) Story Grammar: Plan

(g) DeepCaption: a woman is walking through a door and then falls

(h) Subtitle: Carrie: It's a little loud

(i) Story Grammar: Initiating Event

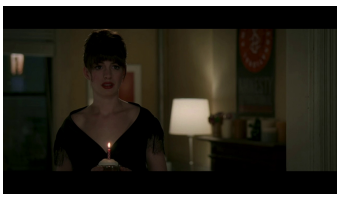
.

(p) DeepCaption: a woman is smiling and looking down

(q) Subtitle: Jenny: Heh heh heh. David: No? Alright, up to you

(r) Story Grammar: Reaction

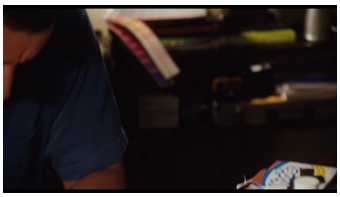
Figure 3.9: Some of the **most** memorable Surrey Story Units segments



(a) DeepCaption:

(b) Subtitle: Nate: You look really pretty.

(c) Story Segment: Consequence



(d) DeepCaption: a man is sitting in a chair and <unk> his head

(e) Subtitle: No dialogue in this sequence

(f) Story Grammar: Setting

(g) DeepCaption: a man laying on a bed with a woman

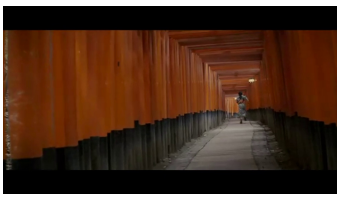
(h) Subtitle: No dialogue in this sequence

(i) Story Grammar: Setting

(j) DeepCaption: a man is looking at something and then looks away

(k) Subtitle: Julian: Uhh... RADIO: ...Sixth Avenue freeway is tied up around Lincoln, but six eighty-five is looking just dandy in both directions... more traffic reports on the 'Five' ... but coming up ...

(l) Story Grammar: Setting



(m) DeepCaption: a man is walking down a hallway and then falls

(n) Subtitle: No dialogue in this sequence

(o) Story Grammar: Setting

(p) DeepCaption: a man is looking at a woman and then she looks away

(q) Subtitle: No dialogue in this sequence

(r) Story Grammar: Reaction

Figure 3.10: Some of the **least** memorable Surrey Story Grammar segments

different segmentations produce very different results. The Story Grammar segmentation's results are more easily interpretable for humans. This could suggest that an adequate segmentation is an important requirement to obtain meaningful results. These results also suggested that using subtitles might be useful to our model, despite not having been trained on it. Finally, the Story Grammar 'setting' is very represented in the least memorable segments.

3.3 New alleys for video memorability prediction: Perplexity, Explainability and Robustness

As presented in Section 2.4, between the 2020 and the 2021 edition of the challenge, a new dataset for video memorability, Memento 10K, was released, as well as a number of papers for memorability, testing their approach on that dataset. This influenced the choices we made for our 2021 edition approach in two ways. First, we noticed that the best approaches for Memento10K are quite close to the best approaches for TRECVid 2019 Video-to-Text dataset, in the sense that they mostly use an ensemble of multimodal and high level features. While the best approaches for the Memento10K reach a score which is close to human consistency, we have seen that this was very far from being the case for the TRECVid 2019 Video-to-Text dataset. While Kleinlein et al. [133] has shown that for Memento10K, some topics are more memorable than others, we investigate if this is also the case for Vine videos while proposing an explainable approach for memorability prediction. We also hypothesize that the difficulty with Vines might be related to the fact that these videos might be more diverse, with creators aiming at breaking the expectations of the viewers. Drawing inspiration from Bainbridge [21] (2.4) who suggests that memorability might be an arrangement rather than a combination of features, we investigate the potential of perplexity as a proxy for scarcity or novelty in video captions. Second, the organisers of the challenge were able to leverage on the introduction of the new dataset, to assess the robustness of the participating teams as well as the transfer learning capabilities offered by the new dataset. Namely, they proposed a subtask, in which participants are asked to train a model on one dataset to predict the memorability of videos from the other dataset. In this section, we report results from our different approaches on the TRECVid 2019 Video-to-Text and the Memento10K dataset and for the transfer learning task, after having presented the Memento10K dataset. It is here worth mentioning that the 2021 TRECVid 2019 Video-to-Text differs a little from the previous edition: the dataset has been expanded and normalised short-term memorability scores are provided. The training set now contains 588 videos, the development set 1,116 videos and the test set 500 videos.

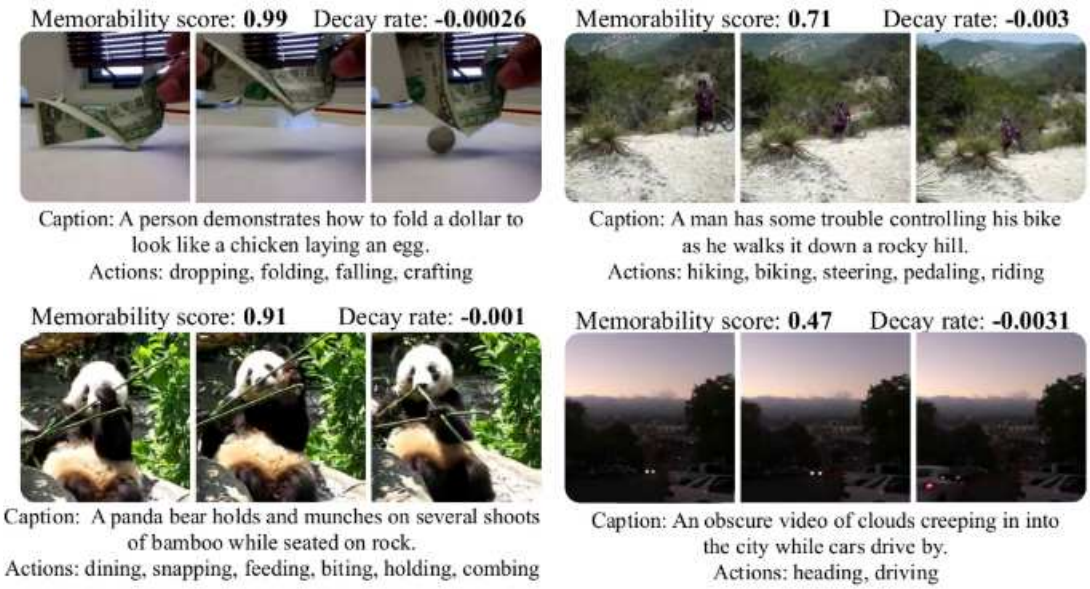


Figure 3.11: A sample of frames of the videos in the Memento10K dataset from [193]

The Memento 10K dataset

Memento 10k [8]is a multimodal memorability dataset that contains both human annotations at different delays and human-written captions. The purpose of Memento10k in this form was to create a dataset that is ideal for studying effects of delay and semantics on memorability. It contains 10,000 video clips augmented with memory scores, action labels and textual descriptions. Each video contains five human-generated captions. In total, Memento 10k contains over 900,000 individual annotations.

Crowdworkers from Amazon's Mechanical Turk (AMT) created the annotations by watching a continuous stream of three-second video clips and pressing space when they saw a repetition. The lag for repetitions has hereby been varied on purpose by the researchers behind Memento 10k. They designed this process according to established findings [126]. Natural videos have been scraped from the internet for Memento 10k, and there is a partial overlap with the Moments in Time dataset [185]. Crowdworkers were asked to identify "home videos", including videos found on social media, and everything else, e.g. professional videos, was discarded, which resulted in 10,000 videos that were respectively split into 7000, 1500 and again 1500 for training, validation and test sets. The Memento 10k contains different motion patterns including camera motion and moving objects, which makes it the most dynamic memorability dataset at the time of its release according to its creators. The mean magnitude of optical flow is almost double that of VideoMem [55](ca. 15.47 vs. 7.296). In addition, Memento 10k

[8]http://memento.csail.mit.edu

contains diverse, natural content in order to enable studies of memorability in an everyday context. Its number of annotations spreads over lags of 30 seconds to 10 minutes and is greater than in VideoMem as well (90 vs. 38 per video). This leads to higher ground-truth human consistency and makes a more robust estimation of the decay rate of a video possible. This human consistency over its 900,000 individual annotations has been measured following Cohendet et al. [55] method: the participant pool was randomly split into two groups and the Spearman's rank correlation has been calculated between the memorability rankings from each group. The rankings were hereby generated by sorting the videos by raw hit rate. Over 25 random splits the average rank correlationis 0.73, compared to 0.68 for images in [126] and 0.616 for videos in VideoMem [55]. A high human consitency speaks in favour of the existence of visual, dynamic or semantic intrisic features of the videos that can be learned to predict memorability.

Memento10k provides also more semantic information such as action labels and 5 detailed captions per video from different crowdworkers. Captions are used to augment all videos in order to relate memorability to semantic concepts. Some sample frames from the dataset are visible in Figure 3.11. Crowdworkers had been asked for a description of the events in the video clip in the form of full sentences. The captions have then been manually and carefully examined for quality and spelling mistakes by the researchers.

Multimodality, Perplexity and Explainability for Memorability Prediction

This section describes several approaches proposed by the MeMAD Team for the MediaEval 2021 "Predicting Media Memorability" task. Along with our best approach based on early fusion of multimodal features (visual and textual), we also explore the feasibility of both an explainable submission and one based on video caption perplexity to predict its memorability. The full description of this task as well as the metrics used for the evaluation is described in [131]. Our code is available at https://github.com/MeMAD-project/media-memorability.

Approach

We have experimented in the past with approaches combining textual and visual features [225] as well as using visio-linguistic models [226] for predicting short and long term media memorability. This year, we have explored other methods ranging from performing early fusion of multimodal features to attempt to explain whether some phrases could trigger memorability or not and to estimate the perplexity of video descriptions.

Early Fusion of Multimodal Features

Textual features. Our textual approach uses the video descriptions (or captions) provided by the task organizers. First, we concatenate the video descriptions to obtain one string for each video. Then, to get the textual representation of the video content, we experimented with the following methods:

- Computing TF-IDF, removing rare (less than 4 occurrences) and stopwords and accounting for frequent 2-grams.
- Averaging GloVe embeddings for all non-stopwords words using the pre-trained 300d version [209].
- Averaging BERT [70] token representations (keeping all the words in the descriptions up to 250 words per sentence).
- Sentence-BERT [228] sentence representations and in particular the distilled version that is fine-tuned for the STS Textual Similarity Benchmark[9]
- Sentence-BERT distilled version that is fine-tuned for the Emotion dataset [237]
- Sentence-BERT with the model fine-tuned on the Yahoo answers topics dataset, comprising of questions and answers from Yahoo Answers, classified into 10 topics.

For each representation, we experimented with multiple regression models and fine-tuned the hyper-parameters using a fixed 6-fold cross-validation on the training set. For our submission, we used the *Sentence-BERT on Yahoo answers topic dataset* model.

Visual features. We extracted 2048-dimensional I3D [41] features to describe the visual content of the videos. The I3D features are extracted from the *Mixed_5c* layer of the readily-available model trained with the Kinetics-400 dataset [125]. These features performance are superior to those extracted from the 400-dimensional classification output and the C3D [275] features provided by the task organizers.

Audio features. We used 527-dimensional audio features that encode the occurrence probabilities of the 527 classes of the Google AudioSet Ontology [87] in each video clip. The model uses the readily-available VGGish feature extraction model [105].

Prediction model. In all our early fusion experiments, the respective features were concatenated to create multimodal input feature vectors. We used a feed-forward network with one hidden layer to predict the memorability score. We varied the number of units in the hidden layer and optimized it together with the number of training epochs. We used ReLU

[9] https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-mean-tokens

non-linearity and dropout between the layers and simple sigmoid output for the regression result. The experiments used the same 6-fold cross-validation on the training set. The best models typically consisted of 600 units in the hidden layer and needed 700 training epochs to produce the maximal Spearman correlation score. We have also experimented with a weighted average to combine modalities, but early fusion turned out to be more successful.

Our final approach is summed up in Figure3.12

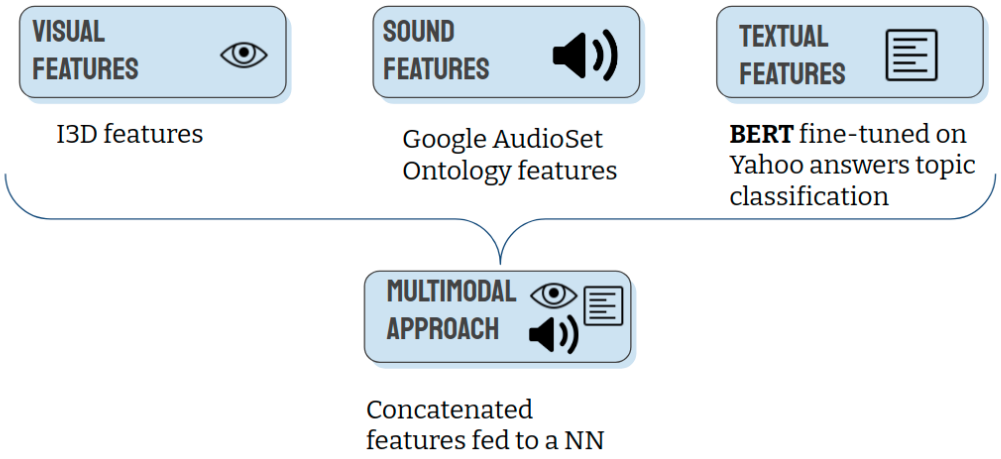


Figure 3.12: Early Fusion Approach for Memorability Prediction

Exploring Explainability

We have experimented with different simple text-based models that offer the possibility to quantify the relation between the caption and the predicted memorability score in an explainable manner. We train the models on the target dataset, i.e. for the short-term memorability predictions, we train the models on the short-term memorability scores.

We compare feeding simple linear models (regressors) interpretable input features: bag of words, TF-iDF, and topic distributions produced by an LDA model [33] trained on the corpus made of captions. Upon evaluating the performance of each model/input feature pair in a cross-fold validation protocol, we obtain the best results using TF-iDF features with a Linear Support Vector Regression (LinearSVR[10]). While this model allows us to somewhat understand the correspondence between some input words and the final score of classification (e.g. that the top words for raw and normalized short-term memorability on both `Memento10K` and `TRECVID` is *woman*), the empirical performance on both subtasks falls significantly behind other models, demonstrating both the non-linear and multimodal nature of memorability.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html

Exploring Perplexity

It has been suggested that memorable content can be found in sparse areas of an attribute space [20]. For example, images with convolutional neural networks features sparsely distributed have been found to be more memorable [174]. Additionally, we observe that the results obtained on the TRECVID dataset (made of short videos from Vine) are considerably worse than those obtained on the Memento10K dataset which may be due to the fact that the TRECVID dataset is smaller but also much more diverse. One hypothesis is that popular vines break with expectations. Backing this hypothesis, we have found, in the TRECVID dataset, that videos depicting a person eating a car, or a chicken coming out of an egg to have a high memorability score. Therefore, inspired by [176] who predicts the novelty of a caption, we wanted to test the hypothesis that the novelty of a caption influences its memorability.

We explore the (pseudo-)perplexity of each video description using a pretrained RoBERTa-large model. The score for each caption is computed by adding up the log probabilities of each masked token in the caption, and the aggregation between captions is done with a max function. We select the caption with the highest perplexity for each video. All runs have identical scores for each dataset as we do not use the training set at all in this method.

Results and Discussion

We have prepared 5 different runs following the task description defined as follows:

- run1 = Explainable (Section 3.3)
- run2 = Early Fusion of Textual+Visual+Audio features
- run3 = Early Fusion of Textual+Visual features
- run4 = Perplexity-based (Section 3.3)
- run5 = Early fusion of Textual+Visual features trained on the combined (TRECVID + Memento10k) datasets

All models except the *run1* use exclusively short-term scores for predicting the long-term score.

We present in Tables 3.6 and 3.7 the final results obtained on the test set of respectively the TRECVID and the Memento10k datasets. We comment on the Spearman Rank scores as this is the official evaluation metrics. The first thing we can observe is that these results are very much higher than those obtained on the MediaEval Vine dataset, getting close to human consistency (0.730). We observe that the early fusion method which uses short term

Method	SpSTr	PeSTr	SpSTn	PeSTn	SpLT	PeLT
run1	0.127	0.153	0.158	0.168	0.016	0.014
run2	0.216	0.212	0.221	0.209	0.060	0.090
run3	**0.220**	0.214	**0.226**	0.218	**0.063**	**0.098**
run4	–0.050	0.013	–0.052	0.018	–0.043	0.024
run5	0.196	**0.215**	0.211	**0.222**	0.062	0.059
GTHUPM [134]	0.291	0.305	0.293	0.295	0.125	0.124
Erika [173]	0.132	0.139	0.123	0.106	0.11	0.116
HCMUS [194]	0.101	0.11	0.06	0.085	0.059	0.067
AIMMLab [59]	0.297	0.312	0.26	0.267	0.097	0.114
DCU [264]	0.105	0.13	0.053	0.071	0.002	0.013

Table 3.6: Results on the TRECVID Test set for Short Term Raw (STr), Short Term Normalized (STn) and Long Term (LT) memorability (Sp = Spearman, Pe= Pearson).

Method	SpSTr	PeSTr	SpSTn	PeSTn
run1	0.464	0.460	0.463	0.458
run2	**0.658**	**0.674**	0.657	0.674
run3	0.655	0.672	**0.658**	**0.675**
run4	0.073	0.064	0.077	0.069
run5	0.654	0.672	0.651	0.671
Erika [173]	0.628	0.635	0.649	0.653
GTHUPM [134]	0.656	0.658	0.657	0.659
HCMUS [194]	0.516	0.534	0.508	0.531
AIMMLab [59]	0.648	0.652	0.648	0.65
DCU [264]	0.523	0.522	0.524	0.522

Table 3.7: Results on the Memento10K Test set for Short Term Raw (STr) and Short Term Normalized (STn) memorability.

Method	SpSTr	PeSTr	SpSTn	PeSTn	SpLT	PeLT
run1	0.076	0.099	0.068	0.091	-0.013	0.021
run2	0.140	0.165	0.146	0.170	0.045	0.042

Table 3.8: Generalisation subtask: results on the TRECVID Test set for Short Term Raw (STr), Short Term Normalized (STn) and Long Term (LT) memorability.

Method	SpSTr	PeSTr	SpSTn	PeSTn
run1	0.196	0.196	0.181	0.184
run2	0.310	0.313	0.320	0.316
DCU [264]	0.116	0.131	0.132	0.145
AIMMLab [59]	0.091	0.22	_	_

Table 3.9: Generalisation subtask: results on the Memento10K Test set for Short Term Raw (STr) and Short Term Normalized (STn) memorability.

scores works the best for both short and long term predictions. Adding the audio modality features did not improve the results. We can also observe that the results for Long Term prediction are always worse than the ones for Short Term prediction and the results for Memento10K are always better. Combining the datasets did not yield better results. This is not very surprising for the Memento10K results since it is a bigger dataset. However, the fact that augmenting the TRECVID dataset did not lead to significant improvement suggests that beyond a size difference, there is a difference in nature between the datasets that leads to a bad generalisation in terms of prediction. This fact is confirmed by the generalisation subtask which yields significantly worse results for both Memento10K and TRECVID (Tables3.9 and 3.8. Finally the scores obtained with the perplexity run were by far the lowest, only reaching 0.073 for Memento10K when our best run obtained 0.658. With this run, rather than obtaining the best results, we want to evaluate the potential for adding a caption perplexity measure. At this stage, these results do not suggest a strong relation between perplexity and memorability. Most of the participating teams (Erika, HCMUS) proposed multimodal models. Which exact features features and with fusion strategy was adopted still differ from team to the other. Erika [173] who included text, audio and visual features adopted an adaptive score fusion strategy. On the other hand, AIMMMLab [59] used a vision Transformer architectures and frame filtering method. DCU [264] preferred to submit unimodal approaches to be able to get insights into the role of each modality. Similarly to their 2020 approach, their best submission for the TRECVid dataset, was a frame based CNN that which was not trained on any data point of the dataset but rather on the Memento10K dataset. These results are in par with some of our previous experiments where the visual modality had been shown to be the most predictive. However, their results also suggest that a model trained on the Memento10K dataset generalise to the TRECVid dataset better than a model trained on the TRECVid dataset. These results do not corroborate with the results obtained by the two other teams who participated to the generalisation subtask. Similarly to us, AIMMLAB rather observed that models trained on dataset and tested on the other perform worse. On the TRECVid dataset, they obtained a Spearman of 0.091 when trained on Memento10K and 0.297 when trained on the TRECVid dataset.

3.4 Conclusion and future of the task

The different approaches we proposed together with other works done in the field, have solidly established the benefit of combining features from the visual and the text modalities. The potential of high level information such as actions or topics has also been shown. We showed that video features work better than image features. When it comes to the audio modality, the conclusion is less firm. While for the Memento10K dataset, the DCU [264] team has found that their best model was a BRR trained on VGGish audio features, despite Memento10k ground-truth scores having been collected with the sound of the videos being muted, the state of the

art scores have been obtained with approaches which do not include audio features. We also got better results on both datasets with our approach which does not include audio features. One can maybe conclude for now with the words of Sweeney et al. [266] who argue that audio information is only relevant if certain conditions are met: “audio plays a contextualising role, with the potential to act as a signal or a trigger that aids recognition”.

With regards to the performance of memorability prediction models, several things can be said to conclude this chapter. First, different approaches have reached scores that are close to human consistency (0.730) (the best score is currently 0.663). On this dataset, our approach ranks first in the challenge and second in the literature with a score of 0.658. However, the results on the TRECVid dataset are very different with the best score being 0.297 for short term memorability and 0.063 for long term memorability. All the experiments presented in this Chapter and other works from the literature confirm that long term memorability prediction remains a challenging task. This is particularly important since, when introducing the task, the organisers of the challenge stated that they consider long term scores prediction to be more useful since most videos creators would rather be interesting in being able to produce videos which are remembered in the long-term.

Both the results obtained on the generalisation subtask of the 2021 challenge and the experiments we conducted to test the robustness of our approaches on the MeMAD datasets, suggest that we are not there yet when it comes to generalisation capabilities. Just like in the case of video summarization, it might well be the case that a difference in video domain impacts the transfer learning possibilities. Vines memorability might also be particularly difficult to predict. At this stage, this is something we can only hypothesise.

In that sense, one possibility for the future of the task might is to explore new video domains such as movies. If memorability was annotated by recognition tests in all the task considered in this chapter, different definitions of memorability could also be considered.

Director Neil Jordan said "I can far more readily think of my favourite bits of movies than my favourite movies...". Consequently, in September 1999, the Observer asked its readers: ’what, in your view, were the most memorable moments in film history?’ They received 15,000 votes with over 2,500 different moments receiving at least one vote [11]. This notion of very long term collective memory has, to the to the best of our knowledge not been used for automatic memorability prediction. We consider that it would be an interesting future work to investigate whether these scenes can be automatically extracted from movies. It would also be insightful to see if these most memorable movie scenes are also scenes which are essential to the story. In the next Chapter, we interrogate this notion of importance for the narrative by proposing several approaches for narrative summarization of TV series episodes.

[11] https://www.theguardian.com/film/series/100-film-moments

Chapter 4

Narrative Summaries

For the second section of this thesis, we will investigate another aspect of relevance for audiovisual content: narrative relevance. In Chapter 3, we explored a mechanical feature which could be used for summarization, i.e., modeling the human brain's capacity to recall a scene that it has previously seen. In this chapter, we are more interested in developing approaches that can extract the important elements of a narrative. We focus on the domain of TV series, using two datasets from different genres: The CSI dataset which contains episodes of a crime series and the BBC Eastenders Dataset which is a soap opera show. We also explore different types of evaluation, from the F1 metrics used in most of the video summarization works, to an evaluation which assesses the ability of a generated summary to answer questions about *what is happening* story-wise.

In this section, we pay specific attention to two topics we pointed out in the related work (Chapter 2) as themes which would require more attention. First, Apostolidis et al. [11] suggested that 'approaches that estimate importance according to both the visual and audio modality of the video' (instead of only estimating importance according to the visual modality) would be an important direction for video summarization. In this chapter, despite not directly using audio features, we leverage on the fact that TV series screenplays both contain stage directions (explaining what happens visually) and speech transcripts, to study the correspondence between 'what is said' and 'what is done' through text. Second, because creating ground-truth video summaries is a time consuming process [229], it has been pointed out that unsupervised methods are particularly relevant. In this chapter, we develop two different unsupervised approaches: one which relies on a matching with fan-written content and one which relies on events zero-shot classification. We start by presenting a supervised approach based on the use of visio-linguistic models, to then present our two unsupervised approaches which were notably designed in the context of our participation in two *TRECVid video summarization challenge* editions (2020 and 2021).

In summary, this chapter reprises the results of the following publications:

1. Alison Reboud and Raphaël Troncy. **What You Say Is Not What You Do: Studying Visio-Linguistic Models for TV Series Summarization**. In 4^{th} Workshop on Closing the Loop between Vision and Language (CLVL), at ICCV, 2021 (Virtual).

2. Harrando, I., Reboud, A., Lisena, P., Troncy, R.
Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization. In *the International Workshop on Video Retrieval Evaluation (TRECVid'2020)*, 17-19 November 2020, Online.

3. Reboud, A., Harrando, I., Lisena, P., Troncy, R.
Zero-Shot Classification of Events for Character-Centric Video Summarization. In *the International Workshop on Video Retrieval Evaluation (TRECVid'2021)*, 7-10 December 2021, Online.

4. Reboud, A., Harrando, I., Troncy, R.
Stories of Love and Violence: Zero-Shot interesting events classification for unsupervised TV series summarization. In *Multimedia Systems - Special Issue on Data-driven Personalisation of Television Content, under review.*

4.1 Datasets for Narrative Summarization of TV Series

The CSI Dataset

The Crime Scene Investigation (CSI) dataset [83, 200] contains 39 CSI video episodes together with their screenplays segmented into scenes, each one being associated to a binary label denoting whether the scene should be part of the summary or not.[1] It also contains word-level labels indicating if the perpetrator is mentioned in the dialogue. An episode scene contains in average 21 sentences and 335 tokens. For the scenes chosen for the summary, the three human annotators had to indicate whether they selected the scene based on one/more or none of the following six reasons to justify why a scene is important: i) it revealed the victim, ii) the cause of death, iii) an autopsy report, iv) crucial evidence, v) the perpetrator, and vi) the motive/relation between perpetrator and victim. The dataset creators considered these reasons to be aspects that should be covered by crime series summaries. An episode contains in average 40 scenes from which 30% are labelled positively. Although 3 episodes (out of 39) contain a second investigation case (instead of just one), we followed the authors in assuming no such prior knowledge considering that TV series and movies often contain sub-plots.

[1] https://github.com/EdinburghNLP/csi-corpus

The TRECVID VSUM task and the BBC Eastenders Dataset

The TREC Video Retrieval Evaluation (TRECVID) aims at fostering the research in content-based exploitation and retrieval of information from digital video via open metrics-based evaluation [16]. One of the tasks proposed for the 2020[2] and 2021[3] editions is the Video Summarization Task (VSUM). The participants have to automatically summarize "the major life events of specific characters over a number of weeks of programming on the BBC EastEnders TV series". The dataset consists in 244 video episodes (464 hours) segmented into 471 527 shots, together with their transcripts. For the 2021 edition of the challenge, for five different characters of the series, the participants had to submit 4 summaries with 5, 10, 15 and 20 automatically selected shots over 10 episodes with a maximum duration of respectively 5, 10, 15 and 20 seconds. These generated summaries are evaluated by the assessors according to their tempo, contextuality and redundancy as well as with regards to how well they contain answers to a set of questions unknown to the participants before submission. Temporability refers to 'how well do the video shots flow together? Do shots cut mid-sentence? Do they flow together nicely so it wouldn't be obvious that this is an automatically generated summary'. Contextuality was defined as: 'Does the content provide the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed?' Finally, redundancy was defined as: Does the video contain content considered to be unnecessary or superfluous? (Here low is best) [16]. We believe this type of evaluation which includes more subjective metrics, complements well the F1 metrics on which the CSI dataset is evaluated.

4.2 What You Say Is Not What You Do: Studying Visio-Linguistic Models for TV Series Summarization

In this section, we generate TV series summaries using both visual cues present in video frames and screenplay (dialogue and scenic textual descriptions). Recently, approaches relying on pre-trained vision and language representations have proven to be successful for several downstream tasks using paired text and images. For TV series summarization, we hypothesize that both scenic information and dialogues are useful to generate summaries. Visio-linguistic models being presented as task-agnostic, we explore if and how they can be used for TV series summarization by conducting experiments with varying text inputs and models fine-tuned on different datasets. We observe that such generic models, despite not being specifically designed for narrative understanding, achieve results closed to the state of the art. Our results suggest also that non aligned data also benefit from this type of visio-linguistics architecture.

[2] https://www-nlpir.nist.gov/projects/tv2020/vsum.html
[3] https://www-nlpir.nist.gov/projects/tv2021/vsum.html

Introduction

The need for automatic multimedia content understanding is exemplary for pushing the research in multimodal machine learning. In a position paper, [205] extends a previously machine-centered definition of multimodality focused on representations, to a broader definition that considers a task to be '*multimodal when inputs or outputs are represented differently* ***or are composed of distinct types of atomic units of information***'. While there has been a substantial amount of work addressing multimodal representations, it is typically not combined with the question of the unit of information. Visio-linguistics tasks are approached with general pre-trained models, said to be agnostic, but created for and tested on tasks where the information between language and vision is redundant: the text generally reflects what is going on visually, therefore, neglecting a vast amount of cases where text and images rather convey complementary aspects of meaning. For example, Figure 4.1-a could be used to evaluate the tasks of automatic image captioning ("A man is lying on the floor") [30] or of visual question answering (Q: "What is the man doing?", A: "Lying on the floor") [9]. In these tasks, the information present in the text is also contained in the image and the challenge consists in aligning the two modalities.

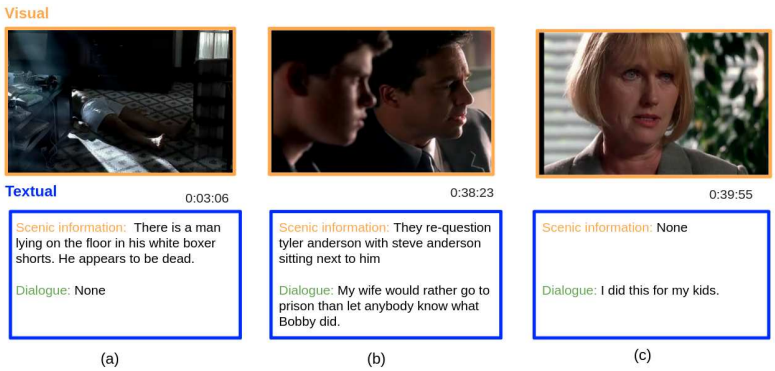


Figure 4.1: Examples of visual and textual segments from the CSI dataset [200]

Summarizing TV series episodes, however, requires to go beyond alignment as dialogue information is not available in the visual scene, and vice versa. We hypothesize that both information are nonetheless essential to this task. In this section, we want to summarize full-length crime TV series by producing shorter video summaries covering their most interesting parts. We use a dataset containing videos of the entire episodes of the CSI crimes TV series as well as their screenplays which are made of dialogues and scenic information. We expect interesting video segments to be characterized by the presence of elements such as remarkable dialogues and/or visual actions. Figure 4.1 presents three possible configurations of information spread for TV series episodes: in (a), we observe that the interesting part is contained in the scenic description, while in (b), it instead lies in the dialogue; Finally, in (c), it seems that none of the modalities is sufficient to grasp the scene content. The combination of the image description

and the dialogue, however, is more interesting: the sentence "I did this for my kids" becomes more dramatic when said in a police office. This case analysis suggests that visio-linguistics models are relevant candidates to push the frontiers of narrative summarization by adding visual information to a task that was previously only based on text [200].

When investigating the notion of complementarity for the task of TV series summarization, our work is also part of a wider reflection on multimodality and the role played by the original source of information. In an effort to assess the task-agnosticity of visio-linguistics models, we aim at shedding the light on assessing the performance of these generic models when used 'out of the box' and in particular in cases where the images and the text say different things. Focusing on TV series, our experimental setup separates the dialogue and the scenic information text which are intertwined in the screenplays. We associate each text inputs to their corresponding video frames and we assess whether both types of texts benefit from the visio-linguistics models. Screenplays contain both redundant and complementary information with respect to the visual content, while enabling to easily separate them using the punctuation signs and are of equal high quality as both types are human produced. It has recently been pointed out that the pre-training choice of these models requires more attention [250]. Consequently, we also consider different pre-training strategies that make use of varying dataset size, domains and quality of image annotations. Our results show that non aligned data can benefit from pre-training too but that the pre-training dataset should be chosen carefully as it does not always help.

The remainder of the section is structured as follows: we first present our motivation, then we describe our approach and the design of our empirical study. We then discuss the results before outlining some future work.

4.2.0.1 Motivation

We have already presented the literature for the field of video summarization and multimodal models, respectively in Section 2.3 and 2.2. In this paragraph, we do a quick recap of the precise works and observations which motivated our contribution.

Video summarization.

Multimodal video summarization is the task of selecting representative video frames or segments using multimodal integration. Recent works have pointed out that multimodal video summarization is still approached by models developed prior to the 'deep learning era' [12, 77, 99]. Deep learning based models for video summarization [95, 254] generally focuses on the visual modality, using images or text captions [35] but neglecting the content of the speech. Summarizing movies and TV series is often done using either visual or textual cues but not their combination. For example, [200] proposed a text based approach using

latent narrative structures knowledge.

Visio-linguistics models and complementarity. As opposed to earlier works in vision and language which designed models with a task-specific architecture, many multimodal approaches use pre-trained generic visio-linguistics frameworks, which are fine-tuned on the downstream task of interest. Pre-training is typically done on image captioning datasets such as Conceptual Captions [246] or COCO Captions [50] and training rely on different self-supervised objectives, such as Image Caption Matching. There are two main types of visio-linguistics architectures: dual-stream models where the two modalities are fused at a later stage such as VilBERT [172] and single-stream models where visual and textual features are directly projected into one embedding space such as VisualBERT [153]. Our work is inspired by [213] who created a new task to push the research in complementarity modelling and who successfully used this type of visio-linguistics model. In terms of approach, our work is closest to [250] who recently created an experimental setup to question common pre-training choices for these models. Noticing that MM-IMDB [13], the out of domain task for which they found no pre-training to work better, also has unpaired data (movie synopsis and posters), we push the analysis further by making the distinction between redundant and complementary modalities.

Approach

Pre-training Datasets Following [250], we select three pre-training datasets which have different characteristics that potentially play a role in finding the most appropriate model for the task: size, domain, and quality of image annotations.

COCO captions [50] contains 200K images from Flickr depicting everyday life situations containing common objects, each associated with five human-written captions in a fixed-style structure, yielding 1M image-caption pairs.

Conceptual Captions(CC) [246] is a collection of 3.3M image-caption pairs automatically scraped from the web using the alternate text of an image for captions. This process results in making CC a dataset with a very large diversity of visual content but suffering at times from noise in the captions. Due to broken links, the version used in this section has 3.1M pairs.

Multimodal IMDb (MM-IMDb) [13] contains the plot (synopsis) and poster image of 26K movies. The task associated to this dataset is to classify each of these pairs according to 23 possible genres. With a total of 3113 movies in the training set, 'Thriller' (the closest genre to CSI) is the fourth most represented genre after Drama, Comedy and Romance. Although synopsis and screenplays do not share the exact same domain, they both tell the story of a movie (so do posters and the episode videos). We include this dataset to test whether sharing the movie domain could be relevant for our task. It is also a dataset where the text and the images are not aligned.

Models Most large-scale pre-trained models have been created to handle static images. We therefore process videos as set of images (frames) and do not consider motion. We experiment with VisualBERT to account for the single-stream type of architecture and with ViLBERT for the dual-stream one. Both models treat images as region features extracted from pre-trained object detection models while text is represented as BERT global text features. In VisualBERT, these embeddings are concatenated and passed through transformer blocks (TRM). In ViLBERT, they go through two parallel transformer streams (a visual and a textual one) connected by co-attention TRM added for certain layers between the visual and textual TRM blocks. For both models, the final representation is contained in the [CLS] token and used for downstream tasks.

Experiments. We uniformly select 6 frames per video scene. We extract features for each of them and we average them afterwards. We use the MMF framework [249] for our experiments which contains, among others, the original implementation of VisualBERT [153, 250] and ViLBERT [172]. For fine-tuning via back-propagation on downstream tasks, we use binary cross entropy loss. The original CSI dataset is split in 10 folds that we re-use for our evaluation. For each fold the episodes used for training, validation and test are specified. We evaluate every 100 updates and report the model with the best loss on the validation set. We use the AdamW optimizer. The learning rate is 5e-5, a batch of size 2 and, due to computation time, we limit the training update steps to 3k (1h 16m for ViLBERT on one of the 10 folds on a NVIDIA TESLA K80 GPU). Due to class imbalance, we assigned respectively (1,3) weights to *not in* and *in* summary classes. These weights were obtained experimentally through a 10-fold cross validation with entire numbers candidates. The maximum length for textual inputs is set to 512. The default configuration as implemented in MMF is kept for the other hyper-parameters. We provide our implementation at https://github.com/alisonreboud/mmf.

Results analysis

Table 4.1 summarizes the results of our experiments using the F1 score as a metric. We also report on the performance of the SUMMER approach [200], the best performing on this dataset. The major observation we can make is that rather than a drop in performance when using complementary data (dialogue), this type of data systematically obtain better results than scenic information. More specifically, when using only dialogue text, we observe that for both ViLBERT and VisualBERT, the pre-trained CC dataset which is the most diverse and noisy dataset gives the best results and achieves near the state of the art performance without adopting a model specifically designed for narrative understanding like SUMMER does. The size of the pre-training dataset does not seem to influence the performance as MM-IMDB (the smallest) beats COCO (a dataset with a limited diversity) and no pre-training beats both. These results suggest that the diversity of the dataset is instead a decisive feature for an effective generalisation on the CSI dataset.

		Dialogue	SI	All text
-	ViLBERT	48.36	**44.84**	48.92
COCO	ViLBERT	46.85	43.98	44.82
CC	ViLBERT	**51.19**	44.01	**50.16**
MM-IMDb	ViLBERT	47.04	44.51	48.73
-	VisualBERT	49.15	46.62	**51.07**
COCO	VisualBERT	46.80	45.91	47.71
CC	VisualBERT	**50.33**	**47.48**	49.66
MM-IMDb	VisualBERT	47.83	42.22	49.79
-	Best SUMMER	-	-	**52.00**

Table 4.1: Results for all text inputs and pre-training configurations in terms of F1 score (SI = Scenic Information). We also report on the state of the art performance on this dataset obtained by SUMMER [200]

For dialogue, ViLBERT and VisualBERT obtain competitive results. Surprisingly enough, for scenic information, despite sharing the caption text domain with COCO and CC, both ViLBERT achieves better results without pre-training and for VisualBERT, only CC beats no pre-training. This suggests that the sensitivity to the domain of the data goes beyond the complementary vs redundant paradigm. The scenic information of this dataset is crime scene descriptions and therefore quite specific (probably more than dialogues). For scenic information, except for the non paired MM-IMDb dataset, VisualBERT outperforms ViLBERT, suggesting that the single-stream architecture is more powerful for aligned data.

For All text (the original screenplay combining both type of information), no pre-training and CC also obtain the best results. All text and dialogue achieve comparable results while scenic information systematically perform worse. Some possible explanations for the latter is that scenic information text is shorter and that TV series summarization benefits from complementary information. Finally, pre-training on MM-IMDb from the movie-domain dataset with non aligned data never achieved the best results. This could be due to the fact that despite sharing the movie domain with the downstream task, screenplays and TV episode videos are not similar to posters and IMDB plots. A major difference which could explain the results of this pre-training method is that posters and IMDB plots, while being indicative of a genre, avoid spoilers and therefore probably do not contain key-scenes type of content.

In summary, we observed that the dialogue text can benefit from visio-linguistics architectures and non-aligned pre-training while pre-training does however not systematically help. These observations are encouraging because they speak in favour of the possibility of relaxing the constraining requirement of having paired data for downstream tasks but also for pre-training datasets.

Conclusion

We conducted a study which isolates text elements of screenplays based on the nature of the information they convey (dialogue versus scenic information) and we tested different pre-training methods on two visio-linguistic models for the task of TV series summarization. We have shown that using a visio-linguistic architecture without paired data and without in-domain pre-training achieves near state of the art results. The fact that even with a small dataset, no pre-training beats some pre-training choices underlines the importance of in-domain and/or diverse pre-training datasets. In the future, our goal is to experiment pre-training with in-domain datasets such as movie captioning datasets [230] and video subtitles, to experiment with a very diverse pre-training dataset where the image-text alignment constraint is relaxed and to work with architectures handling videos and their temporal information [261]. In order to get more insights into the benefits of introducing images, we also plan to compare the performance of these visio-linguistic models with a text-only, general-purpose architecture such as BERT [71]. Finally, while our results suggest that the use of task-agnostic visual-linguistic models without paired data is a promising direction to look at, both for pre-training and downstream dataset, the conclusions about the possible use of complementary data need to be corroborated by more experimental results on other downstream tasks (than TV series summarization). Using visualisation techniques would also allow for a better understanding of the type of relation that the model learns between images and text, especially for complementary data.

4.3 Unsupervised TV Series Summarization: a method based on synopsis alignment

In this section, we describe our first unsupervised approach to TV series summarization. It is a character-centered approach which we developed in the context of the 2020 TRECVid [16] Video Summarization Task. The challenge was described in more details Section 4.1 and in [16]. Our approach relies on fan-made content and, more precisely, on the BBC EastEnders episode synopses from its Fandom Wiki[4]. This additional data source is used together with the provided videos, scripts and master shot boundaries. We also use BBC EastEnders characters' images crawled from the Google search engine in order to train a face recognition system. All our runs use the same method, but with varying constraints regarding the number of shots and the maximum duration of the summary. The shots included in the summaries are the ones whose transcripts and visual content have the highest similarity with sentences from the synopsis.

For all submitted runs, the redundancy score improved with the number of shots included in

[4] https://eastenders.fandom.com/wiki/EastEnders_Wiki

the summary while the relation with the scores for tempo and contextuality seem to vary more. The scores are lower for the question answering evaluation part. This is rather unsurprising to us as we realized while deciding on a similarity measure score that it is challenging for humans to choose between two potentially interesting moments without knowing beforehand the questions included in the evaluation set. Overall, we consider that the results obtained speak in favour of using fan-made content as a starting point for such a task. As we did not try to optimize for tempo and contextuality, we believe there is some margin for improvement. However, the task of answering unknown questions remains an open challenge.

Approach

Our fan-driven and character centered approach is presented in Figure 4.2.

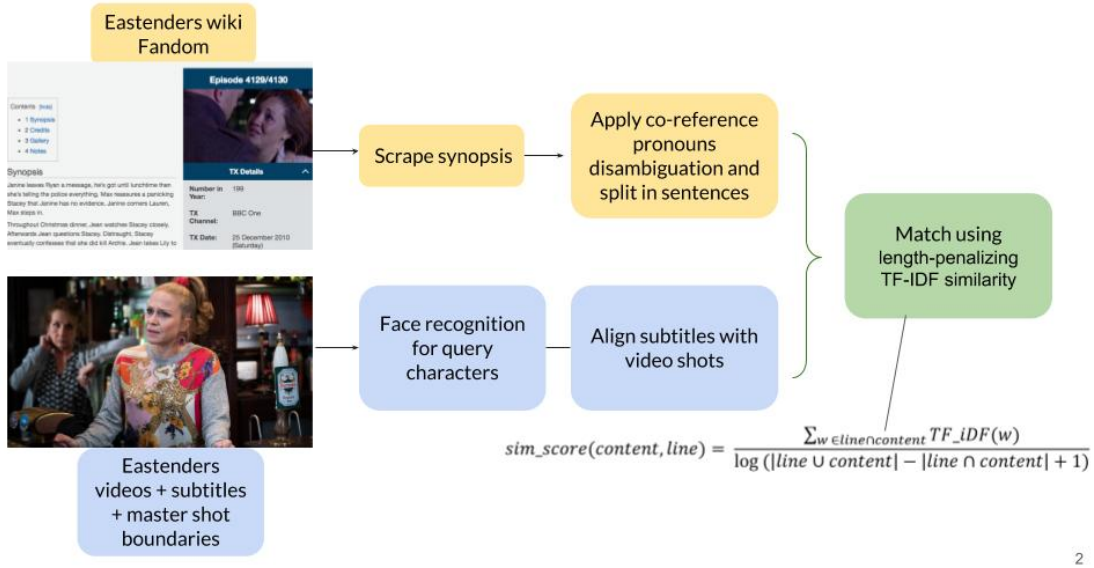


Figure 4.2: TRECVid 2020 - Wiki-driven and character-centered approach illustration.

Scraping Synopses From the Fandom Wiki and Selecting Shots

The first step of our approach consists in scraping synopses available on the Fandom EastEnders Wiki[5].

Our main hypothesis is that every sentence (ending with a period) represents an important event to be added to the final video summary. We scrape the Synopsis and the Cast sections for each episode broadcasted between the dates of the provided episodes. The mapping between

[5]https://eastenders.fandom.com/wiki/EastEndersWiki

the episodes and their dates is in `eastenders.collection.xml` provided by the challenge organizers.

In parallel, we extract the shots in which the three characters of interest appear from the video. We run the Face Celebrity Recognition library[6], a system that relies on pictures crawled from search engines using the actor's name as search keyword. In our experiments, we have added "EastEnders" to the character names in order to avoid retrieving pictures of different people with the same name. For each picture, faces are detected using the MTCNN algorithm and the FaceNet model is applied to obtain face embeddings. Following the assumption that the majority of faces are actually representing the searched actor, other faces – e.g. person portrayed together with the actor – are automatically filtered out by removing outliers until the cosine similarity of face embeddings has a standard deviation below a threshold of 0.24 which has been empirically defined.

The remaining faces are used to train a multi-class SVM classifier, which is used to label the faces detected on frames. For more consistent results between frames, the Simple Online and Realtime Tracking algorithm (SORT) has been included, returning groups of detection of the same person in consecutive frames.

We select the shots displaying any of the the three characters of interests, keeping only those detection having a confidence score greater than 0.5. We also tried to use speaker diarisation to corroborate the visual information about the characters. However, given the limitations of the current technologies in terms of number of characters and the difficulty of identifying the character corresponding to each voice, we could not pursue the idea further.

Synopses and Transcript Pre-Processing

A synopsis for each episode was created using the provided files *eastenders.collection.xml* and *eastenders.episodeDescriptions.xml*. Since these were "EastEnders Omnibus" episodes, they correspond to multiple actual weekday episodes. We use the dates and the continuation to generate one synopsis for each "long" episode (typically made of 4 episodes). We then split the synopses into sentences and performed coreference resolution on the synopses to explicit character mentions [7]. In parallel, the provided XML transcripts were also converted into timestamped text and aligned with the given shot segmentation. Finally, both the synopses sentences and shot transcripts were lower cased, stop words removed and lemmatized.

We also produced automatically-generated visual captions following the method presented by the PicSOM Group of Aalto University's submissions for the TRECVid2018 VTT task [251]. The hypothesis is that by describing the visual information of a shot, visual captions could

[6] https://github.com/D2KLab/Face-Celebrity-Recognition
[7] https://github.com/huggingface/neuralcoref

complement well the dialog transcript and therefore allow for a better matching between the shots and synopses sentences.

Matching and Runs Generation

We perform a synopsis sentence / shot transcript pairwise comparison by generating a similarity score. We define similarity between two sentences as the sum of TF-IDF weights (computed on the transcript) for each word appearing in both of them, divided by the log length of the concatenation of both sentences, thus penalizing long sentences that match with many transcript lines.

Next, we order the shot by similarity score, picking only the best match for each shot (but not the other way around). This gives us scenes we are sure to appear in the summary, but not necessarily any guarantee about how important these scenes are. We also performed the pairwise comparison adding the automatically generated captions. A qualitative assessment revealed, however, that the captions were too noisy to complement well the transcript. We also make sure that if a line of dialog runs through the next shot, we include the next shot as well to improve the smoothness of the viewing. However, this heuristics was only relevant for the longest run (20 shots). Each run is made by selecting the N most matching shots out of the top, in chronological order.

Results and Analysis

The final results for the two teams which have participated in TRECVid VSUM are presented in Table 4.2 while the detailed scores of our approach are presented in Table 4.3. NII_UIT, the other participant team to the 2020 edition, relied purely on visual features [143]. Their importance score is an average of a face recognition (of the characters of interest) and a representation score. The later is obtained from the sequence to sequence model VASNet [80] which takes as input features extracted from GoogLeNet [268] trained on ImageNet [234]

Our method obtains the best overall score for each of the 4 required runs. The mean scores (range 1 - 7. High is best) for tempo, contextuality and redundancy are all above average (respectively 4.75, 4.75, 4.1) despite the fact that our method does not specifically attempt to optimize these metrics. However, in terms of question answering, the results show that the shots selected did not allow to answer more than two (at best) of the five questions. More specifically, Table 4.4 shows (in bold) the questions that were answered in at least one of ours runs. We notice that most of the questions started either with 'What' or 'Who' and that our approach performed equally for both types of questions.

TeamRun	Percentage
MeMAD1	31%
MeMAD2	31%
MeMAD3	**35**%
MeMAD4	32%
NIIUIT1	9%
NIIUIT2	8%
NIIUIT3	8%
NIIUIT4	6%

Table 4.2: TRECVid 2020 average score for each run and team [143].

Query	Tempo	Contextuality	Redundancy	Q1	Q2	Q3	Q4	Q5
Janine1	6	4	5	No	No	No	No	Yes
Janine2	5	5	6	No	No	No	No	Yes
Janine3	5	5	6	No	No	No	No	Yes
Janine4	5	5	7	No	No	No	No	Yes
Ryan1	4	5	3	No	No	No	No	Yes
Ryan2	5	5	3	No	No	No	No	Yes
Ryan3	3	4	5	No	No	No	Yes	Yes
Ryan4	2	3	5	No	No	No	Yes	Yes
Stacey1	6	5	2	No	Yes	No	No	No
Stacey2	6	5	2	No	Yes	No	No	No
Stacey3	6	6	2	No	Yes	No	No	No
Stacey4	4	5	4	No	Yes	No	No	No

Table 4.3: TRECVid 2020 detailed score for MeMAD's approach.

Discussion and Outlook

This work describes a character centered video summarization method based on fan-made content, subtitles and face recognition. One of the key contribution of this paper is to have demonstrated that despite some noise from face detection and recognition, this method enables to capture multiple important plot points for all three query characters. We also conclude that adding more shots to the summaries did, quite surprisingly, not always allow to answer more key moments related questions. Finally, we would like to pinpoint the fact that the task of choosing important sequences that would answer unknown questions, is very challenging for humans. Indeed, when generating the runs, having read the summaries but not having watched the videos, we find it challenging to decide which sequences should be included in the summary. It would be interesting to know how much the score would improve if we would know the questions before evaluation.

Character	Q#	Question
Janine	Q1	What is causing Ryan to be sick in bed?
Janine	Q2	How does Janine attempt to kill Ryan while in the hospital?
Janine	Q3	What happens when Janine attempts to play recording of Stacey?
Janine	Q4	Who stabbed Janine?
Janine	Q5	**Who gives Janine the recording of Stacey?**
Ryan	Q1	How does Janine attempt to kill Ryan in the hospital?
Ryan	Q2	What does Ryan do when Janine is lying in the hospital?
Ryan	Q3	Where is Ryan trapped?
Ryan	Q4	**What does Ryan tell Phil he can do for him?**
Ryan	Q5	**Who is Ryan with when going to put his name on the babies birth cert?**
Stacey	Q1	Who climbs up the roof to talk Stacey out of jumping off?
Stacey	Q2	**What does Stacey reveal when in a cell with Janine, Kat, and Pat?**
Stacey	Q3	What does Stacey admit to her mum in bedroom when mum is upset?
Stacey	Q4	Who confronts Stacey in restroom where Stacey finally admits to killing Archie?
Stacey	Q5	Who calls to Stacey's door to tell her to get her stuff and go after Stacey's mum had called the police?

Table 4.4: TRECVid 2020 questions used for qualitative evaluation.

4.4 Unsupervised TV Series Summarization: a method based on event classification

Introduction

The entertainment domain, which includes movies and TV series, constitutes a particularly rich collection of videos and a good target for video summarization. It is indeed more and more via streaming platforms that the public discovers new audiovisual content and it becomes interesting for them to be able to display key segments of a program in order to facilitate the user search and browsing experience. While there is a high interest for general-purpose video summarization methods [12], a line of work around genre-specific (video summarization) also exists, as outlined in Sreeja et al. [257]. The authors, for example, underline that if specific actions play a major role for sport videos, the presence of the main characters in a video segment might instead be important for movies. Our method then proposes to leverage on the specifics of the entertainment domain. Namely, we exploit the fact that TV series episodes are often associated to transcripts and/or screenplays. The complex narrative of this type of material is an interesting case study from a computational linguistics point of view, and we argue that their summarization can benefit from the progress made in natural language

processing in the last years. For text summarization as well, many approaches leverage domain-specific features [139]. For example, the best approaches aiming at summarizing news articles are based on the observation that the main points of an article are presented at the beginning of the document [300]. Similarly, summarizing scientific articles is best done when taking into account the very specific structure of this genre of document [6].

In this section, we tackle the task of TV series summarization which aims to produce shorter summaries covering the episodes' most interesting scenes, by proposing a text-based unsupervised method, using screenplays or transcripts previously segmented into scenes or shots. We test our approach on two different genres: crime (from the *CSI: Crime Scene Investigation* [83, 200]) and soap opera (from *BBC EastEnders)*. We show that it is possible to rely on a very general unsupervised model (Zero-Shot text classification), using the right label instead of focusing on the architecture of the model.At first, the usage of text classification may seem counter intuitive for summarization as in many settings, we do not know the semantic content of a text beforehand. However, because some themes, events and words often appear together, there is a long tradition of classifying movie and series into genres [26]. We follow Ben-Ahmed et al. [29] in their hypothesis that the most interesting moments of a series episode should be semantically close to its genre or to events recurrent in the considered genre.

Our work also leverages on the fact that large language models boosted the performance of Zero-Shot classification which is the task of classifying textual inputs using only the label information without seeing any training examples of that label [295]. Therefore, we consider zero-shot classification models to be a good opportunity to test our hypothesis about the importance of genre with an unsupervised model that can easily be used for other genres in the future. To the best of our knowledge, this method has not been yet explored for the task of TV series summarization. Screenplays containing mixed information (dialogues and scenic information describing what the spectator sees and hears), we ask ourselves what is the most relevant text type for a text classification approach based on genre? This work also aims at answering the following questions: Can TV series summarization benefit from zero-shot classification methods? How can we find classification relevant labels for the task of summarization? Are different zero-shot models yielding different results? When coupled with existing approaches, does this method provide complementary information?

Our main contribution consists in showing that with the right label, it is possible to obtain results on par with other state of the art approaches, at the task of unsupervised TV series summarization, with an 'out-of-the-box' tool. We show how to find that label and observe that our method yields even better results when ensembled with centrality measurements developed by Papalampidi et al. [200]. Because we test our general approach on two different genres and datasets with complementary evaluation methods, the specifics of our methods vary with the dataset. The remainder of the section is therefore structured as follows: we first

present how the work presented in this section fits with the related work presented in the Related Work Chapter 2 (Section 4.4). In Section 4.4, we present our general approach. In Section 4.4, we detail our experiments and discuss the results on the CSI dataset, while we present our experiments on the BBC EastEnders dataset in Section 4.4. Finally, we conclude and outline some future work in Section 4.4.

Context

We presented the literature review in video summarization in Section 2.3. Here, we briefly, remind the reader of the works which are the most related to our approach and explain how they related to our approach. We also include some extra references, which do not directly belong to the field of video summarization but which are related to narrative summarization.

One inspiring line of work from general-purpose video summarization, for our particular use case, is multimodal and semantic/category-driven methods [146, 305] which aim to increase the similarity between the semantics of the summary and of the associated metadata, action or video category with a reward system. Instead of video categories, our approach aims to create summaries semantically close to some named events. Concerning the movie/ TV series domain, some early approaches attempted to generate movie trailers, relying on a combination of multimodal low-level features such as motion, contrast, statistical rhythm [296] spatio-temporal saliency, AM-FM speech and part of speech tagging [77], with the goal to draw a multi-modal saliency curve. The MediaEval benchmarking initiative [69], for which higher-level features were also used, helped fostering the research in the field. Interestingness in movies has been approached with the help of related concepts such as movie genre classification [29, 106] or emotional resonance [99, 294]. However, in this paper instead of extracting salient moments (like in movie trailers), we wish to build summaries which cover the whole narrative arc with its major events. For this task, some considered important characters identification [37, 236], while our proposed approach is rather event-centric. Papalampidi et al. [201] took upon the challenge of formalising narrative structure. Based on expert knowledge on narratives, they consider that movie scripts contain five turning points (Opportunity, Change of Plans, Point of no Return, Major Setback and Climax) and show that its is feasible to automatically identify them from screenplays.[8]. The authors also release the so-called TRIPOD dataset that contains movie screenplays and Turning Points annotations. On a follow-up work, they propose a sparse movie graph which indicates the similarity between scenes using multimodal information [202], while Lee et al. [145] identified these turning points with a supervised transformer approach. As we want to develop a method which can be applied to TV series episodes that are not self-contained, and therefore do not necessarily follow such a defined structure, we do not use the five turning points identification as a proxy

[8]The authors report a 17.33% Partial Agreement score on the percentage of turning points where there is an overlap of at least one scene between the prediction and the ground truth

task for TV series summarization. Instead, closer to our work is Papalampidi et al. [200] which demonstrated that these turning points can also be used as a latent representation when gold standard TV series summaries are available. We compare our unsupervised approach to theirs, using the same metrics. Another important contribution in the field, is the TRECVID VSUM challenge for which participants had to develop unsupervised methods to summarize episodes of the BBC Eastenders TV series. For this challenge, one team proposed a purely visual approach [143]. With the exception of our 2021 submission, the other teams all based their methods on fan-written text [100, 216, 276]. On the contrary, our current approach is not dependant on the availability of such external data. Addressing the summarization task from a slightly different perspective, some works generated text summaries from movies or TV series (abstractive summarization) [10, 298]. Finally, despite being mostly interested in user-generated content on social media and review sites, another line of work related to our task is spoiler detection [45, 119]. Such work include a model [119] based on the writing style of the online comments (tense, degree of objectivity) and on named entity recognition. Closest to our work is a a deep neural spoiler detection model with a genre-aware attention mechanism approach [44]. The authors also conducted a spoiler characteristics analysis where they extracted semantic frames from spoiler sentences in the dataset. They found frames associated with "*killing*" to be frequent in thriller spoilers, while romance had more frames linked to personal relationships. We directly use these results to define our text classification candidate labels.

We use the Entail and ZeSTE models throughout the section to test our hypothesis of the potential of zero-shot text classification for TV content summarization. While previous works have used zero-shot approaches for abstractive summarization [168], to the best of our knowledge, extractive narrative summarization for audiovisual data through zero-shot classification is a novel direction of work.

Approach

In this section we present our zero-shot classification method. Screenplays contain mixed information: dialogues and scenic information describing what is visually happening. Dialogues are a transcription of the speech and scenic information are visual instruction in the screenplay indicating the movement, position, or tone of an actor, or the sound effects and lighting. For the CSI series, thanks to an homogeneous formatting across the episodes screenplays, we were able to write a script which separates these two types of texts. Our main motivation for this step, is that while dialogue can be automatically obtained with Automatic Speech Recognition techniques, scenic information cannot. Getting insights into which type of data is the most relevant to the task, allows to somehow assess how automatic the methods is, how it would perform if we would only have access to the raw video. We ultimately use three types of text inputs: dialogue only, scenic information only and original screenplay (mixed

information). For each text input and every scene, our approach consists in obtaining a score denoting the probability that it belongs to the candidate label of interest. We then select the scenes with the highest confidence as the ones that we predict to be part of the summary.

Candidate Labels

One of our hypothesis being that the scenes included in a summary are representative of a TV series or movie genre, we select different ways to choose candidate labels related to a genre.

Genre-based method The candidate label(s) chosen corresponds to the name of the series genre(s).

Event-based method Beyond the genre name, the idea of this method is to obtain candidate labels that are representative of events often happening in a specific genre. As mentioned in Section 4.4, Chang et al. [44] conducted an analysis that provides genre specific words for the Romance and Thriller genres in order to develop supervised genre-aware spoiler detection models. More precisely, they use Framenet [22], a tool built on the semantic frame theory, for sentences semantic role labeling where sentences are parsed and associated to semantic frames according to their structure. Semantic frames are descriptions of a type of event, relation, or entity and the participants in it. For the sentence 'John drowned Martha', it would for example tag 'John' as 'killer', 'drowned' as 'killing' and 'Martha' as victim. The authors used the SEMAFOR parser to extract semantic frames from spoiler sentences for different genres including Thrillers and computed their normalised frame frequency (NFF = count of each frame divided by the total number of frames). Figure 4.3 shows the difference of NFFs for each frame and shows the most contrastive 10 frames for the two genres thriller and romance.

For our approach, as we are interested in making summaries that capture the key events of a narrative, we select as candidate labels the frames names describing an event, among the 10 frames displayed. Hence, for the genre "thriller", we select the labels "killing", "death" and "attempt". The authors interpret the contrast in the distribution of the frames as a significant relationship between the genre and contents of a spoiler sentence. As ultimately the key scenes we want to extract could probably qualify as spoilers, these results also give more empirical grounding to our hypothesis that genre could be used for summary scenes retrieval.

Models

To tackle the task of key narrative event extraction, we choose two state-of-the-art approaches for Zero-shot text classification that use two different sources of knowledge: latent knowledge

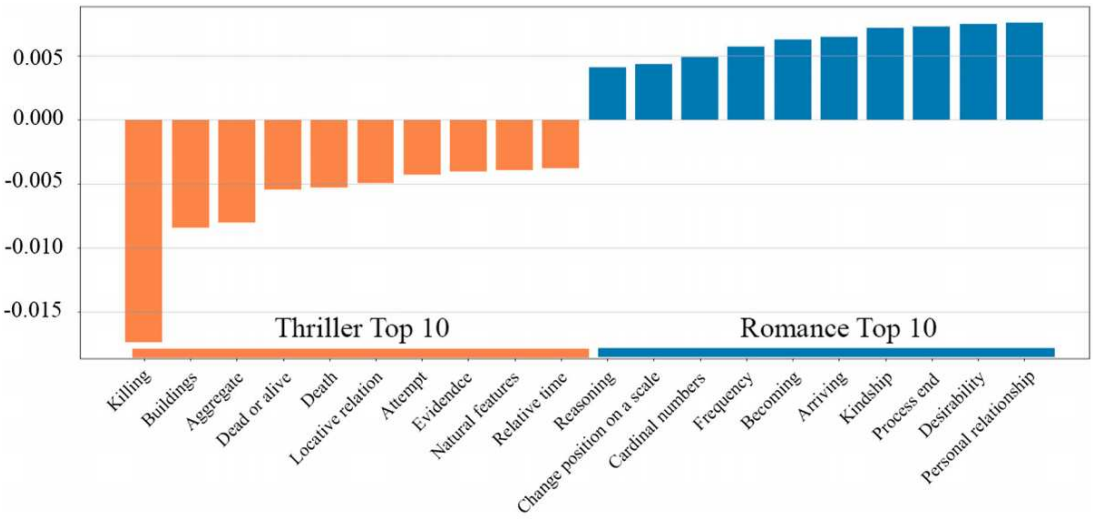


Figure 4.3: Top 10 differentially expressed frames between Thriller and Romance (NFF of Thriller frames - NFF of Romance frames) from violet Chang et al. [44]

from pretraining on big textual datasets through transformers (Entail), and another that uses explicit knowledge about genres through ConceptNet (ZeSTE). Both approaches perform well on several text classification benchmarks, and are freely available and open-source. Because the goal of our section is to illustrate the potential of zero-shot classification for narrative summarization, we forgo the investigation and comparison of more models to make our analysis more focused and concise. We consider this investigation as the future work that follows the promising empirical results of our work.

Entail Given a sentence acting as a *premise*, the task of Natural Language Inference (NLI) aims at determining its relation with an *hypothesis* sentence as either true (entailment), false (contradiction), or undetermined (neutral). NLI datasets consist of sequence-pairs that are generally approached by a transformer architecture such as BERT [71]. Both the premise and the hypothesis are the inputs of a model which classification head predicts one of the following labels: contradiction, neutral, entailment. The method developed by Yin et al. [295] consists in using a model pre-trained on that task as zero-shot text classifier. More precisely, the text to be labeled is the *premise* and the candidate labels are injected in the sentence "This text is about" + label, to form an *hypothesis*.

The confidence with which the Entail model predicts the hypothesis to be entailed by the premise is interpreted as the confidence of the label to be true. While, in the original section, the label-weighted F1 obtained was 37.9% on Yahoo Answers with the smallest version of BERT, fine-tuned on the multi-genre NLI (MNLI) corpus [289], we use the HuggingFace imple-

mentation[9] which reports a F1 of 53.7% by using the Bart model pre-trained on MNLI [148].

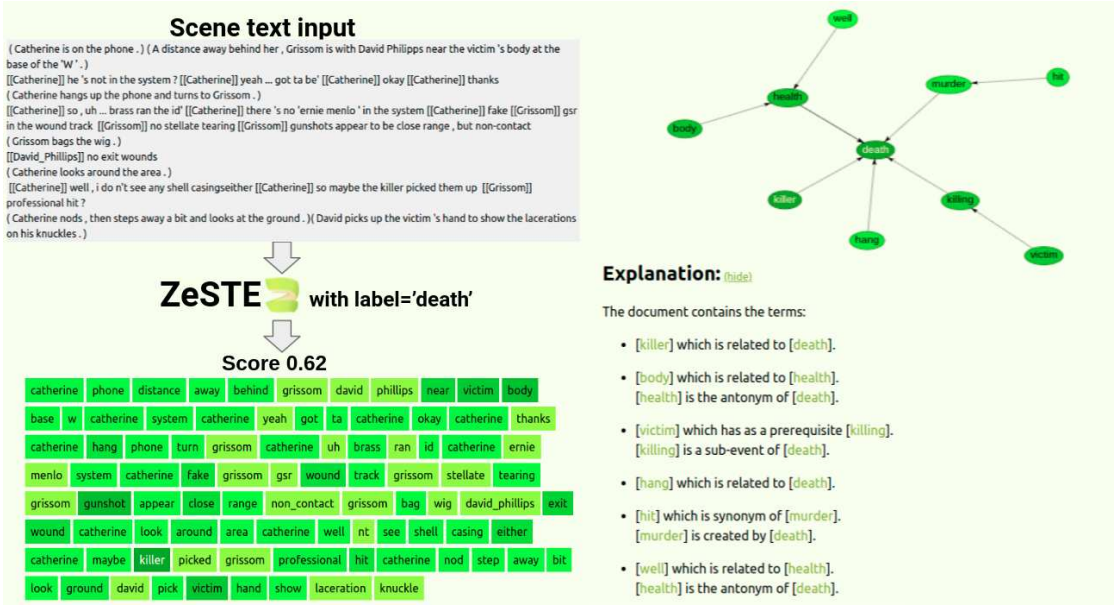


Figure 4.4: Text and explanation of a scene classified by ZeSTE as 'death' as the label with the highest confidence

ZeSTE ZeSTE is a different approach based on the assumption that a document about a topic such as "crime" will probably also mention other words from the same lexical field such as "victim" or "perpetrator". ConceptNet is used to produce a "topic neighbourhood", which is a list of candidate words related to the candidate labels. More specifically, the topic neighborhood is created by querying every node that is N hops away from the label node. Using ConceptNet Numberbatch (ConceptNet's graph embeddings), a cosine similarity score is computed between each node and the candidate label. This similarity score (4.4) represents the relevance of any term in the neighborhood to the candidate label. Next, the documents to classify are assigned a score following the same method (the Numberbatch concept embedding for word w is denoted by nb_w).

$$numberbatch\_score(doc, label) = \sum_{token \in doc \cap LN(label)} sim(nb_{token}, nb_{label})$$

The original authors tried different ways to choose a neighborhood. Following their evaluation procedure on the BBC News dataset, we select their best configuration: 3-hops neighborhoods (for the best performance/computation power ratio), using all the relations (47 relations defined in ConceptNet). Finally, as shown in the example in Figure 4.4, all predicted document labels can be explained by the model by showing the path between the nodes. The darker the colour of the child node, the highest the similarity with the parent node.

[9] https://huggingface.co/zero-shot/

Evaluation

We first evaluate our results with F1 scores, a metric which 'has been adopted by the vast majority of the state-of-the-art works on video summarization' [12], including SUMMER, the method we compare ourselves to on the CSI dataset. Besides this metric, several works also include a human-evaluation [200, 201]. Following Awad et al. [16] who investigated different aspects of human evaluation for the task of video summarization, for the soap opera genre, we assess temporality, contextuality, redundancy and the number of questions a method allowed to answer. For the genre of crime we also investigate the number of crime aspects covered by our method.

Summarizing Crime TV Series Episodes

In this section, we evaluate our genre-based summarization approach, on the CSI dataset [83, 200], which is, according to the authors, associated to the crime genre and which was thoroughly described in Section 4.1[10].

Experiment

We perform the text classification on every scene. In order to compare ourselves with the original SUMMER approach [200], we configure our model to include 30 percents of the scenes in the episode summaries. Applying the genre-based method, the candidate labels are "thriller" and its sub-genre"crime" (as described in the dataset). For the event-bases method, the candidate labels are "killing", "death" and "attempt" (see Section 4.4).

To assess whether our approach yields complementary results to the SUMMER ones (obtained on the mixed information, not separating dialogues from scenic information), we also combined our results. As explained in Section 4.4, SUMMER is an approach that computes centrality measures between scenes to identify turning points and that chooses the scenes with top centrality measures. After min-max scaling these scores, we average them with our ZSC scores (4.4).

$$ensemble\_score(scene) = \sum_{scene} 0.5 \times ZSCscore_{scene} + 0.5 \times SUMMERscore_{scene}$$

Results and Discussion

Table 4.5 presents the results of our experiments on the CSI dataset, where SUMMER corresponds to the state of the art results on this dataset.

[10] The experiments presented in this section can be reproduced using the code published at https://github.com/alisonreboud/screenplay_summarization

Table 4.5: F1 for different text inputs (ZSC = Zero-Shot Classification, SI = Scenic Information, MI= Mixed Information)

		ZSC			ZSC+SUMMER		
		Dialogue	SI	MI	Dialogue	SI	MI
Genre-based method							
crime	Entail	37.32	39.13	38.01	38.75	42.074	41.09
thriller	Entail	39.53	35.91	36.76	40.00	40.84	38.24
crime	ZeSTE	37.44	36.61	40.98	44.14	45.20	44.11
thriller	ZeSTE	36.98	40.52	41.20	45.36	45.08	45.013
Event-based method							
killing	Entail	41.53	**45.49**	41.03	46.34	**48.55**	45.089
death	Entail	40.92	44.77	40.80	45.30	48.97	47.013
attempt	Entail	26.71	32.69	25.45	33.28	40.52	30.89
killing	ZeSTE	40.14	39.17	43.66	46.43	45.14	47.95
death	ZeSTE	43.67	43.25	**46.21**	47.74	46.28	**48.59**
attempt	ZeSTE	37.22	36.95	38.49	43.72	43.44	44.19
	SUMMER			44.70			

First, comparing the results obtained for the genre-based method to the results obtained for event-based method, we observe that for both the Entail and the ZeSTE models, the results obtained with the genre-based method are inferior, suggesting that the name of the genre is not the best candidate label for the summarization via text classification. The F1 scores of the genre-based method reaches a maximum of 41.21% which is under the SUMMER performance. When combined with SUMMER results, the results outperform SUMMER alone in four out of six cases for the ZeSTE model. For the genre-based method, ZeSTE slightly outperforms the Entail model.

For the event-based method, our approach yields the highest mean F1 with the label “killing” using Scenic Information and the Entail model (F1 = 45.49%) and for the label “death” with mixed information and the ZeSTE model (F1 = 46.21%). These labels are semantically close to each other and are the two most representative of the event frames of the genre “Thriller”. On the other hand, the label “attempt” performs the worst of all keywords, across methods (a) and (b). This could probably be explained by the fact that “attempt” is the least domain-specific word among the labels we tried. In a CSI episode context, the word is probably to be understood as “murder attempt”, but the two general zero-shot classification models we use miss the information that our interest only lies in this specific context.

The fact that the words “killing” and “death” are successful labels for crime cases summarization makes intuitive sense from a human point of view. Indeed, this type of crime cases we try to summarize has also be called ’Whodunit’[11] where the word “it” stands precisely for a killing or murder. For these two labels, it is also always the case, for mixed information input types and models, that the combination of our approach and SUMMER always obtains a higher F1

[11] https://en.wikipedia.org/wiki/Whodunit

mean than SUMMER and zero shot classification alone, reaching a F1 score up to 48.59%. In order to assess the statistical significance of our results, we perform a t-test (1) between the F1 scores obtained by SUMMER and the F1 scores of our best approach (ZeSTE with label 'death') (2) between the F1 scores obtained by SUMMER and the ZeSTE (label 'death')+ SUMMER approach. Our null hypothesis is that the two distributions are identical. We respectively obtain p-values of 0.626 and 0.098, which are both above a significance level of 0.05. For such a significance level, these results do not allow us to reject the null hypothesis and we therefore consider our approaches to be on par with the state-of-the-art.

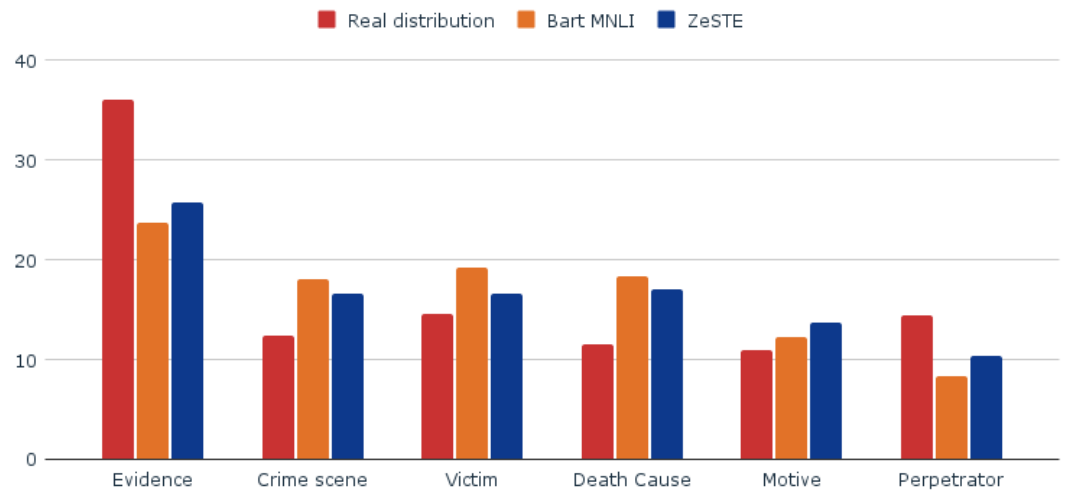


Figure 4.5: Average composition of the scenes correctly predicted as being part of the CSI summary by the best performing Entail and ZeSTE models

In terms of models, there is no clear winner between ZeSTE and Entail. However, they do present differences in terms of the text input it deals with the best. We observe that for Entail, scenic information systematically outperform the other text types with mixed information performing the worst. For ZeSTE, instead, mixed information always yields the best results. Despite not having a direct intuitive explanation for that, the results do suggest that the role played by visual (scenic descriptions) and audio (dialogue) information is an axis worth investigating and that there is a point in isolating texts of different nature, describing two different modalities. We observe that both models are quite sensitive to the label choice and the label ranking (in terms of performance) is quite similar for both models.

Since our goal is to produce informative summaries and given that the SUMMER dataset creators gave some cues about what they consider to be a good summary for this genre – a summary that covers different crime-related aspects which they define to be Evidence, Crime scene, Victim, Death Cause, Motive, Perpetrator of an episode – we compare in Figure 4.5 the distribution of aspects for the scenes chosen by our method with the true distribution

of the dataset. We choose to plot the best performing labels for Entail and ZeSTE, which are respectively “killing” and “death”. First, we observe that the distribution of aspects obtained for ZeSTE and Entail are quite similar. In the ground truth summaries (real distribution), the aspect ’Evidence’ is twice more represented than the other aspects. While ’Evidence’ is also the most frequent aspect in the two models predictions, the frequency of aspects is more evenly distributed with the other aspects. This shows that the summaries created with the approach presented are diverse, covering different aspects of crime plots.

Finally, a small exploration of the scenes wrongly included in the summaries by our method revealed some examples where the error does actually not come from the classification itself: we observe that the scene which was included is indeed strongly associated to the label from a human point of view. Figure 4.4 illustrates such a case. This particular example is an autopsy scene that ZeSTE (rightly) associates strongly to the keyword ’death’ because it contains among others, the words ’body’, ’victim’ and ’killer’ which all are in the ConceptNet neighborhood of ’death’ via the relations mentioned in Figure 4.4. This association to the label is however not sufficient to make the scene relevant enough to be included in the summary. Some more examples of errors can be found in the following section 4.4.

Additional examples

The following section complements Figure 4.4, in presenting the scenes that were wrongly labeled by ZeSTE for season 4 episode 22, when using the label ’death’. Highlighted, are the words that were identified by ZeSTE as being the more closely related to the label [12]

Examples of CSI false positives

Scene 4

(ROBBINS examines ERNIE MENLO’S feet as WARRICK watches.)

Robbins: for what it’s worth, these bruises correspond to the holes in his sock

(ROBBINS walks around the body toward the head-side of the table.)

Warrick: well, he’s been worked over pretty good. He’s got a nice fat lip. Robbins: yeah. there was a good clot in the wound, and the tissues were contused. I’d say it occurred at least an hour or two before death. I teased out a couple of small-caliber projectiles from his brain.

(He holds out the bullets.) (He gives them to WARRICK.)

Robbins: one was embedded in the right frontal cortex. The other lodged in the first cervical vertebra’ Warrick: it’s copper-washed lead. must be a .22.Robbins: you know, historically, .22s were the hit man’s bullet of choice

(Quick CGI POV: The gun shot sounds and the bullet hits the brain and swishes around inside.)

Robbins: they have the energy to enter into the cranial vault, but not enough to exit, so they just ricochet around

[12] https://zeste.tools.eurecom.fr/

inside, shredding the gray matter until they stop.
(End of CGI POV.)(Resume to present.)
Warrick: nice.Robbins: there's also extensive crush injury to both hands, with fractures of the metacarpals and the phalanges
(ROBBINS picks up the body's wrist to show WARRICK the knuckles)
Robbins: bruises appear perimortem. Warrick: any idea what might have caused that kind of damage?
(ROBBINS indicates the x-rays up on the view box behind WARRICK who turns around to look at them.)
Robbins given the fracture pattern, i'd guess it was some sort of blunt object. Warrick: maybe a ball peen hammer. Robbins: what gets you to that? Warrick: they used to tell me back in the days, the first time you got caught cheating, they'd give you a couple whacks on the hand with a ball peen hammer. Robbins: ow Warrick: the second time, you'd lose a limb. Robbins: third time? Warrick: a long walk in the desert with a shovel

Scene 6

(BRASS and GRISSOM interview SAM BRAUN with his LAWYER next to him.)
Lawyer: as much as my client appreciates your flair for the dramatic, the show's over, gentlemen. what do you have? Grissom: the tire patterns at the scene of teddy keller's murder are consistent with the wheel base and turning radius of your client's limousine Lawyer: as well as every other limo in vegas. we also found neon glass embedded in all four tires Lawyer: the whole town's a construction site Lawyer: it's a tenuous link, at best. well, then ... how did his blood end up in the back of your client's limousine?
(THE LAWYER doesn't say anything.)
Brass: you waited until teddy cleared the security cameras
(Quick flashback to: [BACKSEAT OF LIMO] The door opens and SAM pulls TEDDY KELLER into the back seat with him.) (The door shuts behind him.) (SAM back hands TEDDY KELLER in the face causing his nose to bleed.)(A large splotch of blood falls to the seat.)
Sam Braun: we're not through talking, kid
(End of flashback.) (Resume to present.)
Brass: and then you took him for a ride... vegas style. Just like the old days, huh
(Quick flashback to: [NEON GRAVEYARD - NIGHT)(They pull TEDDY KELLER out of the limo's backseat.)
Teddy Keller: please, please, let me go
(The limo door shuts and they lead TEDDY along the neon graveyard.)
Teddy Keller: please. please
(TEDDY'S shirt is opened and the fat suit he's wearing is revealed.)
Sam Braun: let me show you what i do to cheaters Teddy Keller: no, no !
(They pull TEDDY over to the sign and he shoots him twice in the back of the head.) (End of flashback.) (Resume to present.) (SAM leans over and whispers something to his LAWYER.) (When done, the LAWYER turns, looks, and smiles at BRASS and GRISSOM.)
Lawyer: my client offered the young man a ride home Lawyer: they stopped briefly at the neon graveyard, where they held a private conversation regarding the ethics of defrauding a casino
(BRASS chuckles.)

Brass: that must have been some chat Brass: we know he left the casino with the money Lawyer: the young man returned the money as a sign of respect for my client and his position in the community. Brass: i'm sure he did (GRISSOM and SAM stare at each other.)
Brass: so, what next? you gonna tell me you're being set up? it happens to you a lot, huh, sam?

Scene 9

(SARA, WARRICK and DAVID PHILLIP work on the victim's body.) (Camera view down on ERNIE MENLO'S body on the autopsy table.) (He's still in the fat suit.) (SARA picks up something off of his forehead and puts it in a clean envelope.)(DAVID PHILLIPS puts the victim's clothes in a package.) (WARRICK works on the victim's lacerated hand.) (SARA removes the rolex watch.) (She looks at it.)(It's 9:41 am.) Sara: no ticks.it's authentic (She flips the watch over and looks at the back.) Sara: logo sticker is n't worn down. watch could be new. Warrick: guy hits the jackpot, has to celebrate. goes and buys some bling-bling to impress the strippers with (DAVID lifts up the body's foot and sees the holes in the sock's heels.) Warrick: what have you got?
(WARRICK and SARA both look at the feet.)
Warrick: air conditioned socks'

Examples of CSI false negatives

Scene 12

(Camera swoops down to show ERNIE MENLO'S dead body at the base of the'W' in the WHISKEYTOWN letter sign.) (BRASS, GRISSOM and CATHERINE stand around the body.)
Brass: two shots to the back of the head Brass: double tap
(GRISSOM shines his flashlight on the wound at the back of the victim's neck.)
Grissom: he's wearing a wig and a fat suit. it's not halloween, is it? Catherine: in this town, it's always halloween
(BRASS picks up the NEVADA DRIVER'S LICENSE.) (It reads:)
Brass:'ernie menlo' Brass: well, he was n't carrying a very'fat' wad Catherine: rolex is still on his wrist Catherine: probably rules out robbery Catherine: what do you think? Grissom: i do n't know
(GRISSOM turns around and looks at the various signs abandoned and thrown away littering the area.)
Grissom: i'm looking for a sign

Scene 16

(DAVID HODGES explains the composition of the glass as GRISSOM looks through the scope at the shards.)
David Hodges: the glass fragments you found at the apartment building are primarily lead-based. Different curvatures and textures with traces of florescent powder, phosphorous and mercury. Grissom: neon glass David Hodges: i checked out that graveyard once. David Hodges: pretty interesting. Grissom: the comparison? David Hodges: your sample's consistent with the glass collected from the first crime scene. Grissom: see? that connects the two murders. we've got a timeline.

Scene 32

(Sirens wail in the distance.) (ERNIE MENLO sits in the chair in the center of the darkened room.) (In front of him

stands two men - one holding a bright light on him, the other interrogates him.)
Interrogator: I'm going to make this really simple. who are you working with? Ernie Menlo: i'm, uh, unemployed at the moment. Interrogator: you got any idea what we did to chumps like you back in the day? Ernie Menlo: uh, no. look, could you put the a.c. on in here or somethin'? that, or just, uh, let me go. i mean, you ca n't keep me in here. it's against the law Interrogator: there's no law in this room
(He looks at both his interrogators.)
Ernie Menlo: you can't touch me

Analysis

Analyzing the scenes which were wrongly selected for the summary of episode 22 season 4, with our method, we see that we have the same type of error as observed in Figure 4.4: the error does not come from the classification itself. Namely, scenes 4 and 9 are autopsy scenes containing words such as 'body', 'autopsy', 'victim', 'bullet or 'brain'. Similarly, scene 6 is an interrogation which gives information about a murder. The scenes are, hence, strongly associated to the label 'death'. However, the relation with the label, is not here a sufficient condition for the scene to be selected in the summary: autopsy or interrogations scenes seem to be quite common in the CSI episodes but a summary should only include the most relevant scenes for the plot. Analyzing the three summary scenes which were not retrieved by our method, we can see that only scene 32, is not semantically close to the label 'death'. It is worth noting that this particular scene was retrieved by the SUMMER method and as well as by the method which averages the ZeSTE and SUMMER scores (the other two scenes were not). While we can assume that a reason to have included scene 16 in the summary might be its last utterance ('see? that connects the two murders. we 've got a timeline'), which reveals that the plot is about two connected murders, we find it generally difficult, when considering the scenes independently, to justify why these scenes are more relevant than the false positive ones. This speaks in favour of an approach, such as SUMMER, which rather put scenes in a more global perspective, computing a centrality score for each scene. It might also partly explain why, on average we obtain better results when averaging our event classification scores with SUMMER scores.

Summarizing Soap Opera TV Series Episodes

In this section, we further evaluate the robustness of our approach by testing it on an different genre, a soap opera TV series, while adapting the evaluation method. In this section, we present the results obtained for the summarization of the BBC EastEnders series with a human evaluation on the criteria of tempo, contextuality, redundancy and the model's capacity to answer a set of questions about the plot. The dataset was described in Section 4.1. The experiments presented in this section can be reproduced using the code published at https://github.com/MeMAD-project/trecvid-vsum.

Experiment

As the task focuses on some specific characters and does not provide a transcript-shot alignment, we enhance our general approach described in Section 4.4 with additional preprocessing steps that we describe below. Furthermore, as we were only allowed to submit one method for evaluation, we reduced the number of experiments we could do: we select the Entail model, using the dialogue text (the full screenplay of this TV series is not made available by TRECVID) and we focus on the event labels (method (b) in Section 4.4) as our first experiments show better results than just the genre label.

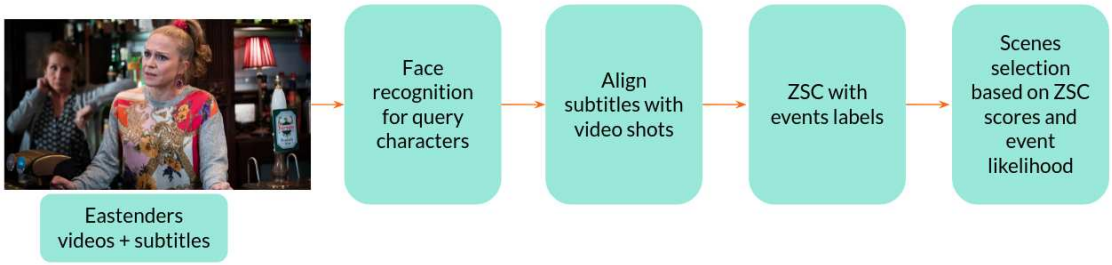


Figure 4.6: Our approach for the VSUM challenge (ZSC = Zero-Shot Classification)

Recognizing Character Faces in Videos The portion of the dataset considered for the challenge contains 10 episodes, that is approximately 19 000 shots, which must be reduced to either 5 or 20 shots (respectively 0.02 or 0.10 percent of the original episodes), while for the previous experiments on the CSI dataset, 30 percents of the scenes were to be selected in the summary. Because of this important compression, we wish to filter out irrelevant scenes and to reduce further the noise. We therefore consider that for a shot to be important for a character, the character needs to be present in this shot. For that, we use the Face Celebrity Recognition library [163], a model that relies on pictures obtained from search engines with the actor's name as the keyword query. In this particular case, we added the word "EastEnders" to the character names to be certain that we do not add pictures corresponding to different people with a similar name. For each image, faces are first detected with an MTCNN algorithm. The system then extracts face embeddings with FaceNet. Based on the hypothesis that most of the faces obtained with the search do depict the actors of interest, other faces (such as people appearing together with the actor) are automatically removed by eliminating outliers to reach a cosine similarity of face embeddings below a standard deviation of 0.24 (this threshold was empirically defined). After this filtering step, a multi-class SVM classifier is trained. To increase the recognition consistency between images, the system also uses the Simple Online and Real time Tracking algorithm (SORT) which returns groups of detection of the same person in consecutive images. In our experiment, we keep the shots with any of the five characters when the confidence score is more than 0.5. Another required preprocessing

step is to transform the given XML transcripts into timestamped text and aligned it with the provided shot segmentation. When a sentence spreads over two shots we include it in both shots as a good summary should probably avoid having scenes with cut utterances. However, this might lead to some noise.

Even if cheaper to produce, we believe that for a better coherence and smoothness of the summary, a segmentation into scenes (like in the CSI dataset) is more appropriate than one into shots. A shot is indeed simply the continuous sequence between two edits or cuts[13], while a scene is the basic unit in a screenplay, usually associated to one main story element [165].

Selecting Events for the Soap Opera Genre Following the semantic parsing method explained in the Section 4.4, we first extract the important events semantic frames for the romance genre. We obtain the following frames: 'process end', 'arriving', 'becoming', 'change position on a scale' and 'reasoning'. All these frames depict quite general events that could refer to many potential sub-events: 'process end' could be the end of a relationship, of a work contract or of an education but also of minor events such as finishing the dinner. For the thriller genre, the event frames extracted ('killing','death' and 'attempt' (murder)) were merely synonyms depicting a single event. The extracted frames for romance rather suggest that this genre encompasses a wider diversity of major events and that we should therefore adapt our method, offering it the flexibility of considering a combination of different event labels rather than a unique one.

Furthermore, as stated in Section 4.4, the task of summarization, even if narrowed down to the specific type of narratives, remains very dependent on the instructions given for the annotation and/or evaluation. For this challenge, it is specifically stated that the model developed for the task should be able to differentiate between meaningful and trivial events, choosing for example 'the birth of a child rather than a short illness'. As the challenge settings does not allow to submit different methods, we anticipate some of the limitations of our semantic parsing approach as a way to extract important events in soap operas and adapt our method to find mode precise soap opera events labels. Specifically, we focus on human knowledge, using the results of a study which aimed to investigate whether soap opera viewers' perceptions of the likelihood of some life events differ from the non-viewers [243]. In this study, the authors select events which they believe are typically happening in soap operas (Table 4.6). Our assumption is that the most likely an event, the least important it is for a summary. For example, we assume that a scene depicting the event 'happily married' is less interesting for a summary than one showing a 'suicide attempt'. We therefore assign to each event a weight equal to the inverse of their perceived likelihood (on a scale from 1 to 5). As we can not assume that the evaluators are especially soap opera viewers, we choose to use

[13] https://en.wikipedia.org/wiki/Shot_(filmmaking)

the likelihood scores given by the non-viewers group. To score each shot, we multiply its confidence score from the zero-shot classifier (which we first normalize for each class using RobustScaler[14]) with the weight of the class (inverse of the perceived likelihood in Table 4.6). Furthermore, to avoid extracting short scenes, and therefore very few information, we further multiply this score with the log of the length of the shot transcript content (Equation r4.4).

$$score(shot_i) = max\_l \in labels(zsc(trans_i, l) * weight(l) * log(len(trans_i))$$

where $shot_i$ is the unique id of the shot, $trans_i$ is its corresponding transcript, $labels$ is the list of events, with their importance expressed with $weight(l)$ for l in labels.

Finally, we select the top N shots for each character based on the max score on all classes as a summary. Because of the constraint on the length of the summary, if the selected shots are too long, we push out the longest scene from the top N and replace it with the N+1th one, and so on until we get a total runtime that fits the summary length requirement.

Table 4.6: Life events labels, their perceived likelihood for non-viewers (scale from 1 to 5 higher is more likely) and their associated weight (inverse of the likelihood) [243]

Label	Likelihood	Weight
extramarital affair	1.98	0.51
get divorced	1.96	0.51
illegitimate child	1.45	0.69
institutionalized for emotional problem	1.43	0.70
happily married	4.05	0.25
serious accident	2.96	0.34
murdered	1.81	0.55
suicide attempt	1.26	0.79
blackmailed	1.86	0.54
unfaithful spouse	2.23	0.45
sexually assaulted	2.60	0.38
abortion	1.41	0.70

Baselines A first comparison we make is with the 2020 edition results presented in Table 4.8. Despite focusing on other characters and episodes, the two editions of the task should have a similar level of difficulty. In 2020, we had proposed a method which relied on data augmentation (MeMAD's method on Table 4.8). As explained in Section 4.4, we had found and scraped one fandom synopsis per episode and we had assumed that each sentence of the synopsis was related to an important moment. After using the same face recognition step as presented in section 4.4, our approach computes the similarity between the synopsis and

[14] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

the episodes transcripts associated to each shot. We defined similarity as being the sum of TF-IDF weights (computed on the transcript) for each word appearing in both the synopsis and the transcript [100]. NII_UIT, the other participant team to the 2020 edition, relied purely on visual features [143]. Their importance score is an average of a face recognition (of the characters of interest) and a representation score. The later is obtained from the sequence to sequence model VASNet [80] which takes as input features extracted from GoogLeNet [268] trained on ImageNet [234].

We also compare our results with the ones obtained by the other challenge participants for the 2021 edition (see Table 4.7). NII_UIT [276], ADAPT [216] both proposed a method relying on the gathering and matching of some external data (respectively Wikipedia articles and fan-written synopsis) about the BBC Eastenders show. More precisely NII_UIT derives an importance score for each shot which is a combination of a face recognition score, a co-appearance face score, a wiki and a virtual event score. The wiki score is obtained from comparing the similarity between the transcripts and the ten most interesting sentences (manually selected) from the characters Wiki page. The virtual event score is a visual one, obtained from an EfficientNet B4 [271] network, which was trained to detect social events in images [1]. The ADAPT approach also relies on a face recognition step and a matching with keywords manually extracted from scraped BBC Eastenders video metadata and fansites. In this table, we also include the results we obtained on a subtask of the challenge, where the evaluation questions are revealed to the participants before submission. These results allow us to assess the gain in performance when the specific important moments are known. For this subtask, we used a longformer pretrained on a QA dataset (Squad-v2) [15]. NII_UIT used the same importance score they computed for their main method, to which they added a question score. This score is obtained after concatenating and embedding the questions (with the Universal Sentences Encoder [42]) to obtain a similarity score. ADAPT relied on the same approach as for the main task, only modifying the list of keywords of interest .

Results and Discussion

Table 4.7 shows the overall results (combining evaluation metrics and characters). for the following constraints:

- Ours_1: 5 shots with highest scores and the total duration of the summary is <150 sec;
- Ours_2: 10 shots with highest scores and the total duration of the summary is < 300 sec;
- Ours_3: 15 shots with highest scores and the total duration of the summary is < 450 sec;
- Ours_4: 20 shots with highest scores and the total duration of the summary is < 600 sec.

Table 4.7: Average score for each run and team (Ours [227], NIIUIT [276], ADAPT [216])

	Team_Run	Main Task	Subtask
in TRECVID VSUM 2021	ADAPT_1	31.20%	15.60%
	ADAPT_2	34.20%	11.40%
	ADAPT_3	27.40%	17%
	ADAPT_4	27.80%	25%
	Ours_1	17.40%	32.20%
	Ours_2	30.40%	31.80%
	Ours_3	32.80%	30.80%
	Ours_4	**37.60%**	34.60%
	NII_UIT_1	7.40%	19.60%
	NII_UIT_2	12.20%	22.40%
	NII_UIT_3	29.60%	28.20%
	NII_UIT_4	22.80%	**49.20%**

Table 4.8: Average score for each run and team (Ours [100] and NIIUIT [143])

	TeamRun	Percentage
in TRECVID VSUM 2020	Ours_1	31%
	Ours_2	31%
	Ours_3	**35**%
	Ours_4	32%
	NIIUIT_1	9%
	NIIUIT_2	8%
	NIIUIT_3	8%
	NIIUIT_4	6%

Our submission with 20 shots reached 37.60%, the best score across the teams. Given that the approach from the ADAPT and NII_UIT teams relied both on the necessary condition of gathering fan-written material specific to the series and on manually selecting important keywords or sentences from them, the results obtained by our approach are encouraging. Indeed, despite not being obtained automatically either, our candidate labels are not specific to the episodes nor to the BBC Eastenders show but could potentially generalise to other series of the soap-opera genre. Our approach is also more minimalistic than the NII_UIT team who combined 5 different scores. We find it interesting that every team relied on different types of text sources (labels, synopsis of the episode, wikipedia of the character) to extract the most interesting shots. However, since all the methods relied on some additional components, which differ between teams (similarity measures, text embeddings, face recognition pipeline...), it is impossible to isolate the text source element and to conclude about their individual relevance. Harmonizing the other components would be an interesting experiment for future work. Another observation we can make is that, contrary to the ADAPT team, the smaller the compression, the better our results. Interestingly, for the substask where queries where known in advance, except from run NII_UIT_4 which obtained 49.20%, no run obtained better results than the best run for the main task. While these results suggest that answering this type of question is still a very challenging task, it might also be that soap opera events are good enough of a proxy for a complete and informed question about the character. Table 4.8 shows the results obtained with the same constraints for the 2020 edition. The fact that our results from the 2020 edition (which ranked first) [100] did not outperform our zero-shot classification method, despite being guided by a fandom synopsis for each episode, might be another cue speaking in favour of genre-events being a good guide to find interesting moments in soap opera episodes.

We present in Table 4.9 all detailed results for the TRECVID VSUM 2021 edition. In particular, the last two rows show the mean for the temporability, contextuality and redundancy metrics. Our results across character and runs are above the mean except for redundancy for which we are slightly under the mean. In general, the results obtained by all teams across all evaluation metrics show that the task remains a challenging one. In Table 4.10, we report the evaluation questions for each character. The questions for which our run 4 (which performed the best and which compression rate is closest to the CSI dataset) was able to answer are marked in bold. We can notice that the majority of the evaluation questions were 'What' questions (16 out of 25), most of them being about events. From these 16 questions, our approach was able to answer 9 of them. On the contrary, our system did not allow to answer any of the other types of questions, namely the 'Who', 'Why', 'Where' questions and one instruction. These results both speak in favour of events/actions as being the first important aspect of a summary but also suggest that our model would benefit from covering other aspects such as

[15] https://huggingface.co/mrm8488/longformer-base-4096-finetuned-squadv2

locations and persons, which might also be genre specific. It is difficult to draw conclusions regarding the type of events that our system was able to capture. However, the questions provide with interesting cues about typical events happening in soap operas. The events we used for candidate labels were all related to love or violence. This is also mainly the case for the events in the evaluation questions. We find a marriage, a kidnapping, a police break, injuries leading to an hospitalization, an attempt murder, etc. These questions cover some events that were not exactly in our candidate list but are nevertheless quite close to them. If we link these results with the ones obtained for the thriller genre were 'killing' was the main event, we can see that despite the fact that romance and thriller are the most distinctive genres according to the semantic parsing experiments [44], in both cases, stories are all about violence and (a little bit of) love. The results are interesting because they point towards a new research question: are love and violent events always the most interesting ones for narratives across genre?

Limitations

In this section, we realised that while the semantic frame extraction step was useful for the experiments on the crime series, it did not generalise well to the soap opera genre. Instead, for that genre, we used a list of events we found in the literature [243]. The fact that our approach is yet missing an automated way to obtain this named events candidates, probably constitutes its current biggest limitation and a direction for future work. Our error analysis also suggested that semantic similarity with the candidate labels is not always a sufficient condition to be included in the summary. As suggested by the increase in performance when averaging the ZSC and SUMMER scores, our method would probably benefit in integrating a step which aims to minimise the redundancy of our summaries, for example by computing a centrality score *a la* SUMMER. Then, in this work, we have worked with two well-defined genres. However, as TV series become more complex, we would need to evaluate our approach on series which genre is not as a clear cut as in the CSI and BBC Eastenders episodes. Finally, in terms of evaluation, more automatic metrics could be investigated in the future. In a classification setting, with a F1 score, a scene is either considered as being important or not. However, it might be worth introducing some nuances. Let's consider two scenes which are not included in the ground truth summary: one, despite not being the one scene chosen in the summary, relate to important event mentioned in the summary, while the other scene is not linked to any important events. An appropriate evaluation should probably account for such a difference. For example, we have seen that the VSUM Trecvid human evaluation, instead of using some specific scenes as ground truth, interrogates whether the selected scenes can answer a specific set of questions. As such, in the future we plan to evaluate our summaries with automatic metrics which assess question-answering capabilities [242]. Following [214], another direction could be to draw inspiration from the text summarization field in which the content similarity between a generated and a ground-truth summary is usually measured. Popular similarity

Table 4.9: All results (T=Tempo, C=Context, R=Redundancy)

Team_Run_Query	T	C	R	Q1	Q2	Q3	Q4	Q5	final_score
ADAPT_1_Archie	5	3	2	Yes	No	Yes	No	Yes	62%
ADAPT_2_Archie	6	5	4	Yes	Yes	Yes	No	Yes	79%
ADAPT_3_Archie	4	6	4	No	Yes	No	No	No	30%
ADAPT_4_Archie	5	5	3	No	Yes	No	No	No	31%
Ours_1_Archie	3	4	5	No	Yes	No	No	No	26%
Ours_2_Archie	3	4	4	Yes	Yes	No	No	Yes	59%
Ours_3_Archie	3	5	5	Yes	Yes	No	No	Yes	59%
Ours_4_Archie	**3**	**5**	**4**	**Yes**	**Yes**	**No**	**No**	**Yes**	**60%**
NII_UIT_1_Archie	3	2	7	No	No	No	No	No	6%
NII_UIT_2_Archie	3	3	5	No	Yes	No	No	No	9%
NII_UIT_3_Archie	4	3	4	No	No	No	Yes	No	27%
NII_UIT_4_Archie	2	2	6	No	No	No	No	No	6%
ADAPT_1_Jack	6	5	2	No	No	No	No	No	17%
ADAPT_2_Jack	6	4	2	No	No	No	No	No	16%
ADAPT_3_Jack	5	5	4	No	No	No	Yes	No	30%
ADAPT_4_Jack	4	5	3	No	No	No	No	No	14%
Ours_1_Jack	6	3	3	No	No	No	No	No	14%
Ours_2_Jack	5	5	4	No	No	No	No	Yes	30%
Ours_3_Jack	4	4	2	No	No	No	No	Yes	30%
Ours_4_Jack	**5**	**4**	**2**	**No**	**No**	**No**	**No**	**Yes**	**31%**
NII_UIT_1_Jack	2	2	5	No	No	No	No	No	7%
NII_UIT_2_Jack	3	2	6	No	No	No	No	No	7%
NII_UIT_3_Jack	4	3	5	No	No	No	Yes	No	26%
NII_UIT_4_Jack	6	4	4	No	No	No	Yes	No	30%
ADAPT_1_Max	3	3	3	No	Yes	No	No	No	27%
ADAPT_2_Max	2	3	5	No	No	No	No	No	8%
ADAPT_3_Max	2	4	4	No	No	No	No	No	8%
ADAPT_4_Max	3	3	4	No	No	No	No	No	10%
Ours_1_Max	4	3	3	No	No	No	No	No	12%
Ours_2_Max	4	3	3	No	No	Yes	No	No	28%
Ours_3_Max	4	3	3	No	Yes	Yes	No	No	44%
Ours_4_Max	**4**	**3**	**4**	**No**	**Yes**	**Yes**	**No**	**No**	**43%**
NII_UIT_1_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_2_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Max	3	3	4	No	Yes	No	No	No	26%
NII_UIT_4_Max	3	3	4	No	Yes	No	No	No	26%
ADAPT_1_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_2_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_3_Peggy	2	3	4	No	No	Yes	No	No	25%
ADAPT_4_Peggy	2	3	3	No	No	Yes	No	Yes	42%
Ours_1_Peggy	3	3	3	No	No	No	No	No	11%
Ours_2_Peggy	3	3	4	No	No	No	No	Yes	10%
Ours_3_Peggy	3	3	5	No	No	No	No	Yes	9%
Ours_4_Peggy	**3**	**3**	**4**	**No**	**No**	**No**	**No**	**Yes**	**10%**
NII_UIT_1_Peggy	2	3	3	No	No	No	No	No	10%
NII_UIT_2_Peggy	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Peggy	3	3	4	No	No	Yes	No	No	26%
NII_UIT_4_Peggy	2	3	4	No	No	No	No	No	9%
ADAPT_1_Tanya	3	2	5	No	Yes	No	No	No	24%
ADAPT_2_Tanya	4	4	5	No	No	No	Yes	Yes	43%
ADAPT_3_Tanya	4	4	4	No	Yes	Yes	No	No	44%
ADAPT_4_Tanya	3	4	5	No	Yes	No	No	Yes	42%
Ours_1_Tanya	4	2	6	Yes	No	No	No	No	24%
Ours_2_Tanya	2	4	5	Yes	No	No	No	No	25%
Ours_3_Tanya	2	2	6	Yes	No	No	No	No	22%
Ours_4_Tanya	**5**	**4**	**5**	**Yes**	**Yes**	**No**	**No**	**No**	**44%**
NII_UIT_1_Tanya	2	1	7	No	No	No	No	No	4%
NII_UIT_2_Tanya	3	3	5	No	Yes	No	No	No	25%
NII_UIT_3_Tanya	4	4	5	No	Yes	Yes	No	No	43%
NII_UIT_4_Tanya	4	4	5	No	Yes	Yes	No	No	43%
Mean	**3.47**	**3.4**	**4.12**						
Ours_mean	**3.65**	**3.5**	**4**						

Table 4.10: Evaluation questions used by assessors in TRECVID VSUM 2021

Archie:
What happens when Phil throws Archie in to a pit?
What happens after Danielle reveals to Archie that Ronnie is her mother?
Where do Peggy and Archie get married?
What happens when Archie arrives at the pub after Peggy invited him?
What happens when Archie is kidnapped?
Jack:
What happens when police break in the door of Jack and Tanya's home?
Where are Max and Jack during the violent confrontation between them when a gun is drawn?
Who does Jack offer to pay in order to withdraw their statement to the police?
Why is Jack a suspect in the hit and run on Max?
What does Jack reveal to Tanya about his dodgy past?
Max:
What were the cause of Max's serious injuries which left him in hospital?
What is/was the relationship between Max and Tanya?
What kind of weapon does Max obtain from Phil?
Where are Max and Jack during the violent confrontation between them when a gun is drawn?
Who is responsible, or who does Max believe is responsible, for the serious injuries which left him in hospital?
Peggy:
Who does Peggy ask to kill Archie?
Where do Peggy and Archie get married?
Show one of the challenges which Peggy faces in her election run.
What does Peggy overhear Archie saying, which causes their marriage to be over?
What is Janine doing to irritate or anger Peggy?
Tanya:
What does Tanya reveal to the police while being interviewed at the station?
What is/was the relationship between Max and Tanya?
What does Jack reveal to Tanya about his dodgy past?
What does Tanya discover in the sink and on Jack's clothes?
What big move were Tanya and Jack planning for the future?

metrics include N-gram based metrics such as ROUGE [160], BLEU [203] or METEOR [142] or neural ones such as BERTscore [302].

Conclusion and Future Work

We have proposed a new method for unsupervised summarization, and we have demonstrated the effectiveness of zero-shot classification with events representative of a genre as candidate labels for crime series and soap operas. When provided with a screenplay, we were able to observe that the Entail model performs best when handling only visual information data. We think our approach is helping to push interpretability: contrary to modelling interestingness without proxies, this approach allows to justify the choice of summary scenes by their closeness to non subjective labels. Another major strength of this approach is its flexibility. Realising that video summarization is subjective, some recent work are interested in producing personalised query based video summaries [112].

In the future, we would like to be able to test how zero-shot classification performs when a user is interested in extracting emotionally interesting scenes or other different concepts related to interestingness. The Entail model is also especially interesting for testing query-based approaches as the pretraining of the model with an 'hypothesis' sentence offer possibilities that go way beyond the sentence we used for classification. The fact that for soap opera and crime, two very different genres, important moments described dramatic events, makes us wonder if an approach based on classification of dramatic events could perform well across genres. While trying to design an approach to find events candidates, we realised that there is a gap in the literature when it comes to classifying events between dramatic and trivial or describing the most common events of a movie genre. In a future work, we plan to close this gap, potentially by relying on human annotation

4.5 Conclusion and Future Work

In this section, we approached the topic of TV series summarization through the angle of multimodality and through unsupervised text-based approaches. In terms of performance, if the supervised approach remains more efficient on the CSI dataset, we showed that a zero-shot classification based method still achieved competitive results. In terms of multimodal supervised approaches, we think visio-linguistic models could benefit from a more diverse pre-training which could include movie captioning datasets [230] and video subtitles, but also in general with datasets where the image-text alignment constraint is relaxed. Then we have shown on the BBC dataset, the potential of an approach guided by named events classification. As we did not have a ground truth for the named-events classification, we could not evaluate the models capacities to correctly categorize scenes. Two observations motivate

us to investigate that further in the next Chapter. First, we saw that it was not necessarily easier to construct a good summary when being provided with the evaluation questions in advance. Second, the method which was guided by episodes synopsis did not outperform the zero-classification one. This came as a surprise and might suggest that the difficulty of the task does not only lies in naming what is interesting but also in matching a description of an interesting moment with the right scene. As such, in the next Chapter we want to evaluate and ameliorate zero-shot models capacities to identify labels related to narrative aspects (such as 'perpetrator', 'crime scene', ... for crime series). Another question we wish to answer is: if Language Models did not allow to answer human produced questions efficiently, what type of questions could they instead answer? To answer this question, we investigate the task of automatic question generation.

Chapter 5

Story Understanding

In the previous Chapter, we developed an approach for unsupervised TV Series Summarization which was based on event classification. In this approach, we had named categories but no or too small annotated resources to train a classifier. We therefore used ENTAIL and ZESTE, two models for zero-shot text categorization. If the results were encouraging, these experiments also allowed us to point out at some limitations of the these two models such as the lack of a domain-adaptation capacities. When we used, the label 'attempt' for example, we realised that we could not specify that this word should be understood within the context of 'crime series' and not within its more global meaning. In this Chapter, we then aim to ameliorate narrative aspects classification, by proposing PROZE a novel method for zero-shot text categorization which allows for domain adaptation. Compared to the previous chapter, we take a step forward towards narrative understanding by classifying text using more fine-grained types of narrative aspects. Because we want to build a system which is not specific to crime aspects, we also show that our model performs well on datasets from very different domains such as silk or emergency situations.

In a second part of this Chapter, we continue our exploration of Language Models capacities when applied to domain-specific datasets, by exploring the task of automatic Question Generation for TV series. In the previous Chapter, we had seen that the VSUM TrecVID challenge proposed to evaluate summarization by assessing a model capacity to answer important questions written by humans. In this Chapter, we investigate whether the writing of the important questions can be automatized.

This section covers the following publication:

1. Ismail Harrando*, Alison Reboud*, Thomas Schleider*, Thibault Ehrhart and Raphael Troncy (*Equal contribution). **ProZe: Explainable and Prompt-guided Zero-Shot Text Classification**. IEEE Internet Computing: Special Issue on Knowledge-Infused Learning, 2022.

2. Reboud, Alison*, Schleider, Thomas*, Troncy, Raphaël
Exploring Automatic Question Generation for In-Domain Text Understanding. (under review).

5.1 Story element extraction through domain adaptation of a zero-shot text classification method

As technology accelerates the generation and communication of textual data, the need to automatically understand this content becomes a necessity. In order to classify text, being it for tagging, indexing or curating documents, one often relies on large, opaque models that are trained on pre-annotated datasets, making the process unexplainable, difficult to scale and ill-adapted for niche domains with scarce data. To tackle these challenges, we propose *ProZe*, a text classification approach that leverages knowledge from two sources: prompting pre-trained language models, as well as querying ConceptNet, a common-sense knowledge base which can be used to add a layer of explainability to the results. We evaluate our approach empirically and we show how this combination not only performs on par with state-of-the-art zero shot classification on several domains, but also offers explainable predictions that can be visualized.

Introduction

The Natural Language Processing (NLP) and Information Extraction (IE) fields have seen many recent breakthroughs, especially since the introduction of Transformer-based approaches and BERT [71], which has become the *de-facto* family of models to tackle most NLP tasks. Over the last years, few-shot and zero-shot learning approaches have gained momentum, particularly for the cases with little data and where uncommon or specialized vocabularies are being used. Fully zero-shot classification approaches do not require any training data and often show respectable performance. An interesting new paradigm is *prompt-based learning* which leverages pre-trained language models through prompts (i.e. input queries that are handcrafted to produce the desirable output) instead of training models on annotated datasets. However, a major downside of all these approaches based on transformer-based language models is that they suffer from a lack of explainability.

One direction of a growing amount of work interested in explainable methods is to generate explanations and to develop evaluations that measure the extent and likelihood that an explanation and its label are associated with each other in the model that generated them [204]. However, none of these techniques totally compensate for the obscurity associated with language models. This is the main reason why the approach presented in this section relies on ZesTE (Zero Shot Topic Extraction) [101], which is not based on a pre-trained language

model, and provides explainability of its classification results using ConceptNet [256] and its explicit relations between words, as a prediction support. With every word being a node in ConceptNet, ZeSTE can justify the relatedness between words in the document to classify its assigned label. While it shows state-of-the-art results in topic categorization, it does not offer ways to specialize the classifier beyond "common sense knowledge" (domain adaptation), nor does it offer the possibility to disambiguate labels.

These challenges are important to solve for text classification of specific domains, especially since zero-shot classification is particularly useful for domain-specific use cases with little data to train a model. As a consequence, this section proposes *ProZe*, a Zero-Shot classification model which combines latent contextual information from pre-trained language models (via prompting) and explicit knowledge from ConceptNet. This method keeps the explainability property of ZeSTE while still offering a step towards label disambiguation and domain adaptation. Previous contributions leverage knowledge graphs [48, 169, 301] and common-sense [192] to improve the performance of several classification tasks. To the best of our knowledge, our approach is the first to use a common-sense knowledge graph to not have a learning component, and uses the KG as is, allowing it to retain explainability.

The remainder of this section is structured as follows. First, we detail our proposed method called ProZE. Next, we present our results on common topic categorization datasets as well as on three challenging datasets from diverse domains: screenplay aspects for a crime TV series [83], historical silk textile descriptions [239], and the Situation Typing dataset [180]. We report and analyze the results of several empirical classification experiments, which includes a comparison to some state-of-the-art Zero-Shot approaches. Finally, we conclude and outline some future work.

Method

Our model can be seen as a pipeline comprising several components. In this section, we explain each step of the process in further details.

ConceptNet A central resource for this work is ConceptNet [256], a semantic network "*designed to help computers understand the meanings of words that people use*"[1]. Broadly speaking, ConceptNet is a graph of words (or *concepts*), connected by edges representing semantic relations that go beyond the lexical relations than can be found in a dictionary such as "Synonym" or "Hypernym". Most importantly, ConceptNet contains relations of general "relatedness" (or `/r/RelatedTo` on ConceptNet), which imply an undefined semantic relation between two concepts, such as "Business" and "Outsourcing": while both terms are

[1] https://conceptnet.io

used in similar contexts, one cannot define such relation as one of containment, usage or typing. It it notable that, unlike semantic similarity between two terms via word embeddings, "relatedness" relations are usually mined for dictionary entries or corresponding Wikipedia articles, thus making them explainable to the user.

Other than the knowledge graph, ConceptNet comes with its set of graph embeddings called "ConceptNet Numberbatch". Computed in a special way to reflect both the connectedness of nodes on the ConceptNet graph and the linguistic properties of words via retrofitting to other pretrained word embeddings [256], these embeddings can better capture semantic relatedness between words, as demonstrated by their performance on the SemEval 2017 challenge (https://alt.qcri.org/semeval2017).

We use both the semantic graph for generating explanations and the Numberbatch embeddings to prune out excessive and noisy relations in our method.

Generating Label Neighborhoods The first step of our approach is to manually create mappings between target class labels and their ConceptNet nodes. For instance, if we want our classifier to recognize documents for the class “sport", we designate the node `/c/en/sport` as our starting node.[2]

Based on these mappings between target labels and concept nodes, we can then generate a list of candidate words (from ConceptNet) that are related to the respective concept. This list can be called the "label neighborhood". Each of the candidate is produced by retrieving every node that is N-hops away from the class label node.

Afterwards, a score can be calculated for each label based on which words are present in the input text or document to classify. To this end, we score every word in the label neighborhood based on its "similarity" to the class label.

Scoring a Document Like ZeSTE, we proceed to score each document by first generating a score for each node in a label neighborhood. To do so, multiple approaches exist. In this section, we present and compare 3 such scoring methods (SM):

1. **ConceptNet embeddings similarity (SM1)**: ConceptNet Numberbatch[3] are graph embeddings computed for ConceptNet nodes. To quantify their similarity, we compute cosine similarity between the embedding of each node on the label neighborhood and the label node itself.

[2] In the remainder of this section, we will omit the prefix `/c/en/` as all labels in our datasets are in English.
[3] https://github.com/commonsense/conceptnet-numberbatch

2. **Scoring through Inference (SM2)**: for this scoring method, we use a model that is pre-trained on the task of Natural Language Inference. In a similar setting to the previous method, we prompt the model with a sentence related to the label or its domain, and then we ask it to score all the words from its neighborhood based on the logical entailment between the prompt (premise) and a template containing the word (hypothesis).

3. **Language Modeling Probability (SM3)**: for this scoring method, we combine the predictive power of language models with the explicit relations that we can find on the label neighborhood. For each label, we supply the language model with a *prompt*, or a sentence that is likely to guide it towards a specific meaning of the label we target (for example, the definition of the label), and then, we ask it to predict the next word in a Cloze statement (a sentence where one word is removed and replaced by a blank). For example, to score words related to the label "sport", we can give the model a definition of the word, and then ask it to predict the blank word in the following Cloze statement: *"Sport is related to [blank]."*. Given that language models, are pre-trained on predicting such blanks, we can use the scores they attribute to that blank to measure the similarity between our label and the candidate words from its neighborhood. For instance, when we give the dictionary definition of sport to the language model, the top predicted words are 'recreation', 'fitness' and 'exercise'. Because the language model outputs a probability for every word in its vocabulary, we score only the words that are originally on the label neighborhood. If a word in the neighborhood does not appear among the predictions of the model (i.e. out of the model's vocabulary), the score from SM1 is used.

Once the scores are computed by one of these methods, we can proceed to score any document given as input to the model. To score such document, we first tokenize it into separate words. We then take all the nodes from the neighborhood of a label that appear in the tokenized document, and we add up their scores to produce a score for the label. We do so for each label we are targeting, and the final prediction of the model corresponds to the label with the highest score. Because all the nodes in the neighborhood are linked to the label node with explicit relations on ConceptNet, we can explain in the end how each word in the document contributed to the score and how it is related to our label.

Prompting Language Models In this section, we explain how we leverage language models to score the label neighbors extracted from ConcepNet, as per the scoring methods SM2 and SM3 described above.

Both SM2 and SM3 methods rely on prompting the language model, i.e. to feed it a sentence that would function as a context to "query" its content (also known as *probing* [58]). As expressed in the related work, prompting language models is an open problem in the literature.

In this work, we explore some potential ideas for prompting to serve our objective of measuring word-label relatedness.

The prompting follows the same scheme for both scoring methods. We vary both the premise and hypothesis templates and report the results for some proposals in the Evaluation section. For the premise, we experiment with two approaches:

1. Domain description: where we prime the model with the name or description of the domain of the datasets, i.e. "Silk Textile", "Crime series", etc.

2. Label definition: where we prime the model with the definition of the label, with the assumption that this will help it disambiguate the meaning of the label and thus come up with better related words. For instance, for the label "space", we provide the language model with the sentence "Space is the expanse that exists beyond Earth and between celestial bodies". We take the definitions from Wikipedia or a dictionary, we generate it using a NLG model etc.

We observed experimentally that using just the description of the domain as a prompts gives better overall performance. Therefore, we only report results on these prompts in the following sections. As for the hypothesis, we provide the model with a sentence like *"[blank] is similar to **space**"* or *"**Space** is about [blank]"* which we use in our reported results.

We note that, while the combination of premise and hypothesis can impact the overall performance of the model, the search space for a good prompt is quite wide. Thus, we only report the performance on some combinations, as we intend this section to only point out the use of such mechanism for this task rather than fully optimize the process.

Tool Demonstrator To explain the decisions of the model, we follow the same method as ZeSTE [101], i.e. we highlight the words which contribute to the decision of the classification as shown in a graph that links them with semantic relations to the label node. The difference is that the scores in ProZe take also into account the scoring from the language model. To illustrate the contribution of the language model, we developed an interactive demonstrator enabling a user to test the effect of prompting the language model to improve the results of zero-shot classification (Figure 5.1). This demonstrator is available at http://proze.tools.eurecom.fr/.

After choosing a label to study, the user is asked to enter a prompt that can help the model to identify words related to the label (e.g. definition or domain). The user is then shown an abridged version of the prompt-enhanced label neighborhood: the connection between any node and the label node is omitted for clarity but it can be trivially retrieved from ConceptNet,

and only the top 50 (based on the used scoring) words are shown to represent the new label neighborhood, with the intensity of the color reflecting higher scores.

The user can view in detail the updates happening before and after introducing the new scoring from the Language Model. For this demonstration, we use the SM3 method to score the nodes as it requires only one pass through the Language Model to generate a score for all words in its vocabulary, whereas the SM2 method requires an inference for every word in the label neighborhood. As a consequence, while the SM2 methods takes up to 7 minutes per label on our hardware, the SM3 method takes less than a second while still delivering good performance.

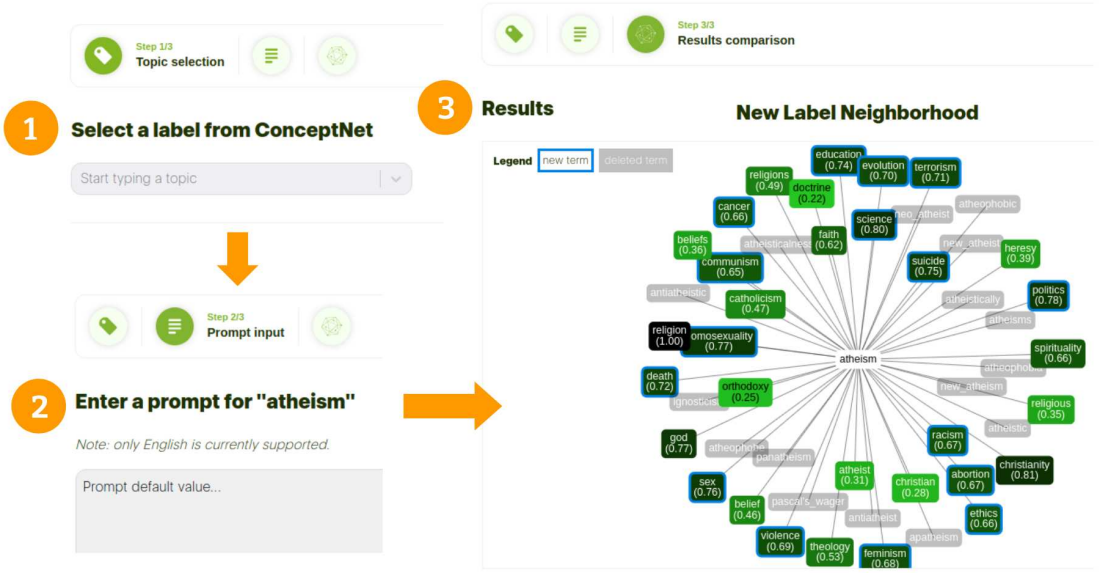


Figure 5.1: ProZe neighborhoods demo. (1) The user is asked to select a label (2) The user can input a text to prompt and guide the language model. (3) The user can visualize the label neighborhood, with added and removed nodes highlighted, and is shown a detailed list of all the changes resulting from the prompt.

Complementary datasets used for domain-adaptation evaluation

In this section, we present three widely used topic categorization datasets in the news domain, as well as three other very different and domain-specific datasets making used of fine-grained labels.

News Topics Datasets Used to benchmark multiple text classification approaches, news datasets are often categorized by topic and are written in simple and common language. In our experiments, we report results on three such commonly-used datasets: AG News, BBC

News and 20NG.

- **20 Newsgroups** [141]: a collection of 18000 user-generated forum posts arranged into 20 groups seen as topics such as *"Baseball", "Space", "Cryptography", and "Middle East"*.
- **AG News** [92]: a news dataset containing 127600 English news articles from various sources. Articles are fairly distributed among 4 categories: *"World", "Sports", "Business" and "Sci/Tech"*.
- **BBC News** [90]: a news dataset from BBC containing 2225 English news articles classified in 5 categories: *"Politics", "Business", "Entertainment", "Sports" and "Tech"*.

Crisis Situations The first low-resource classification dataset we use is the Situation Typing dataset [180]. The goal is to predict the type of need (such as the need for water or medical care) required in a specific situation or to identify issues such as violence. Therefore, this dataset constitutes a real world, high-consequence domain for which explainability is particularly important. The entire dataset contains 5,956 labeled texts and 11 types of situations: "food supply", "infrastructure", "medical assistance", "search/rescue", "shelter", "utilities, energy, or sanitation", "water supply", "evacuation", "regime change", "terrorism", "crime violence" and a "none" category. In our experiment, we use the test set (2343 texts), where we only select texts that represent at least one of the situations and we consider it a success if the model predicts at least one correct label.

Silk Fabric Properties This dataset is an excerpt from the multilingual knowledge graph of the European H2020 SILKNOW research project[4] aiming at improving the understanding, conservation and dissemination of European silk heritage. The SILKNOW knowledge graph consists of metadata about 39,274 unique objects integrated from 19 museums and represented through a CIDOC-CRM-based set of classes and properties. This metadata about silk fabrics contains usually both explicit categorical information, like specific weaving techniques or their production years, but also rich and detailed textual descriptions. Our goal is to try to predict categorical values based on these text descriptions.

The SILKNOW Knowledge Graph dataset can be divided into using "material" and "weaving technique" subsets. More precisely, we slightly extend the dataset used in [240], and after removing objects with more than one value per property, we obtain 1429 object descriptions making use of 7 different labels for silk materials, and 833 object descriptions with 6 unique labels for silk techniques. The chosen labels have also to be mapped to ConceptNet entries to work with this approach. Table 5.1 shows the final selection of thesaurus concepts and their mapping to ConceptNet nodes.

[4] https://silknow.eu/

Property	SILKNOW Concept	ConceptNet
Material	Cotton	`/c/en/cotton`
Material	Wool	`/c/en/wool`
Material	Textile	`/c/en/textile`
Material	Metal thread	`/c/en/metal`
Material	Metal silver thread	`/c/en/silver`
Material	Silver thread	`/c/en/silver`
Material	Gold thread	`/c/en/gold`
Technique	Damask	`/c/en/damask`
Technique	Embroidery	`/c/en/embroidery`
Technique	Velvet	`/c/en/velvet`
Technique	Voided Velvet	`/c/en/velvet`
Technique	Tabby (silk weave)	`/c/en/tabby`
Technique	Muslin	`/c/en/tabby`
Technique	Satin (Fabric)	`/c/en/satin`
Technique	Brocaded	`/c/en/brocaded`

Table 5.1: Mapping between the concepts used in the SILKNOW knowledge graph and ConceptNet (ProZe and ZeSTE)

Evaluation

We evaluate ProZe on these 6 datasets. In this section, we present the results of this evaluation.

Baselines

We compare our model with:

- *ZeSTE*: this approach solely relies on ConceptNet to perform Zero-Shot classification;

- *Entail*: this model was originally proposed in [295]. We use `bart-large-mnli` as the backend Transformer model, which it is a version of **BART** [149] that was been fine-tuned on the Multi-genre Natural Language Inference (MNLI) task, as per the implementation we use for our experiments (can be tested at https://huggingface.co/zero-shot/). Given a text acting as a *premise*, the task of Natural Language Inference (NLI) aims at predicting the relation it holds with an *hypothesis* sentence, labelling it either as false (contradiction), true (entailment), or undetermined (neutral). Generally, the labels are injected in a sentence such as "This text is about" + label, to form an *hypothesis*. The confidence score for the relation between the text to be labelled and the premise to be 'entail' is the confidence of the label to be correct. We use the implementation provided at https://github.com/katanaml/sample-apps/tree/master/01)

Datasets	20 Newsgroup		AG News		BBC News	
	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg
ZeSTE	63.1%	63.0%	**69.9%**	**70.3%**	84.0%	84.6%
Entail	46.0%	43.3%	66.0%	64.4%	71.1%	71.5%
ProZe-A	62.7%	62.8%	68.5%	69.1%	83.2%	83.7%
ProZe-B	**64.6%**	**64.6%**	69.0%	69.6%	**84.2%**	**84.8%**

Table 5.2: Prediction scores for the news datasets (the top score in each metric is emboldened).

Quantitative Analysis

We limit the size of the label neighborhoods to 20k per label for each experiment, except in cases where querying ConceptNet returns less nodes than that. Then, we resize all the other neighborhoods to be all equal in size to the smallest one (by eliminating the nodes with the lowest similarity), as we found that having neighborhoods of different sizes skews the predictions towards the larger ones (by virtue of having more nodes to contribute to the score). This can be circumvented by increasing the number of hops (thus boosting the size of smaller neighborhoods before filtering), but according to our observations, this hurts the quality of the kept nodes as they get less semantically relevant as we hop further. Resizing the neighborhoods eliminate the bias against the in-domain labels that may not have so many related words in the first place.

Table 5.3 and Table 5.2 show a score comparison of the ProZe approaches to the baselines of ZeSTE and the Entail approach. **ProZe-A** refers to scoring the nodes using a combination of SM1 and SM2, whereas **ProZe-B** uses a combination of SM1 and SM3. We tested several ways to combine the scores from ConceptNet (SM1) and language models (SM2 and SM3), including taking the sum of the two scoring methods, their product, their max, or a weighted average. Empirically, we obtain the best empirical results by multiplying the two scores (both normalized to be between 0 and 1). The main advantage of multiplication is that it penalizes disagreement between the language model and the KG over how close two terms are. This also means that the explainability layer reflects accurately the decisions of the model, as words that are not scored well by the language model will not contribute significantly to the classification score.

Table 5.2 contains the accuracy and weighted average scores for the 3 news datasets that consist of general knowledge texts. ProZe has similar performance, but not beating ZeSTE, which is in line with our expectations: both approaches are based on the ConceptNet commonsense knowledge graph, and the vocabulary does not need or cannot be guided into a more fitting direction with the prompts. For all three news datasets, however, ProZe performs better than Entail.

Datasets	Silk Material		Silk Technique		Crime aspects		Crisis situations	
	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg
ZeSTE	34.3%	39.0%	46.9%	47.2%	31.2%	32.3%	46.3%	45.8%
Entail	29.0%	33.3%	**64.0%**	**65.8%**	**43.7%**	**43.7%**	46.7%	48.1%
ProZe-A	**39.0%**	40.1%	50.8%	57.6%	36.3%	37.6%	**50.1%**	49.7%
ProZe-B	37.4%	**41.7%**	48.5%	48.7%	29.8%	31.1%	**50.1%**	**49.8%**

Table 5.3: Prediction scores for the domain-specific datasets (the top score in each metric is emboldened).

Table 5.3 shows the results for the 3 domain-specific datasets. We observe that ProZe is consistently outperforming ZeSTE, which we take as a confirmation that the guidance through the prompt is effective for specific domains. For two datasets, silk material and situations, ProZe even beats the non-explainable baseline scores of the Entail approach. This is not the case for the silk technique and the CSI screenplay datasets as some labels from these datasets have very limited neighborhoods in ConceptNet. Nevertheless, our approach is still close and retains in all cases its higher degree of explainability.

Qualitative Analysis

To illustrate why a re-ranking of related words induced by a domain prompt improves the score, we analyse a concrete example. Taken from the silk technique dataset, the top 10 candidate terms of the ConceptNet label neighborhood for the weaving technique "embroidery" are as follows: "Embroidery, overstitch, running stitch, picot, stumpwork, arresene, couture, fancywork, embroider, berlin work". While these words are clearly related to the concept of embroidery, they are not necessarily relevant in the context of silk textile. For example, "picot" is a dimensional embroidery related to crochet. The intuition is then that this neighborhood can be improved by specifying the domain.

In comparison, the top 10 candidate terms of the pre-trained BART language model, guided by a prompt that included the term "silk textile" are: "Craft artifact sewn, fabric, embroidery stitch, embroidery, detail, embroider, mending, embellishment, elaboration, filoselle". These terms are more general even if also related to silk textile. Words such as "detail", "mending", "elaboration" or "embellishment" seem useful for classifying texts that are not only consisting of details about different types of embroidery. When combining the scores from ConceptNet and the language model, the ProZe method increases its F1 score of circa 8%, from 61% to 69%.

Conclusion and Future Work

In this section, we demonstrated the potential of fusing knowledge about the world from two sources: First, a common-sense knowledge graph (ConceptNet), which explicitly encodes knowledge about words and their meaning. Second, pre-trained language models, which contain a lot of knowledge about language and word usage that is latently encoded into them. We explored several methods to extract this knowledge and leverage it for the use case of zero-shot classification. We also empirically demonstrated the efficiency of such combination on several diverse datasets from different domains.

This work is experimental and does not fully explore all possibilities of this setup. As future work, we want to study the effect of prompt choice in more detail, and seeing how such choice impacts not only the quality of the predictions but also that of the explanations. Different language models can also be tried to measure how such choice can improve the overall classification, especially for specific domains such as e.g. medical documents.

Another potential improvement over this method is to filter out words unrelated to the label using the slot-filling predictions from the language model. From early experiments, this method seems to give good results by restricting the neighborhood nodes to ones that almost exclusively relate to the label in some way.

A natural direction of work is to involve the user in the creation of the label neighborhood (human-in-the-loop) by asking whether some words that only the Language Model and not ConceptNet suggests pertain to the target label. This allows to inject the extracted knowledge from the language model back into the zero-shot classifier, and fill in the gaps of knowledge from ConceptNet.

Finally, some existing limitations of the original work can be still improved upon such as letting the language model inform the label selection and expansion, handling multi-word labels, and integrating more informative concepts from ConceptNet beyond word tokenization (e.g. *'crime_scene'*, *'tear_gaz'*).

5.2 Exploring Automatic Question Generation for Narrative Summarization

Introduction

Several Natural Language Processing (NLP) tasks observe a drop in performance when applied to domain-specific datasets. Recently, Text-to-Text Transfer Transformer (or T5) made a strong impact for tasks like Question Generation, which require the model to identify important text parts without necessarily requiring prior fine-tuning. We investigate such an approach with

two different applications: information extraction on a dataset about European silk fabrics and summarization on a dataset about scripts of a TV series. We evaluate our approach both qualitatively and in case of the silk fabric dataset also quantitatively. We show how transformer-based Question Generation represents a promising supplement for adapting NLP tasks to domain-specific datasets.

How to assess text understanding in the field of Artificial Intelligence (AI) remains an open question. In order to evaluate if a human understands a concept, we often test their capacities to answer questions, but also to produce meaningful questions about the subject matter [53]. Question Generation (QG) [233] has been a relevant task inside the field of Natural Language Processing (NLP) for many years. Just as with human text comprehension, within Artificial Intelligence (AI) a model's ability to ask meaningful questions is considered to be central to evaluate its text comprehension ability [190].

In recent years, nearly all models for Question Generation were deep learning-based, particularly since the emergence of Seq2Seq [263]. Afterwards, a huge breakthrough in the whole field of NLP came with the emergence of Transformer-based models, particularly with the introduction of BERT [71]. Transfer learning is another sub-domain for which transformer-based approaches have been very relevant since years, for example Text-to-Text Transfer Transformer, or T5 [220]. T5 can not only easily be used for Question Generation, the performance of models based on it are also on par with other approaches.

In this section, we investigate if generating meaningful questions out of an input text could possibly imply a good text understanding and if it would be possible to leverage on this for other downstream tasks. We identified two applications with domain-specific texts to which these models could be particularly helpful. The first one is the task of TV series script summarization for which the content needs to be reduced to the most important questions. The summarization use-case is especially relevant since a new line of work around query-based summarization has appeared in response to voices arguing against the traditional evaluation methods.

The second one consists in identifying the most central parts of rich textual descriptions of silk fabrics. Since this particular application is not directly related to the topics of this thesis, the experiments and results for this domain were placed in an Appendix Ato this thesis. In this section, we focus on answering the following question: How can such an approach be used to build and evaluate summaries? The remainder of the section is structured as follows: we first describe other works with investigate the role of question answering for summarization (Section 5.2). We detail our experiments on TV series summarization in Section 5.2. Finally, we conclude and outline some future work in Section 5.2.

Question Answering and Summarization

Traditional metrics for both extractive and abstractive summarization have been criticized for their lack of correlation with human judgements [11, 196]. For abstractive summarization, n-based metrics such as ROUGE fail at accounting for factual consistency and evaluating whether only the most important content is included in a summary [195, 212]. For this reason, alternative evaluation methods based on questions answering have been designed. One of these methods is developed in the context of the TRECVID VSUM evaluation campaign, which aims at producing summaries consisting of important shots of a TV series [16]. They proposed an a posteriori evaluation based on tempo, contextuality and redundancy, as well as with regards to how well they answered a set of 5 questions per character. Surprisingly, unveiling the evaluation questions for the sub-task did not allow the participating teams to improve on the results, which have been obtained without the questions being revealed in the main task [227]. This observation motivates our work, raising further questions related to two hypotheses:

- Questions automatically generated from the source text have a formulation closer to the original text. Answering them could be an easier task, but are such questions relevant for summarization? We aim at getting insights into the type of questions that can be generated and therefore answered and compare them to the experts questions.
- Training a question-answering model for TV series summarization requires the availability of summarization questions. To the best of our knowledge, the dataset created for the VSUM Trecvid challenge (which only contains 15 questions) is unique and is the only one with questions specifically designed to cover summaries aspects of TV series. Question creation being an expensive process, is it possible to automate this part?

Another related work which also underlines the need for domain-specific important questions datasets is [242]. The authors developed a summarization evaluation method which relies on assigning a score to queries with regards to their importance and to how well they were answered in a summary. However, the authors focus on a factoid question answering dataset and underline that their 'query weighter' (defined as a module which distinguishes important questions from anecdotal ones) is domain and application dependant. Consequently, our work takes a step towards improving query-based TV series summarization, by evaluating the capacity of general models to generate relevant questions from different text inputs: transcripts and fan-written synopsis.

Approach

Dataset We choose to work with the subset and questions of the VSUM Trecvid 2020 edition because for those episodes, external data (fan written synopsis) is available. For the 2020

edition of the challenge, the participants had to submit 4 summaries with respectively 5, 10, 15 and 20 automatically selected shots over 10 episodes for three different characters of the series. That year, the best team only managed to find shots that answered 4 out of 15 shots [100]. TV series transcripts indeed offer a number of challenges that differ from the datasets on which the question generation model has been trained: these texts are very with complex stories where the important information is not necessarily at the beginning of the text (contrary to news articles) and they are dialogues which in our case does not include speaker names. Consequently, for a more complete evaluation of the question generation method, we compare the questions obtained from the transcripts to questions generated from the fan-written episode synopses, taken from the Eastenders Fandom Wiki[5] scraped and used by [100] in their approach for the 2020 edition. We are aware that such resources are not always available (as it was the case for the 2021 dataset) and expensive to produce. However, next to giving more insights to the capacity of the question-generation models, we see a potential in leveraging this existing resource for our second line of work, which is about producing datasets with questions and context to train models for the domain of TV series.

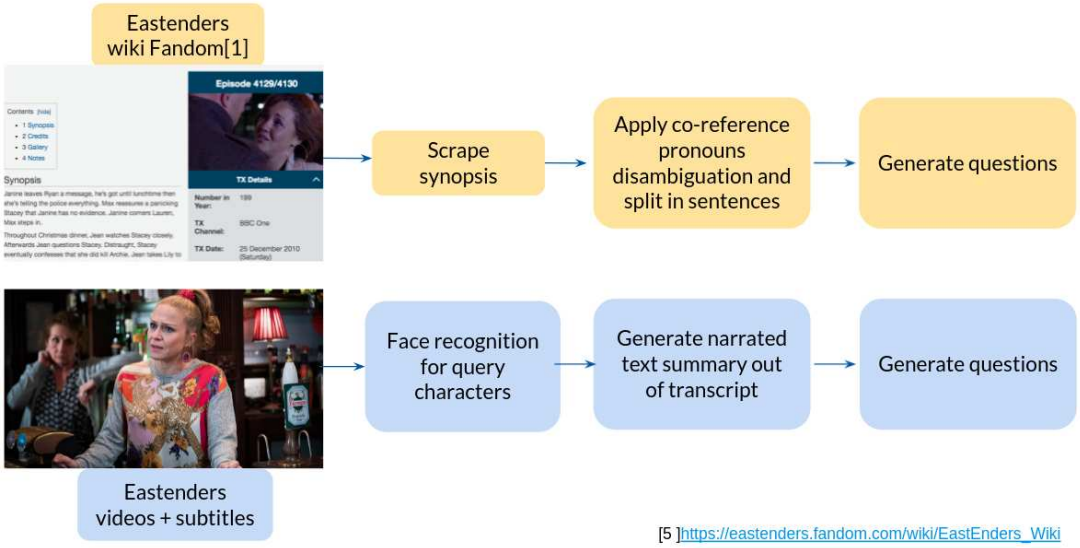


Figure 5.2: From the original text to the generated questions

Method In order to generate questions, we first need to follow some preprocessing steps which are presented in Figure 5.2. Most of them are similar to the ones taken by [100] for their challenge submission[6]. Similarly to this approach, we apply co-reference pronouns disambiguation to the synopsis to explicit character mentions using https://github.com/

[5]https://eastenders.fandom.com/wiki/EastEnders_Wiki
[6]https://github.com/MeMAD-project/trecvid-vsum

Character	Questions-nbr	Question
Janine	Q1	What is causing Ryan to be sick in bed?
Janine	Q2	How does Janine attempt to kill Ryan while in the hospital?
Janine	Q3	What happens when Janine attempts to play recording of Stacey?
Janine	Q4	Who stabbed Janine?
Janine	Q5	Who gives Janine the recording of Stacey?
Ryan	Q1	How does Janine attempt to kill Ryan in the hospital?
Ryan	Q2	What does Ryan do when Janine is lying in the hospital?
Ryan	Q3	Where is Ryan trapped?
Ryan	Q4	What does Ryan tell Phil he can do for him?
Ryan	Q5	Who is Ryan with when going to put his name on the babies birth cert?
Stacey	Q1	Who climbs up the roof to talk Stacey out of jumping off?
Stacey	Q2	What does Stacey reveal when in a cell with Janine, Kat, and Pat?
Stacey	Q3	What does Stacey admit to her mum in bedroom when mum is upset?
Stacey	Q4	Who confronts Stacey in restroom where Stacey finally admits to killing Archie?
Stacey	Q5	Who calls to Stacey's door to tell her to get her stuff and go after Stacey's mum had called the police?

Table 5.4: Questions used for summaries evaluation

huggingface/neuralcoref. As the task focuses on some specific characters and requires an important compression rate for the transcripts, we also follow their filtering choice which eliminates all shots where the character of interest is not present, using the Face Celebrity Recognition library [163].

Then, we follow [75], who observed an improvement for the TV shows question answering task, when converting dialog inputs to narrated text with a dialogue summarization model. This step is relevant to our task because the questions generation models we used were not trained on dialog text, which tends to be a more informal and repetitive type of text than structured narrated paragraphs [47]. Therefore, similarly to the question generation part, we use a T5 model, this time fine-tuned on the SAMSum corpus [88] which contains around 16K chat conversations together with speaker names (which is not the case for our dataset) and corresponding summaries. The style of these daily life chats spans from formal to informal also containing some emoticons, slang words and typos.

We produced a summary for every 1792 text characters (approximation of 512 tokens) and kept the default parameters of the model [7]. We only kept sentences where the name of at least one of the query characters appears (both for the generated summaries and for the fan synopsis), then produce questions for every chunk of 10 sentences as advised by [171] and use the three question generation models described earlier.

Qualitative Analysis In this section we perform a qualitative analysis, guided by questions about the number of questions generated, their type and closeness to the annotators questions.

[7] https://huggingface.co/henryu-lin/t5-large-samsum-deepspeed

Event	Questions VSUM	Questions dialogue
Killing Archie	Stacey Q4	What does Lauren hate about Bradley killing Archie? Who killed Bradley and Archie?
Ryan at the hospital	Janine Q2+Ryan Q1	Janine's wife is in what hospital?
Stabbing	Janine Q4	Who stabbed Ryan and Lily?
Changing certificate	Ryan Q5	Charlie is going to the Town Hall to get Lily's birth certificate changed?

Table 5.5: Events appearing in dialogue and annotators questions (Questions VSUM refer to the questions number and characters in Table 5.4)

Event	Question VSUM	Questions dialogue
Killing Archie	Stacey Q4	Who swears on Lauren's life that Stacey didn't kill Archie?
Killing Archie	Stacey Q4	Who accuses Lauren of not just killing Archie, but as good as pushing Bradley off the roof?
Stabbing	Janine Q4	What did Janine grab when she stabs herself?
Give/play a recording	Janine Q3+ Q5	What does Lauren give Janine to play the recording of?
Climbing on the roof	Stacey Q1	Who follows Stacee as she climbs onto the Vic roof?
Unknown Event	Ryan Q4	Who gives Phil some car keys and is given cash in return?

Table 5.6: Events appearing in synopsis and annotators questions (Questions VSUM refer to the questions number and characters in Table 5.4)

Is the number of sentences generated reasonable for a summary? After turning the transcript into narrated text, we obtain 299 sentences that we intended to reduce to 111 sentences by removing the ones that do not contain the names of the query characters. However, comparing the dialogue text to the narrated text, we realized that the dialogue-to-narrated text model sometimes mixes up characters, wrongly inferring who did what: for example, the model outputs 'Stacey stabbed Ryan and Lily' when in reality it was Janine who was stabbed. This is an issue which is most probably imputable to the absence of speaker diarization in the transcript. To avoid missing import life events of the query characters, we then kept all the 299 sentences for the question generation.

The Multi-task QA-QG and Single-task models create either one or multiple questions for every sentence in the source document, whereas the end-to-end model generates only a few questions per paragraph. We only report on the question generation results from the end-to-end model. Indeed, generating too many questions defeats the purpose of summarization, increases the number of anecdotal questions and therefore decreases precision. The challenge organisers considered 15 questions (5 per query character) to be a good number of summary questions for 10 episodes. We, instead, obtain 101 generated questions for the transcripts and 86 for the synopsis. The reason why we obtain more questions than the annotators is that the end-to-end model does not allow to directly control the number of questions. This option would be an interesting path for future work. For now, to comply with the VSUM task, the 2058 shots found by the face recognition model to contain at least one query character, would need to be reduced to 15 most important shots for the summary. This means that less than 1 percent of shots should be kept. Whether this number is to be used as an absolute reference, is, however, a question that remains open. We find it indeed useful to state that a look at the TV series summarization literature shows that there is no reduction agreed upon rate. Another TV series summarization dataset [83, 200]) rather keeps 30 percents.

Recall: are the topics of the expert questions covered by the generated questions? Table 5.4 displays the questions written by the annotators for the challenge evaluation. Table 5.5 and Table 5.6 display events mentioned in the annotators questions which can be respectively found in the generated questions from the transcript and the synopsis. It teaches us that the events of 5 out of 15 questions from the annotators are mentioned in the questions generated from the dialogue. This number reaches 5 or 6 for the synopsis questions. Indeed, the last row of table 5.6 is only an hypothetical match, since the expert question is 'What does Ryan tell Phil he can do for him?' and whether the question 'Who gives Phil some car keys and is given cash in return?' refers to the same event, remains ambiguous. When assessing whether this overlap is high, we need to keep in mind that we used an automated face recognition tool to filter out some shots and the performance of that tool has necessarily an impact on the final results. Given that these results were obtained with a question generation model that was not fine-tuned on extracting important questions for stories at all, we argue that these results are encouraging.

A somehow surprising observation is that the questions generated from the dialogues match with the expert ones as well as the ones generated from the synopsis do. One could have rather expected that the questions generated from synopsis would contain more import events as the synopses are written by humans and therefore supposed to be of better quality than the automatically generated narrated summaries. One reason for that might be that summarization remains a subjective task: fans and annotators do not necessarily agree on what is interesting. The challenge's recommendation was to extract important life events. When observing the synopsis based questions, we find that many questions could be considered as candidates for important questions. Examples include: 'What does Ryan give Janine when she doesn't get the job?', 'What does Stacey do when Ryan staggers out the front door and collapses at her feet?','Max reassures a panicking Stacey that Janine has no evidence of what?'. The transcript based questions include equally dramatic events such as 'Ryan cheated on Janine and poisoned her, what did she want to have in the new year?'. This observation might suggests that an assessment based on a comparison with expert questions has limitations: an additional human evaluation of the relevance generated questions might be needed.

In terms of error analysis,the questions in the tables confirm the early observations we made that the dialogue-to-narrated-text model sometimes mixed up the names and/or roles. For example in "Janine's wife is in what hospital?", 'wife' should be replaced by 'husband' to be correct. Another question states wrongly that Bradley killed Archie (correct would be Stacey). Another type of error that is sometimes observed is when the answer is included in the generated question (e.g.: "What did Max promise to look after Stacey after Bradley's death?").

What type of questions did the model generate? For the synopsis generated questions we find: 42 What question, 38 Who questions, 3 When questions, 2 Where questions, 1 How question For the transcript generated questions: 33 What questions, 30 Who questions, 0 When question, 5 Where questions, 5 How questions. In this expert questions, we find: 6 Who questions, 6 What, 0 When question, 1 Where question, 2 How questions. The type of questions is therefore quite similarly distributed between the generated questions, with 'What' and 'Who' questions being the most frequent ones. Finally, with regards to the style of the questions generated, the phrasing remains very close to the original text. This should make it easier for a question to be answered. For example the sentence:'Kat persuades Stacey and Jean to come to R&R with Kat and Kim?' becomes 'Who persuades Stacey and Jean to come to R&R with Kat and Kim?' (note that the repetition of Kat and Kim instead of 'them' comes from the co-reference step)

Conclusion and Future Work

For the summarization application, we found reasons to be enthusiastic about the possibility of creating non expensive datasets for query based summarization, with questions phrasing close to the original text. Despite using a real-world challenging dataset (no speaker indication nor ground truth face recognition), the types of questions are similar to the expert ones and in terms of events, many of the questions generated seem to include important life events. However, this first exploration needs be confirmed by a proper human evaluation which would assess the relevance of these questions for summarization. Going back to the topic of question generation for text understanding, this study has also highlighted that an absence of speaker diarization results in a mismatch between actions and the characters. Besides fixing this issue, in the future, we plan to go beyond these first out-of-the box experiments, by fine-tuning a question-generation model for the task of summary questions generation.

5.3 Conclusion

In this section, we established the power of merging knowledge about from two sources: First, a common-sense knowledge graph (ConceptNet), which encodes knowledge about words and their meaning in an explicit way. Second, pre-trained language models, which latently capture knowledge about word usage and language. We examined different ways to use this knowledge for zero-shot classification. We also established the efficiency of combining both representations on different datasets from several domains. One of the major direction outlined for the future of this work is human-in-the-loop: to engage humans in the formation of the label neighborhood. In general, given that there is a large amount of research in narratology applied to cinema or literature, we believe there is room to integrate expert knowledge to summarize stories, to create more datasets annotated for narrative aspects and

eventually to augment ConceptNet with more in-domain terms.

In this section, we have also contributed to pushing the effort towards evaluation of narrative understanding with objective tasks such as story QA. We have started exploiting the potential of Language Models for narrative question generation, the next steps would consists in further exploring prompting methods for this task and evaluating our results with human experts. In general, we have started using prompting in this section and it would interesting to further investigate this paradigm, for factual probing, investigating what type of narrative knowledge pre-trained LM internal representations bear.

Chapter 6

Conclusion and Future Work

6.1 Summary of the thesis

In this thesis, we have uses techniques from both the field of *NLP and Information Extraction*, and *multimodal content analysis*, to tackle a number of tasks about videos and the stories they convey. We have covered different aspects of storytelling in multimedia content: What makes a video memorable? How to extract important moments for narrative summarization? How extract story aspects from screenplays? How to formulate meaningful questions from TV series transcripts? In summary, the thesis lead to the following contributions:

- Obtaining leading results on the MediaEval Memorability Benchmark for 3 consecutive years, by leveraging pretrained deep models and combining different content representations (as text, as visual features, as multimodal embedding, as audio). Investigating the robustness of our memorability prediction models by testing it on 5 datasets.
- Conducted a study which isolates text elements of screenplays based on the nature of the information they convey (dialogue versus scenic information) and we tested different pre-training methods on two visio-linguistic models for the task of TV series summarization. We have shown that using a visio-linguistic architecture without paired data and without in-domain pre-training achieves near state of the art results.
- Demonstrated, notably through our participation in the TrecVID VSUM challenges, how the textual component of media which can be easily obtained automatically, can help tackle the task of character-based summarization: either by leveraging fan-made synopses, or using zero-shot classification to capture the major life events of characters.
- Developed a model that rely on external knowledge (common-sense knowledge from ConceptNet and linguistic knowledge from pretrained language models) to perform text classification in a zero-shot fashion, i.e., given just a list of labels. Showed that the

domain-adaptation capacities of our model is beneficial to the field of story understanding in multimedia content.

- Started to explore the potential of Language Models to automatically generate questions from TV series transcripts as mean to create non expensive datasets for query based summarization.

- Participated to the Tweet Engagement Prediction (RecSys Challenge '20) task. As this work is only loosely connected to the thesis topic, it was included in Appendix B

6.2 Future Work

Multimodal zero-shot classification We approached both summarization and narrative aspects extraction through the lense of text zero-shot classification. We have limited ourselves to leveraging on the textual representation of TV series. In the future, it would be good to integrate visual and audio cues in this process. In general, better integration of modalities for the representation of multimedia content. Follow the progresses made in the direction of modality-free representation. Beyond single-track improvement in each modality, Transformer-based architectures seem to be approaching the maturity point where they can be used on all modalities and perform just as well as the modality-specific ones (e.g. CNN for vision)[1].

Representing stories as graphs In this thesis, we have started pointing out at the fact that deep representations offer limited options of explanations. We saw that one solution was to work with the ConceptNet knowledge graph. Similarly, we think an exciting direction would be to use sota deep models to populate knowledge graph which explicitly represent stories in multimedia content. The TRECVid *Deep Video Understanding (DVU) Challenge* [62] explores this path by formulating the task of *Video Understanding* as one of extracting knowledge from all available modalities (speech, image/video and text) of the videos to solve different types of queries related to story understanding.

Further exploring the link between what is said and what is done In Chapter 4, we have introduced in the context of TV series screenplays, the topic of complementarity between what is said and what is happening visually. We also have started investigating the power of text generation models like Text-to-Text Transfer Transformer (T5) [220] on tasks such as question generation. We would find interesting to ask whether Languge Models allow to generate stage directions from dialogues only? We believe screenplays, offer a unique opportunity to explore

[1] The first high-performance self-supervised algorithm that works for speech, vision, and text

further the complementary link between dialogues and stage directions, with the goal to aid animation generation but also other tasks related to understanding the link between speech and visual in stories [117]. The link between dialogue and stage directions in screenplays has been used to obtain weak labels for action recognition [189]. However, the richness of information conveyed by stage directions goes beyond action verbs. Among others, it contains information such as location indications or sentiments. Such experiments could be particularly relevant to the line of research which aims to include a storytelling dimension to video captioning [38, 113]

Publications list

The research carried out during this PhD thesis has lead to the publication of the following scientific papers:

Journal Papers

1. Ismail Harrando*, Alison Reboud*, Thomas Schleider*, Thibault Ehrhart and Raphael Troncy (*Equal contribution). **ProZe: Explainable and Prompt-guided Zero-Shot Text Classification**. IEEE Internet Computing: Special Issue on Knowledge-Infused Learning, 2022.

2. Alison Reboud, Ismail Harrando, Pasquale Lisena and Raphael Troncy. **Stories of Love and Violence: Zero-Shot interesting events classification for unsupervised TV series summarization**. In *Multimedia Systems - Special Issue on Data-driven Personalisation of Television Content, under review*, 2022.

Conference and Workshop papers

1. Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphael Troncy, Hector Laria Mantecon. **Combining Textual and Visual Modeling for Predicting Media Memorability.**. In *10^{th} MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)*, 27-29 October 2019, Sophia Antipolis, France.

2. Ismail Harrando, Alison Reboud, Pasquale Lisena, Jorma Laaksonen and Raphael Troncy. **Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization**. In *International Workshop on Video Retrieval Evaluation (TRECVID)*, 17-19 November 2020, Online.

3. Alison Reboud, Ismail Harrando, Jorma Laaksonen, Raphael Troncy. **Predicting Media Memorability with Audio, Video, and Text representation**. In *11^{th} MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)*, 14-15 December 2020, Online.

4. Alison Reboud and Raphael Troncy. **What You Say Is Not What You Do: Studying Visio-Linguistic Models for TV Series Summarization**. In 4^{th} Workshop on Closing the Loop between Vision and Language (CLVL), at ICCV, 2021, Online.

5. Alison Reboud, Ismail Harrando and Raphael Troncy. **Zero-Shot Classification of Events for Character-Centric Video Summarization**. In *International Workshop on Video Retrieval Evaluation (TRECVID)*, 7-10 December 2021, Online.

6. Alison Reboud, Ismail Harrando, Jorma Laaksonen and Raphael Troncy. **Exploring Multimodality, Perplexity and Explainability for Memorability Prediction**. In *12^{th} MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval)*, 13-15 December 2021, Online.

Technical Reports

- Ismail Harrando, Benoit Huet, Dejan Porjazovski, Alison Reboud, Michael Stormbom, Raphael Troncy and Tiina Lindh-Knuutila. **MeMAD Deliverable D3.2 - TV Moments Detection and Linking, version 1**. January 2020. https://memad.eu/wp-content/uploads/D3.2-TV-moments-detection-and-linking-revised.pdf

- Amine Dadoun, Ismail Harrando, Pasquale Lisena, Alison Reboud and Raphael Troncy. **Two Stages Approach for Tweet Engagement Prediction**. arxiv:2008.10419, 2020. https://arxiv.org/abs/2008.10419

- Thibault Ehrhart, Ismail Harrando, Ilkka Koskenniemi, Mikko Kurimo, Jorma Laaksonen, Tiina Lindh-Knuutila, Pasquale Lisena, Dejan Porjazovski, Alison Reboud and Raphael Troncy. **MeMAD Deliverable D3.3 - TV Moments Detection and Linking, version 2**. April 2021. https://memad.eu/wp-content/uploads/D3.3-TV-moments-detection-and-linking-v2.pdf

Chapter A

Question-Answer Generation For Silk Text Classification

How does such an approach compare with Zero-Shot classification for extracting specific type of information (e.g. about silk fabrics)? In Section A.0.0.1, we present our experiments and results for the task of silk information extraction.

Question-Answer Generation and Text Classification Leveraging question answering or question generation for information extraction is not new, as it has been studied even before the emergence of deep learning or transformer-based models [288], but it is still rather rarely studied. Despite this, recently several promising models have been recently proposed such as e.g. QuAChIE for the Chinese language [231].

A unified multi-task learning framework for joint extraction of entities and relation that consisted importantly on a sub-task including question generation based QA with a transformer-based Seq2Seq model is another new example [303]. An important feature was the detection of subjects and objects without relying on NER models in this pipeline. In this section, we only consider a pipeline for text classification and could therefore not use this framework, but consider it relevant that an information extraction task has been pre-processed through question generation.

Finally, we would like to present one more recent approach which leverages question generation for entity and relation extraction [93]. In this case, the question answer model was created by training BERT on the SQuAD dataset. The input texts are pre-processed with a NER model and then uses a phrase generation method to frame the questions. In general, these few recent examples show promising results, but we are not aware of any recent work about pipelines that consist of question generation and text classification as in our case.

A.0.0.1 Question generation for key information extraction from texts about silk fabrics

Method Our approach consists in generating questions and answers from textual descriptions of silk made objects. For reasons of self-evaluation, we choose only objects with an associated property, such as the material, to see if this known categorical value is included in the generated output. To verify if this is the case, we both perform simple string matching and fuzzy string matching by first measuring an edit distance similar to the classical Levenshtein distance [147] between two tokens, the label and the tokenized version of the full output of one of the T5 models with an empirically defined threshold of 0.9.

Preliminary results have shown that this edit distance measurement is showing equal or better scores than calculate the semantic similarity after converting all input into word vectors with the most recent large English model of Spacy [1], whose word vectors are trained with GloVe [208] on the Common Crawl [2]. This is why we do not include the latter results of the semantic similarity with word vectors in this section.

This dataset in this form is only a slight extension of one already used together with the model ZeSTe [240], which makes quantitative evaluation possible. ZeSTE [101] (or Zero-Shot Topic Extraction) uses ConceptNet and the nodes neighborhoods of the knowledge graph to compute similarity between the tokens of an input text and the target concept classifying the document. The candidate ranking of ZeSTe got hereby updated after prompting the language model BART [150] with a sentence related to the domain of the word (in this case, e.g. "silk textile"). We provide a comparison with both this updated and prompting-guided Zero-shot classification method as well as the results of ZeSTe itself on a very similar dataset. As a final baseline we provide the class distribution, which illustrates the multi-label classification setup.

To put the results into perspective, we compare our predictive scores with three baselines. Finally, we also qualitatively analyze if our approach has the ability to predict values that those other methods could not.

Quantitative Analysis Table A.1 contains the scores of our auto-evaluation experiments. We observe that both the scores obtained through the edit distance measurement and simple string matching in almost all cases beat the class distribution baseline, but are not coming very close to the two predictive performances of the Zero-Shot Classification models. The Multi-task model achieves the best performances between the three T5-based models, for both SILKNOW properties.

This might come from the added complexity of the multi-task model as it is more fine-tuned on the separate tasks of answer extraction, question generation and finally question answering.

[1] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.2.0
[2] https://nlp.stanford.edu/projects/glove/

The single-task model is second best for materials, but behind end-to-end for the techniques condition. The end-to-end model is the only one that achieves a score lower than one of the class distribution baselines, but only when we apply simple string matching, which consistently yields worse results then edit distance measurement. On average the end-to-end model still produces output that scores mostly comparatively to the other models, despite it not producing any answers, but only questions.

Between all conditions we can also observe, that the scores for the SILKNOW Techniques are consistently higher, despite this condition having one more class and the baseline being accordingly lower. The reason for this may simply lie in the average length or quality of the original input texts that are ultimately coming from different museums. Potentially the difference could also stem from the semantic similarity or dissimilarity between the class labels inside of one condition. The two Zero-Shot models confirm this discrepancy between the data for the different properties as well.

Model	Measurement	SILKNOW Materials	SILKNOW Techniques
Single-Task T5	**Edit Distance**	**19.40%**	**25.10%**
Multi-Task T5		**24.50%**	**28.80%**
End-to-End T5		**17.40%**	**26.30%**
Single-Task T5	String Matching	15.30%	19.80%
Multi-Task T5		17.30%	23.00%
End-to-End T5		12.50%	20.00%
Prompt-guided ZS Classification	Accuracy	39.00%	50.80%
ZeSTE*		34.3%	46.9%
Baseline	Class Distribution	14.00%	12.50%

Table A.1: Auto-evaluation scores based on matches between the target label and the generated question(-answers). Comparison with the label prediction accuracy of two Zero-Shot classification methods that have been performed on the same dataset (*For ZeSTE the results of its application on a minimally different, but comparable dataset are stated here). The baseline is representing the class distribution.

Qualitative Analysis For the SILKNOW dataset, we also investigate if the output of the question-answer generation models could be a supplement to the output of text classification models. For this, we compare if the output of the T5-based models explored in this study could be matched with the labels of objects for which the ZeSTe-based model could not predict them.

Table A.2 shows some examples that are solely selected based on beating the prediction of our baseline model. In most cases, we still got proper English sentences that ask relevant

questions. An exception is hereby the last row which shows the rather strange question "What scroll appears to have read 'Benedetto Ghalilei'?". We also have some examples of very technical question-answer pairs whose use might be quite limited, for example: "Question: How many threads per in? - Answer: 36-38". Nevertheless, this might be quite an interesting detail which is not yet explicitly available in the knowledge graph. An automatic extraction of it might be complex, but this output still emphasizes the highlighting abilities of the model which could further be leveraged in the future.

Next to that we have several examples of locations or time-spans in some of the answers that could easily be linked, like "Kerman, Iran", "India" or "17th century". Linking of such further properties can be used with several other applications of SILKNOW, for example a spatio-temporal map that an end-user could use. Given future expert evaluation we see great potential here for use of these question and answers not only for further enrichment of the knowledge graph, but directly in some web applications.

As a common pattern we could observe that the T5-based models described in this section were almost never better than the stated Zero-Shot methods at predicting one of the two respective majority labels (Textile for material, and embroidery for technique), but did occasionally so for some of the smaller labels. For example the one displayed in table. We could not find out a proper reason for this, but this also hints at a potentially useful complimentary function of these models next to other better performing (Zero-shot) classification methods for a future work.

Conclusion For the application on a dataset with metadata about historical European silk fabric we can conclude that the output of the question(-answer) models is not directly surpassing the state-of-the art of zero-shot classification, but showcases promising highlighting abilities when it comes to producing questions or answers from relevant text sections which goes beyond random selection. As far as we can qualitatively analyze without expert confirmation, we consider the output in most cases to be grammatically correct and useful. We can also observe that despite the lower classification performance of these models, some results appear to be complimentary to the main baseline model: some labels were matched that could not be predicted before.

Matched Label	Prompt-guided ZS classification	Property	Selected Output
Single-Task T5			
Cotton	Wool	Material	{'answer': 'Kerman, Iran', 'question': "Where was the 'Vase Carpet' lattice design located?"}, {'answer': 'silk', 'question': 'Along with cotton weft and wool knotted pile, what textile is used in Persian carpets?'}, {'answer': 'wool', 'question': 'What type of fiber is the carpet made of?'}
Velvet	Brocaded	Technique	{'answer': 'red', 'question': 'What color is the cut and uncut velvet?'}
Multi-Task T5			
Wool	Silver	Material	{'answer': '36-38', 'question': 'How many threads per in?'}, {'answer': '16', 'question': 'How many knots per in?'}, {'answer': 'wool', 'question': 'What is the Pile made of?'}
Muslin	Embroidery	Technique	{'answer': 'embroidered muslin', 'question': 'What is the girdle made of?'}, {'answer': 'India', 'question': 'In what country is the girdle of muslin embroidered with silk and silver threads?'}, {'answer': '17th century', 'question': 'When was the girdle of embroidered muslin made?'}
End-to-end T5			
Cotton	Silver	Material	'How many threads per inch does white cotton have?', 'How many shoots of weft do gold-coloured cotton and gold coloured silk have per inch?', 'What color are the lilies in the center of the present gragment?', 'Where do the white lily veins meet on the horizontal plane?', 'Which leaves form a square frame?'
Velvet	Brocaded	Technique	'What is the coat of arms of the Galilei family of Florence represented by?', 'How many rungs are under the cross in the center of the velvet?', 'What scroll appears to have read "Benedetto Ghalilei?"', 'What may have been used in a set of ecclesiastical vestments for a family chapel?'

Table A.2: Two generated output texts per T5-based model. All examples represent cases in which the target label could be matched with the output and the Prompt-guided ZS classification method used on the same dataset predicted a wrong label.

Chapter B

Two Stages Approach for Tweet Engagement Prediction (RecSys Challenge '20)

Introduction

Dealing with a constantly increasing quantity of information is one of the challenge of modern computer science. The growing amount of content posted on social networks requires the introduction of algorithms that provide end-users with the most relevant content in order to improve their experience. Predicting if a given user would actively interact with a post is a key goal for optimising these algorithms that aim to sustain engagement of the user on the platform.

This paper describes our approach for the task of engagement prediction for the 2020 RecSys Challenge [28]. The target dataset – released in the context of the challenge [27] – includes 160M[1] public engagements from the Twitter timeline, including both positive (like, retweet, reply, retweet with comment) and negative (absence) examples of engagements. Our method can be described as a two stages approach:

- In the first stage, different learning modules extract heterogeneous features from the dataset. Those modules are: handcrafted features extractor, knowledge graph embedding, sentiment analysis and engagement predictions based on tweet content as represented by BERT tokens;
- In the second stage, these features are combined in input to an ensemble system, implemented using XGBoost [49].

The implementation of this approach is publicly available at https://gitlab.eurecom.fr/dadoun/RecSys_Challenge_2020.

[1] It is worth to note that more than 10% of the data has been deleted during the course of the challenge and was not processed

The remainder of this paper is organised as follows. Section B presents our approach, while Section B details its application to the challenge dataset, together with an internal evaluation protocol and the obtained results. Finally, Section B outlines some conclusions and future work.

Approach

The approach that we propose for predicting the engagement on tweets relies on two subsequent stages, shown in Figure B.1. In the first stage, from the set of features D contained in the challenge dataset, we select 4 subsets D_i. Each of those is processed by a different learning module i, which gives in output the set of features D'_i. The four modules are detailed respectively in Section B, B, B and B. The second stage implements the engagement prediction task. An XGBoost classifier [49] is trained on the previously generated features $D^{'}_{i}$ acting as an ensemble classifier. It returns in output the probability that the user u performs an engagement for a tweet (like, retweet, reply, retweet with comment) $P_{engagement}(u,t|D)$. This stage is detailed in Section B.

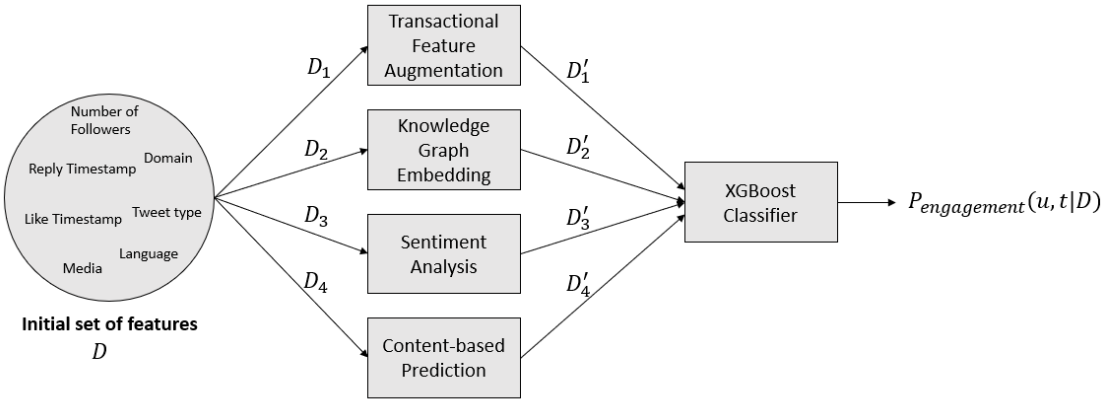


Figure B.1: Two stages approach for tweet engagement prediction

Transactional Feature Augmentation

In addition to the list of transactional (interaction between users and tweets) features[2] provided in the released dataset [27], we compute some additional features (feature augmentation) that contains more information about the tweets and the transaction (user, tweet). For the sake of explanation, we distinguish between two kinds of users, the *reader* which represent

[2]This list of transactional features includes: `present_domains`, `tweet_type`, `language`, `present_media`, `engagee_follows_engager`, `hashtags`, `engaging_user_follower_count`, `engaging_user_following_count`, `engaged_with_user_follower_count`, `engaged_with_user_following_count`, `engaging_user_account_creation`, `engaged_with_user_account_creation`, `engaged_with_user_is_verified`, `engaging_user_is_verified`.

the `engaged_with_user` and the *author* with represent the `engaging_user`. We detail below the list of added features:

- Number of engagements per reader: For each engagement (e.g. like, retweet, etc.), we compute the number of times this specific engagement has been performed by the reader before encountering the current tweet.
- Number of engagements per author: For each engagement (e.g. like, retweet, etc.), we compute the number of times this specific engagement has been received by the author before posting the current tweet.
- Number of engagements per user towards an author: This feature represents the number of times a reader has made an engagement to an author in the past (before seeing the current tweet).

Knowledge Graph Embedding

Several connections can be seen in the challenge dataset: the relationship *follower-followed* between users, the authorship of a tweet, the interaction with another one, the sharing of hashtags or domains among tweets. Knowledge graphs (KG) provide a suitable way for representing these connections and they have already largely been used to model social networks [78, 280]. KGs have also successfully been exploited in recommender systems, in particular using graph-embeddings [164, 186, 199].

We used the information coming from the dataset for populating a KG, whose structure is illustrated in Figure B.2. The core of this structure is made of the tweet and of the user. The latter can be, in different moment, either the author of a tweet – to which it is linked through a `write` edge – or the one that reads a tweet – eventually linked to it through an interaction edge, such as a `like`. When a user follows another user, a specific edge links the two. In addition, a class is assigned to each user depending on her or his number of followers following the distribution presented in Table B.1. Apart from the edges connecting users, a tweet node is also linked with five literal nodes:

- `has type`: TopLevel, Quote, Retweet, or Reply (the value corresponds to the `tweet type` column in the dataset);
- `has media`: Photo, Video, GIF, or Photos (when there is more than 1 photo);
- `has lang`: language identifier;
- `has hashtag`: hashtag identifier;
- `has domain`: domain identifier.

CLASS	MAX FOLLOWERS	CLASS	MAX FOLLOWERS
0	150	4	100,000
1	500	5	1,000,000
2	1,000	6	10,000,000
3	10,000	7	200,000,000

Table B.1: Classification of users depending on their number of followers

The KG is populated reading the dataset tsv file line by line and creating node and edge instances when required. For example, the *has domain* link would be present only if the tweet contains a domain link. As a consequence, not all edges are created for each row. Figure B.2 represents always-present edges with a continuous arrow, while dashed arrows mark optional edges.

For being used in input to machine learning algorithms, the graph embedding process transforms the graph structure in a set of multi-dimensional vectors. For this purpose, we used *node2vec* [91], a state-of-the-art algorithm that generates random walks between the nodes of the graph, on which it computes the transition probabilities between nodes, which are mapped into the vector space. In other words, nodes sharing more connections are more likely to be part of the same random walk and consequently are more likely to be close in the computed embedding space. We assigned to each kind of edge a different weight, which impacts on the possibility of its nodes to appear in the same random walk.

The limitations of this approach are: the required resources since the machine needs to load the entire graph in memory, and the long computation times, which grows non linearly with the number of nodes and edges. In order to obtain results in a reasonable time for the challenge, we performed the training of these embeddings only on a subset of 40M dataset entries, taking into account only some kind of edges, namely `follow`, `write`, `like`, `has domain`, and `has hashtag`.

Sentiment Analysis

The task of sentiment analysis aims at attributing a predefined sentiment category to a text sequence. In our particular case, it means assigning a positive or negative polarity to each tweet. BERT is very effective for text classification. However, the available pre-trained BERT models for sentiment analysis have been trained on general corpora such as the BooksCorpus (800M words) [308] and English Wikipedia (2,500M words), while a fine-tuned model on in-domain data may increase its efficiency. We searched for a domain dataset, containing annotations for the task of sentiment analysis, with two main requirements: *i)* it must contain tweets as they represent a specific type of expression, written in a certain style and with a length constraint and *ii)* the dataset must contain more than 30 languages.

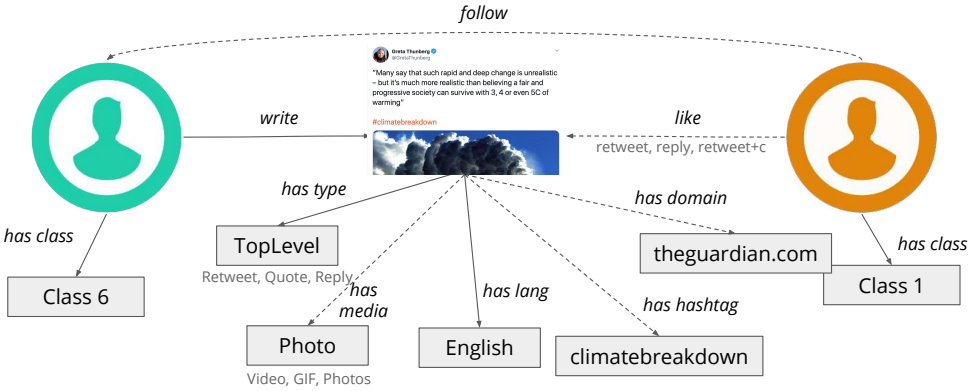


Figure B.2: Excerpt of the knowledge graph. Most of the values are de-anonymised for simplicity, while they are in reality identified with an alphanumeric code (i.e. domain, language).

If we have been able to find a sentiment analysis dataset for English tweets called Sent140 [89], to the best of our knowledge, a sentiment analysis dataset matching the language distribution of the challenge dataset does not exist. Given that English still represents approximately 40% of the tweets in the dataset, we used the Sent140 dataset in the prediction of sentiment labels for those tweets, ignoring those written in other languages.

In [262], a fine-tuning method for pre-trained BERT models achieves state of the art results for a variety of tasks and datasets. This approach consists in two steps:

1. further training BERT on within-task training data, in our case the tweets;
2. fine-tuning BERT for the target task using labels, in our case sentiment polarity.

The authors use this method for fine-tuning a pre-traind BERT, providing as input a dataset of film reviews from the IMDB Dataset [178], annotated with the sentiment polarity. The result is a BERT model specialised for sentiment analysis. We performed a further fine-tuning to this model on the Sent140 dataset, in order to capture Twitter-specific expressions and style. The final model has been used on the BERT tokens from the challenge dataset, once decoded to plain text. We used both the sentiment analysis labels predictions and corresponding logits as predictive features for the ensemble.

Content-based prediction from BERT Tokens

The content of tweets was provided in the dataset as a list of BERT [71] token IDs corresponding to multilingual words (e.g. “$token21601$" → "$spiel$") or sub-words ("$token21603$" → “ – $sation$"). These tokens can be decoded into strings using the appropriate BERT Tokenizer (`bert-base-multilingual-cased`) or alternatively used as-is to represent the tweet content.

In our work, we implemented two distinct methods for exploiting these tokens:

- We fed them into a pre-trained multilingual BERT to generate a fixed-length representation of the tweet textual content, either by pooling (e.g. averaging) the transformed token representations at the output of the BERT model, or by taking the `[CLS]` embeddings[3] which somehow represents the entire input sequence. Both representations are dense 768-dimensional vectors.
- We apply the list of tokens as a bag of tokens in input to a TF-IDF model, which uses the count of each token in the tweet and normalises it by the token count in the entire dataset. Since decoding the tokens into their original form increases the vocabulary significantly, we opted for directly using the tokens as represented by their IDs. We also keep the highly-occurring token n-grams ($n <= 3$). This generates a 1M-dimensional sparse tweet representation.

Both fixed-size representations are then fed into models to predict the interaction with the tweet (one classifier for each interaction, with a binary output). We use a SVM classifier with the (sparse) TF-IDF features and a feed-forward neural network with the BERT embeddings. The output of these models have been used as a feature (D_4') into the ensemble system.

XGBoost

XGBoost [49] is an implementation of gradient boosting decision trees. As presented in Figure B.1, XGBoost takes as input the outputs of the first stage modules $D_i^{'}$. We detail below the different outputs of stage 1:

- $D_1^{'}$: The output of **transactional feature augmentation** module, represented by features coming from the challenge dataset and from extracted transactional features (Section B);
- $D_2^{'}$: The **Knowledge Graph Embedding** computed in Section B representing readers, authors and tweets;
- $D_3^{'}$: The outputs of the **sentiment analysis** module are labels predictions and corresponding logits (Section B);
- $D_4^{'}$:The softmax score of the **SVM classifier** trained on the TF-IDF model representing the text tokens (Section B).

[3]When given a sentence as input, the BERT model outputs a contextual embedding for each token of that sentence, as well as a "sentence-wide" representation for classification purposes (represented by the special token [CLS]). According to both the original paper and our experiments, both representations generate comparable results when fed to a model to predict user engagement.

For each engagement prediction task, an ablation study is performed for features selection in order to train the model on a subset of features $D_{engagement} \subset (D_1^{'} \cup D_2^{'} \cup D_3^{'} \cup D_4^{'})$. This helps speed up the training and also improve slightly PRAUC & RCE scores. Moreover, we performed a grid-search to find optimal hyper-parameters of XGBoost classifier.

Experiments

In this section, we discuss how we have implemented our approach and we comment on the obtained results. Following the challenge rules, the evaluation relies on two metrics: the area under precision-recall curve (PRAUC) and the relative cross-entropy (RCE) [27].

Development Pipeline

The challenge dataset consists of a training (121M public engagements) and validation set (12M), together with a final submission set (12M) released in the last part of the challenge. Only the training set contains the information about the engagements. To enable a faster computation and evaluation of planned experiments, we relied on an development pipeline composed of three stages represented in Figure B.3. In each stage, training and evaluation are performed in different subset of the original training set:

1. In the first stage, we use a randomly extracted subset from the training set that including 2 million rows (i.e. public engagements) and split it into a training set (90%) and a development set (10%). This phase allowed to perform experiments with different methods and feature sets.

2. When an improvement was observed in both PRAUC and RCE, we moved to a second stage where the full original training set was used, split again into local training set (90%) and local development set (10%). This phase helped us to understand if the method developed during the first stage is generalizable to a larger dataset.

3. Finally, in the last stage, we use the trained models to compute predictions on the original validation set. These results are submitted to the (public) challenge leaderboard.

Because of time and hardware constraints, we used in Stage 2 a local training set corresponding to only 40% (instead of 90%) of the original dataset. This additional sampling has enabled us to study the results obtained by the KG embedding and content-based prediction from BERT tokens modules. In fact, we could not scale to the entire training set with these modules due to memory limitations.

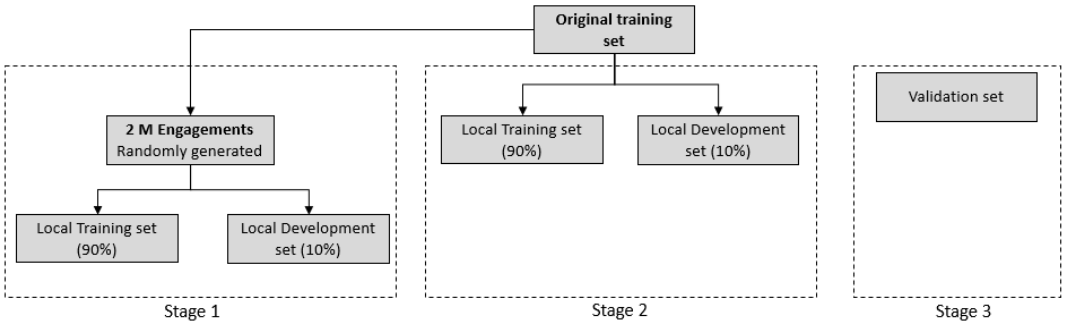


Figure B.3: Development Pipeline

Results and Discussion

Table B.2 presents the results obtained on local development dataset from the different implemented models trained in Stage 2:

- Model 1: XGBoost trained on only $D_1^{'}$.
- Model 2: XGBoost trained on $D_1^{'}$ & $D_2^{'}$.
- Model 3: XGBoost trained on $D_1^{'}$ & $D_3^{'}$.
- Model 4: XGBoost trained on $D_1^{'}$ & $D_4^{'}$.

Model	PRAUC Retweet	RCE Retweet	PRAUC Reply	RCE Reply	PRAUC Like	RCE Like	PRAUC Retweet with Comment	RCE Retweet with Comment
Model 1	0.66	40.03	0.33	**26.44**	0.86	38.63	0.18	**17.71**
Model 2	0.16	-21.35	0.22	-56.22	0.43	-12.84	0.05	-205.19
Model 3	0.64	38.65	0.27	22.14	0.80	34.23	0.16	15.20
Model 4	**0.68**	**42.03**	**0.33**	25.92	**0.89**	**41.08**	**0.19**	16.51

Table B.2: Models evaluated on our local development set (10% of the original training set)

We only submitted to the public leaderboard the models 1 and 4 which were the most accurate ones. As we can see from the values in Table B.3, PRAUC and RCE scores were remarkably different. We believe that the reason for this is the different variables distribution between the training and validation sets based on an exploratory data analysis we performed on the datasets.

We finally used our Model 4 for the final submission on the test data. Even if it is not the better performing one, we believe that it best represents our approach which is to combine features coming from different data sources. This approach ranked at position 22 in the final

Model	PRAUC Retweet	RCE Retweet	PRAUC Reply	RCE Reply	PRAUC Like	RCE Like	PRAUC Retweet with Comment	RCE Retweet with Comment
Model 1	0.28	7.82	0.07	5.25	0.66	10.25	0.02	-25.07
Model 4	0.32	7.98	**0.14**	**7.82**	0.66	12.21	0.02	-22.35
Model 1+	**0.39**	**15.94**	0.12	6.45	**0.68**	**12.44**	**0.03**	**-16.71**

Table B.3: Models evaluated on the validation set. Model 1+ was trained using the entire local training set.

leaderboard[4] (Table B.4). In contrast to the top ranked systems, we observe that our method is able to obtain a better balance between RCE and PRAUC for all predictions.

Model	PRAUC Retweet	RCE Retweet	PRAUC Reply	RCE Reply	PRAUC Like	RCE Like	PRAUC Retweet with Comment	RCE Retweet with Comment
Model 4	0.2924	5.95	0.0789	0.25	0.6145	0.37	0.0186	-11.40

Table B.4: Results on the final test set

Conclusion and Future Work

Predicting user engagement facing tweets on a very large dataset is a challenging task, both in terms of leveraging the massive information available at scale and coming up with representative and significant features to feed the different models. In this paper, we have experimented with a broad set of features we extracted and exploited for this task.

A key takeaway is the importance of leveraging the entire dataset to have a well-performing model, as our more advanced models, while showing very good results on our local training subset, did not manage to perform as well on the public leaderboard. The effort in studying and understanding the data and its distribution according to key factors has proven to be crucial for realising subsets that are representative of the final validation set. This would also stand as a straightforward solution to the scalability limitation of some of the used models. Other improvements include the capture of time variant information (for example the start and stop of a user following action), which can be represented both in the transactional features and in the KG.

[4] https://recsys-twitter.com/final_leaderboard/results

Chapter C

Resume en francais

EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

PHD THESIS

In Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy from Sorbonne University

Specialization: Data Science

Vers la compréhension automatique de contenu audiovisuel narratif

Alison REBOUD

Comité:

Rapporteur	**Christophe GRAVIER**, Université Jean Monnet, Saint-Etienne, France
Rapporteur	**Vasileios MEZARIS**, Centre for Research and Technology Hellas, Thessaloniki, Greece
Examinatrice	**Claire-Hélène DEMARTY**, InterDigital Inc., Rennes, France
Examinatrice	**Maria ZULUAGA**, EURECOM, Sophia Antipolis, France
Directeur de Thèse	**Ulrich FINGER**, EURECOM, Sophia Antipolis, France
Co-Directeur de Thèse	**Raphäel TRONCY**, EURECOM, Sophia Antipolis, France

Abrégé

Qu'il s'agisse de films et séries produits par l'industrie du divertissement et distribués sur des plateformes de streaming, ou de médias sociaux où les utilisateurs affichent les histoires de leur vie avec la fonctionnalité 'story', la narration moderne est numérique et basée sur la vidéo. Comprendre les histoires contenues dans les vidéos reste un défi pour les systèmes automatiques. Avec la multimodalité comme thème transversal, cette thèse décompose la tâche de "compréhension" en différents défis qui couvrent différents aspects du concept.

1. **Prédire le degré de mémorabilité d'un contenu multimédia.** Face à la multiplication des vidéos, la capacité à identifier et à créer un contenu mémorable suscite un intérêt croissant. La mémorabilité est une idée particulièrement intéressante car, contrairement à d'autres concepts associés tels que l'"intérêt", elle peut être mesurée objectivement par des tests de reconnaissance. Nous explorons la tâche de prédiction automatique de la mémorabilité des vidéos dans le premier chapitre, en utilisant des modèles multimodaux avec des indices visuels, textuels et audio pour différents types de vidéos.

2. **Résumer du contenu multimédia.** Après avoir extrait les moments mémorables, nous étudions comment extraire les moments qui sont importants pour l'histoire de séries télévisées. En raison du coût élevé de l'annotation pour cette tâche, nous avons décidé de capitaliser sur la richesse de la composante textuelle qui accompagne généralement ce type de contenu pour développer des approches non supervisées.

3. **Modélisation de la narration dans des contenus multimédia** Enfin, le dernier chapitre fait un pas de plus vers la compréhension narrative. Pour ce faire, il (i) propose PROZE, une nouvelle approche explicable, pour la catégorisation de textes, qui s'avère prometteuse pour la tâche de classification des aspects narratifs. (ii) découvre comment les modèles de langage peuvent être utilisés pour générer des questions importantes sur les intrigues des séries télévisées, auxquelles un résumé construit automatiquement devrait pouvoir répondre.

C.1 Contexte

Selon le paradigme narratif de Fisher [82], raconter des histoires est un trait humain naturel. Des aventures d'Ulysse aux blogs en ligne, c'est une activité de longue tradition, qui prend la forme de son époque. Si la crise du Covid nous a rappelé notre attachement aux cinémas, cafés, bars et autres lieux où l'on partage traditionnellement aventures et anecdotes, elle a aussi accéléré l'explosion de la consommation de contenus multimédias numériques. Cette tendance vaut pour les médias narratifs produits par l'industrie du divertissement, mais aussi pour les contenus créés par les utilisateurs sur les plateformes sociales, où l'on se voit offrir la possibilité de transformer sa vie en histoires (*story* étant littéralement le nom d'une fonctionnalité sur Instagram, Snapchat et Facebook). Dans le secteur du cinéma, Disney a dépassé les 100 millions d'abonnés dans le monde, moins de deux ans après le lancement de sa plate-forme [1]. Malgré un ralentissement du nombre d'abonnés en raison de l'essoufflement du boom de la pandémie, Netflix n'a pas perdu les 36 millions de nouvelles inscriptions qu'elle a obtenues grâce au verrouillage, dépassant ainsi les 200 millions de clients dans le monde [2].

Cette entreprise, ainsi que ses concurrents (HBO, Amazon Prime...) ont su surfer sur l'engouement pour les séries télévisées, un format longtemps méprisé [36], qui a aujourd'hui complètement gagné en reconnaissance, comme le démontre la création du Festival international des séries de Cannes en 2018 [3]. Le poids de Netflix sur le secteur du divertissement est tel qu'il est désormais également un acteur majeur de la production cinématographique, produisant des films acclamés par la critique directement sur Internet [4]. En ce qui concerne les vidéos créées par les utilisateurs, on estime qu'environ une personne sur six aux États-Unis utilise TikTok chaque semaine [5], une plateforme de partage de courtes vidéos, lancée il y a 5 ans. Pour faire face au succès de ces réseaux sociaux vidéo, Meta a créé la fonctionnalité reels, leur propre version des vidéos TikTok. De son côté, Youtube, le géant de la vidéo détenu par Alphabet, dont le nombre d'utilisateurs mensuels est estimé à deux milliards [6], a développé Shorts, sa plateforme de vidéos courtes de type TikTok [7]. Enfin, comme le montre la campagne " Storytelling Goes Here ", Meta a maintenant l'ambition de développer davantage ses capacités de partage de vidéos, en facilitant le chargement de vidéos de longue durée grâce à de nouvelles options de segmentation. [8].

Tous ces chiffres confirment que les gens regardent et produisent des histoires, en utilisant des

[1] The Guardian - Disney forecast to steal Netflix's crown as world's biggest streaming firm

[2] Netflix records dramatic slowdown in subscribers as pandemic boom wears off

[3] Site web du Festival international des séries de Cannes

[4] Netflix décroche 7 prix, doublant presque son total historique aux Oscars

[5] Croissance massive de TikTok : Plus 75 % cette année, 33 fois plus d'utilisateurs que le concurrent direct le plus proche

[6] TikTok dépasse YouTube pour le temps de visionnage moyen aux États-Unis et au Royaume-Uni

[7] YouTube souligne les principaux domaines de croissance, notamment l'essor de Shorts et l'expansion de son économie des créateurs

[8] Meta lance Facebook Reels pour tous les utilisateurs, élargissant ainsi son offre de vidéos de courte durée

outils digitaux et que les video courtes sont particulièrement populaires sur les médias sociaux. En bref, le storytelling moderne est numérique et passe par la vidéo. Une telle évolution est particulièrement intéressante pour les chercheurs en IA : tout en offrant des cas d'utilisation concrets - le secteur a besoin de divers outils pour l'aider à naviguer sur une mer de vidéos qui continue de s'étendre -, la richesse du contenu multimédia posent des questions de recherche d'une grande complexité.

Le spectre des tâches liées à la compréhension du multimédia est large et englobe différents niveaux de complexité. Si la localisation, les visages ou la classification d'images ont été abordés avec succès par des systèmes automatisés [136, 241], certaines tâches plus globales telles que la compréhension de l'intrigue d'une histoire ou le résumé de vidéos restent un défi.

Par exemple, pour créer des résumés de séries télévisées, un système automatique devrait capturer les scènes qui sont importantes pour la narration. On peut imaginer qu'un modèle traditionnel générant une bande-annonce sur la base de caractéristiques de bas niveau ne parviendra pas à saisir la sémantique de la vidéo et qu'il y aura un écart entre les modèles et la manière dont les humains traitent le contenu. Quelles sont les scènes que les humains considèrent comme des parties essentielles d'un média narratif ? De quels types de vidéos se souvient-on ? Peut-on les prédire automatiquement ? Ce sont quelques-unes des questions liées à la compréhension du multimédia, auxquelles nous souhaitons répondre dans cette thèse. En suivant cette voie, nous devrons faire face aux défis spécifiques posés par le contenu multimédia, à savoir sa nature multimodale et sa diversité. Outre le flux visuel, les vidéos contiennent souvent du son, des paroles et sont parfois accompagnées de métadonnées telles que des transcriptions, des titres ou des descriptions. Idéalement, un modèle automatique serait capable de combiner les informations provenant de ces différentes représentations, de la manière la plus efficace possible. Le sujet de la multimodalité ainsi que la relation entre ce qui est dit et ce qui est fait (ce qui se passe visuellement) dans une vidéo, sera un thème transversal dans cette thèse.

Pour certaines tâches, comme la création automatique de résumés audiovisuel (qui doit être considérée ici comme la tâche de classification binaire des scènes comme intéressantes ou non), la grande diversité dans les domaines vidéo peut être considérée comme un défi. "Interessant" est est en effet un qualificatif assez flou qui dépend souvent du domaine. Par exemple, les moments intéressants pour la narration d'une série télévisée qui s'étend sur plusieurs épisodes seront différents des moments intéressants d'un match de football. De même, il existe une hétérogénéité dans la quantité et la nature des métadonnées liées à une vidéo : si un film ou une série TV est souvent accompagné d'une profusion de textes supplémentaires tels que des synopsis, des fandom, des critiques ou des articles wiki, ce n'est pas nécessairement le cas pour les vidéos créées par les utilisateurs. Dans cette thèse, nous avons décidé de travailler avec des séries TV et des films ainsi qu'avec des vidéos courtes

générées par les utilisateurs.

C.2 Le Projet MeMAD

Ce travail a été réalisé dans le cadre du projet de recherche MeMAD, financé par l'UE (H2020). L'acronyme signifie "Méthodes de gestion des données audiovisuelles" et son objectif était de développer des méthodes pour une réutilisation et une réaffectation efficaces du contenu audiovisuel multilingue afin de révolutionner la gestion vidéo et la narration numérique dans la radiodiffusion et la production de médias.

La création de MeMAD a été motivée par l'augmentation du volume de données audiovisuelles et la nécessité de les traiter et de les utiliser plus efficacement dans les industries du divertissement, notamment la télévision, le cinéma et les services de streaming. Plus concrètement, les méthodes automatiques basées sur le langage pour la gestion, l'accès et la publication de contenus vidéo afin de faciliter leur réutilisation faisaient partie des principaux axes et objectifs de recherche de MeMAD.

Outre les objectifs du projet, MeMAD a formulé quatre cas d'utilisation :

- Services de fourniture de contenu pour la réutilisation par les utilisateurs finaux/clients grâce à l'indexation des médias et à la description vidéo.
- Création, utilisation, réutilisation et réaffectation de nouvelles séquences et de contenus archivés dans la production de médias numériques grâce à l'indexation des médias et à la description vidéo.
- Amélioration de l'expérience des utilisateurs grâce à l'enrichissement des médias par des liens vers des ressources externes
- Sous-titrage et description visuelles ou audio automatisées. Conversion de la parole et des sons en texte, ainsi que du contenu visuel en texte, dans plusieurs langues de sortie, pour un usage général et pour les personnes sourdes, malentendantes, aveugles et malvoyantes.

Les partenaires du projet MeMad étaient quatre instituts de recherche, l'Université Aalto et l'Université d'Helsinki de Finlande, l'Université de Surrey du Royaume-Uni et EURECOM de France, quatre entreprises, YLE de Finlande, Limecraft de Belgique et Lingsoft plus Lingsoft Language Services de Finlande, ainsi que l'Institut national de l'audiovisuel français. Le projet de recherche a débuté en 2018 et s'est terminé en 2020. Notre contribution au projet s'inscrit dans le cadre (i) du domaine "Analyse automatique de contenu multimodal" qui a développé des outils pour l'analyse multimodale, la description et l'indexation de contenu vidéo (ii) du domaine "Enrichissement des médias et hyperliens" qui est centré sur l'utilisation du traitement du langage naturel et des technologies sémantiques pour prédire quels moments de télévision susciteront l'intérêt des téléspectateurs et comment ces moments devraient être

enrichis.

C.3 Questions de recherche

Comment identifier les moments mémorables de contenu médiatique ?

Le 20 avril 2022, juste avant le débat télévisé [9] entre Emmanuel Macron et Marine Le Pen (les deux candidats du second tour de l'élection présidentielle française de 2022), le journaliste politique Maxence Lambrecq explique à la radio [10], qu'en préparation du débat, une grande partie de l'attention des conseillers politiques d'Emmanuel Macron a été consacrée à "gérer son sourire", à trouver une expression faciale qui évoquerait la convivialité plutôt que l'arrogance. Il justifie cette attention portée au physique et aux gestes en affirmant que les images sont plus facilement mémorisables que les discours. Nous voyons ici que l'identification du type d'indices dont les gens se souviennent est de la plus haute importance pour quiconque veut raconter et contrôler une histoire. Cela inclut traditionnellement des acteurs de différents domaines tels que la publicité, l'éducation, la politique... Maintenant que la création de contenu vidéo s'est démocratisée, nous pouvons facilement imaginer que les utilisateurs de médias sociaux bénéficieraient également de la possibilité de prédire automatiquement le potentiel de mémorisation de leur vidéo, avant de la mettre en ligne. De même, pour ces plateformes digitales, le fait de pouvoir afficher les vidéos les plus mémorables améliorerait leur expérience utilisateur. La mémorisation est ici définie comme la qualité ou l'état d'être facile à retenir. C'est une notion particulièrement intéressante pour la science des données, car contrairement à d'autres concepts associés tels que l'"intérêt", elle peut être mesurée objectivement par des tests de reconnaissance. Après le succès de la prédiction de la mémorabilité des images, la tâche de prédiction de la mémorabilité des vidéos a ensuite été formalisée quelques mois seulement avant le début de cette thèse, en 2018, avec la première édition du *MediaEval Memorability Challenge*. Dans le chapitre 3 de la thèse, nous explorerons ce qui rend une vidéo mémorable, quelles modalités (textuelles, audio, visuelles) sont pertinentes et nous évaluerons les capacités de généralisation de nos approches à d'autres jeux de données.

Comment résumer les histoires de contenu médiatique?

Après avoir exploré la capacité de mémorisation, le chapitre 4 aborde une autre dimension de l'intérêt : nous cherchons à extraire les parties d'une vidéo qui sont *essentielles à l'histoire*. Dans ce contexte, les moments intéressants sont ceux qui sont décisifs pour la narration et résumer devient la tâche de sélectionner automatiquement les scènes qui sont des élé-

[9] Le débat est disponible sur https://www.france.tv/actualites-et-societe/politique/3264511-le-debat-de-l-entre-deux-tours.html

[10] "Edition spéciale : Débat de l'entre-deux tours" sur https://www.franceinter.fr/emissions/le-telephone-sonne/le-telephone-sonne-du-mercredi-20-avril-2022

ments importants de la structure narrative de la vidéo. Après avoir travaillé avec du contenu créé par les utilisateurs dans le chapitre précédent, nous utilisons ici des vidéos créées par l'industrie du divertissement, en nous concentrant sur le résumé des histoires des épisodes de séries télévisées. Comme Bost [36] l'a souligné, les séries TV modernes offrent un cas d'utilisation réaliste pour le résumé narratif, car contrairement aux séries TV classiques composées d'épisodes autonomes, leurs intrigues s'étendent sur de nombreux épisodes. Les séries sont généralement divisées en un ensemble d'épisodes appelés saisons, dont la sortie est annuelle ou semestrielle. Par conséquent, lorsqu'une nouvelle saison sort, les téléspectateurs sont souvent déconnectés de l'intrigue. Bost [36] a constaté que 60% des personnes interrogées ressentaient le besoin de se faire rappeler les principaux éléments narratifs des saisons précédentes avant de regarder la nouvelle. Ce cas d'utilisation est donc un exemple du type de problèmes que le développement d'outils automatiques pour le résumé des intrigues de séries TV, peut résoudre. Comme sous-thème, nous interrogeons spécifiquement l'utilisation de modèles visio-linguistiques et le potentiel des approches non supervisées pour cette tâche.

Comment extraire automatiquement les éléments d'une histoire dans un contenu médiatique ?

Après avoir essayé d'isoler les moments les plus importants de la narration des épisodes de séries télévisées, nous explorons dans le chapitre 5 la compréhension générale des histoires dans les séries télévisées à partir de systèmes automatisés. Étant donné que le résumé du contenu médiatique, plutôt que d'être une tâche autonome, est lié à un large éventail d'autres tâches telles que l'extraction de caractéristiques " liées au contenu ", le développement d'outils d'analyse vidéo liés aux histoires complète directement l'objectif du chapitre précédent. Dans ce chapitre, en utilisant le texte de scénarios, nous demandons spécifiquement comment les tâches de génération de questions-réponses et de classification de texte permettent l'extraction d'éléments spécifiques de l'histoire. L'un des aspects de la compréhension d'une histoire est en effet de pouvoir poser et répondre à des questions significatives et globales sur l'intrigue. Ces questions peuvent porter sur des sujets tels que la relation entre les personnages ou le motif d'une action : pourquoi quelqu'un a-t-il été tué ? Nous examinons si et comment les modèles de langage sont capables de générer de telles questions. En particulier, nous nous interrogeons sur la possibilité d'un système qui s'appuierait à la fois sur la puissance des modèles linguistiques et sur l'explicabilité des bases de données de sens commun.

C.4 Contributions

Les travaux menés au cours de cette thèse ont abouti aux contributions suivantes :

- Contribuer à l'avancement de l'état de l'art en matière de prédiction de la mémorabilité

des médias en participant au *MediaEval Memorability Challenge*. [57, 60, 85, 131] en 2019, 2020 et 2021. Au cours de cette thèse, nous avons approfondi différentes facettes de la prédiction de la mémorabilité, notamment la multimodalité (*comment combiner au mieux des caractéristiques provenant de différentes modalités*), les choix de caractéristiques visuelles, textuelles et audio, ainsi que l'impact de la perplexité comme indicateur de la nouveauté. Nous avons montré que la mémorisation à court terme peut être mieux prédite - nous avons obtenu un score de Spearman de 0,658 sur le jeu de données Memento10K - avec des modèles multimodaux et que la dégradation de la mémoire reste une tâche difficile. Dans cette thèse, nous avons également consacré une attention particulière à l'étude de la robustesse de nos approches en les testant sur un total de 5 jeux de données couvrant une grande variété de genres, y compris des films ou des vines. En particulier, outre les 3 jeux de données de référence, nous avons utilisé deux ensembles de données MeMAD différents contenant des programmes TV provenant de deux fournisseurs de contenu : Yle (*Yleisradio Oy*, la société nationale de radiodiffusion publique de Finlande) et INA (*Institut National de l'Audiovisuel*, un dépôt de toutes les archives audiovisuelles de la radio et de la télévision françaises). Le code est publié sur https://github.com/MeMAD-project/media-memorability

- Développement de PROZE, un modèle pour la classification de textes explicables et guidés par des promptes, qui exploite les connaissances provenant de deux sources : des modèles de langage pré-entraînés et promptes, ainsi que ConceptNet, une base de connaissances de sens commun qui peut être utilisée pour ajouter de d'explicabilité aux résultats. Nous évaluons notre approche de manière empirique et nous montrons comment cette combinaison non seulement obtient des performances comparables à celles de la classification de pointe à zéro coup dans plusieurs domaines, mais offre également des prédictions explicables qui peuvent être visualisées. Une démo est disponible à http://proze.tools.eurecom.fr/.
- Proposition de deux approches non supervisées pour le résumé de séries télévisées. La première est une approche axée sur l'exploitation de contenus rédigés par des fans et par l identification des personnages principaux. Cette approche s'est classée première à la tâche de résumé vidéo TRECVID [16] de 2020. Après avoir sélectionné les plans d'intérêt par le biais d'une étape de reconnaissance des visages, un score de similarité est calculé entre les phrases issues du contenu créé par les fans (synopsis des épisodes d'EastEnders de la BBC provenant de son Fandom Wiki[11]) et les transcriptions. La deuxième approche s'appuie sur la création de grands modèles de langage qui ont permis à la classification de texte Zero-Shot de fonctionner efficacement dans certaines conditions. Nous explorons si et comment de tels modèles peuvent être utilisés pour le résumé de séries télévisées en menant des expériences avec des entrées de texte variables. Notre hypothèse principale étant que les moments intéressants dans les

[11] https://eastenders.fandom.com/wiki/EastEnders_Wiki

récits sont liés à la présence d'événements intéressants, nous choisissons des étiquettes candidates pour être des événements représentatifs de deux genres : le crime et le feuilleton et obtenons des résultats compétitifs. Le code est publié sur https://github.com/alisonreboud/screenplay_summarization et https://github.com/MeMAD-project/trecvid-vsum.

- Étude de l'utilisation de modèles visio-linguistiques et des choix de pré-entraînement pour le résumé supervisé de séries TV. Les modèles visio-linguistiques se sont avérés efficaces pour plusieurs tâches en aval utilisant du texte et des images appariés. Présentés comme agnostiques par rapport à la tâche, nous explorons si et comment ils peuvent être utilisés pour le résumé de séries télévisées en menant des expériences avec des entrées textuelles variées (dialogue et texte scénique à partir de scénarios) et des modèles affinés sur différents ensembles de données. Nous observons que ces modèles génériques, bien qu'ils ne soient pas spécifiquement conçus pour la compréhension narrative, obtiennent des résultats proches de l'état de l'art. Nos résultats suggèrent également que les données non alignées bénéficient également de ce type d'architecture visio-linguistique. Nous fournissons notre implémentation à l'adresse https://github.com/alisonreboud/mmf.

C.5 Organisation de la thèse

Le reste de cette thèse est organisé en quatre chapitres. Nous pouvons récapituler les contributions à cette thèse, vues à travers les trois lentilles de compréhension du multimédia, comme indiqué ci-dessus :

1. Dans le chapitre 2, nous commençons par présenter l'état de l'art sur la compréhension multimédia. Nous commençons par donner un aperçu du côté multimodal et du côté NLP, pendant la période de rédaction de cette thèse. C'est une période qui est définie par deux choses : l'avènement des gros modèles de langage pré-entraînés et l'émergence du prompteur. Nous présentons ensuite les domaines du résumé vidéo et de la prédiction de la mémorisation.

2. Dans le chapitre 3, nous nous plongerons dans la tâche de prédiction automatique de la mémorabilité de vidéos, en montrant que cette tâche bénéficie de l'utilisation d'approches multimodales. Nous testerons les capacités de généralisation de nos modèles en utilisant un total de 5 jeux de données dans cette section. Enfin, nous explorerons de nouvelles pistes telles que la perplexité ou l'explicabilité.

3. Dans le chapitre 4, nous nous concentrons sur le résumé de séries TV avec une approche multimodale et deux approches textuelles non supervisées.

4. Enfin, nous consacrons le chapitre 5 à l'extraction des éléments de l'histoire en développant une méthode de classification de texte zéro adaptée au domaine. Nous commençons également à étudier les possibilités de génération automatique de questions pour la compréhension des histoires.

C.6 Première Partie

La première question traitée est: "Comment identifier les moments mémorables dans le contenu médiatique ?". Travailler avec un indicateur objectivement quantifiable tel que la mémorisation pour la tâche de résumé est intéressant car il limite la subjectivité associée au concept de "moments intéressants". Dans une tentative de formalisation de la notion d'intérêt visuel, Constantin et al. [61] avance que plutôt qu'un concept autonome, il est étroitement lié à de nombreux aspects des perceptions tels que les émotions, l'esthétique ou la mémorisation. La mémorisation, en particulier, a été décrite comme "une propriété intrinsèque des images" [40, 115] en raison de son accord inter-annotateur élevé et a été utilisée pour créer des résumés vidéo [81]. Les deux concepts sont liés mais ne se chevauchent pas : un segment vidéo peut être mémorable sans être essentiel. être mémorable sans être une partie essentielle à inclure dans un résumé [61]. En dehors de la résumé, l'analyse de la mémorisation de la vidéo est en soi pertinente pour de nombreuses applications, telles que la recherche de contenu, l'éducation, la gestion de l'information et la gestion de l'information. applications telles que la recherche de contenu, l'éducation, le résumé, la publicité, le filtrage de contenu et les systèmes de recommandation [60]. et les systèmes de recommandation [60]. D'après l'initiative de benchmarking pour l'évaluation multimédia (MediaEval) : "Des modèles efficaces de prédiction de la mémorisation feront également progresser la compréhension sémantique du contenu multimédia". compréhension sémantique du contenu multimédia". Ces dernières années, le concours MediaEval Memorability Challenge [57, 60, 85, 131], auquel nous avons participé en 2019, 2020 et 2010, a certainement été l'acteur le plus actif dans la promotion de la recherche sur la prédiction automatique de la mémorisation des vidéos. Au cours des trois éditions du défi, nous avons pu explorer différents aspects de la prédiction de la mémorabilité, tels que les méthodes de fusion de différentes modalités, la sélection des caractéristiques (visuelles, textuelles et audio) et le rôle de la nouveauté. En participant à de participer aux défis, nous avons également étudié la robustesse en testant notre approche sur deux ensembles de données MeMAD différents. Ces vidéos MeMAD correspondent à des programmes de radio et de télévision qui proviennent de deux fournisseurs de contenu : Yle (Yleisradio Oy, la société nationale finlandaise de radiodiffusion publique) et l'INA (Institut National de l'Audiovisuel). Au total, nous avons obtenu des résultats pour 5 jeux de données différents ensembles de données différents dans une variété de genres, des vines aux films.

Dans la deuxième partie de cette thèse, nous allons étudier un autre aspect du résumé de

contenu audiovisuel : la narration. Dans 3, nous avons exploré un aspect mécanique du résumé, c'est-à-dire la modélisation de la capacité du cerveau humain à se souvenir d'une scène qu'il a vue précédemment. Dans ce chapitre, nous nous intéressons davantage au développement d'approches capables d'extraire les éléments importants d'un récit. Nous nous concentrons sur le domaine des séries télévisées, en utilisant deux jeux de données de genres différents : le jeu de données CSI qui contient des épisodes d'une série policière et le jeu de données Eastenders de la BBC qui est un feuilleton. Dans ce chapitre, nous explorons également différents types d'évaluation, de la métrique F1 utilisée dans la plupart des travaux sur le résumé vidéo, à une évaluation qui évalue la capacité d'un résumé généré à répondre aux questions sur le *que se passe-t-il* de l'histoire.

C.7 Deuxième Partie

Dans ce chapitre, nous accordons une attention particulière à deux sujets que nous avons signalés dans le chapitre sur les travaux connexes2 comme des thèmes nécessitant une plus grande attention. Tout d'abord, Apostolidis et al. [11] ont suggéré que " les approches qui estiment l'importance en fonction de la modalité visuelle et de la modalité audio de la vidéo " (au lieu de n'estimer l'importance qu'en fonction de la modalité visuelle) seraient une direction importante pour le résumé vidéo. Dans ce chapitre, bien que nous n'utilisions pas directement les caractéristiques audio, nous nous appuyons sur le fait que les scénarios de séries télévisées contiennent à la fois des indications scéniques (expliquant ce qui se passe visuellement) et des transcriptions de discours, pour étudier la correspondance entre " ce qui est dit " et " ce qui est fait " à travers le texte. Deuxièmement, comme la création de résumés vidéo véridiques est un processus qui prend du temps [229], il a été souligné que les méthodes non supervisées sont particulièrement pertinentes. Dans ce chapitre, nous développons deux approches non supervisées différentes : une qui s'appuie sur une mise en correspondance avec le contenu écrit par les fans et une qui s'appuie sur la classification des événements zéro-shot. Nous commençons par présenter une approche supervisée basée sur l'utilisation de modèles visio-linguistiques, pour ensuite présenter nos deux approches non supervisées qui ont notamment été conçues dans le cadre de notre participation à deux éditions du *TRECVid video summarization challenge* (2020 et 2021).

C.8 Troisième Partie

Dans le chapitre précédent, nous avons développé une approche pour le résumé non supervisé de séries TV qui était basée sur la classification d'événements. Dans cette approche, nous avions des catégories nommées mais pas ou trop peu de ressources annotées pour entraîner un classificateur. Nous avons donc utilisé ENTAIL et ZESTE, deux modèles pour la catégorisation

de textes sans annotation. Si les résultats sont encourageants, ces expériences nous ont également permis de pointer du doigt certaines limites de ces deux modèles comme l'absence de capacités d'adaptation au domaine. Lorsque nous avons utilisé, par exemple, le label " tentative ", nous nous sommes rendu compte que nous ne pouvions pas spécifier que ce mot devait être compris dans le contexte des " séries policières " et non dans son sens plus global. Dans ce chapitre, nous avons donc cherché à améliorer la classification des aspects narratifs, en proposant PROZE une nouvelle méthode pour la catégorisation des textes de type "zero-shot" qui permet une adaptation au domaine. Par rapport au chapitre précédent, nous faisons un pas en avant vers la compréhension de la narration en classant les textes en utilisant des types d'aspects narratifs plus fins. Parce que nous voulons construire un système qui n'est pas spécifique aux aspects du crime, nous montrons également que notre modèle est performant sur des ensembles de données provenant de domaines très différents tels que la soie ou les situations d'urgence.

Dans une deuxième partie de ce chapitre, nous poursuivons notre exploration des capacités des modèles de langage lorsqu'ils sont appliqués à des ensembles de données spécifiques à un domaine, en explorant la tâche de génération automatique de questions pour les séries télévisées. Dans le chapitre précédent, nous avions vu que le défi VSUM TrecVID proposait d'évaluer le résumé en évaluant la capacité d'un modèle à répondre à des questions importantes écrites par des humains. Dans ce chapitre, nous étudions si l'écriture des questions importantes peut être automatisée.

C.9

Conclusion et travaux futurs

C.9.1 Résumé de la thèse

Dans cette thèse, nous avons utilisé des techniques issues à la fois du domaine du *NLP et de l'extraction d'information*, et du *analyse de contenu multimodale*, pour aborder un certain nombre de tâches concernant les vidéos et les histoires qu'elles véhiculent. Nous avons abordé différents aspects de la narration dans le contenu multimédia : Qu'est-ce qui rend une vidéo mémorable ? Comment extraire les moments importants pour le résumé narratif ? Comment extraire les aspects de l'histoire à partir de scénarios ? Comment formuler des questions significatives à partir de transcriptions de séries télévisées ? En résumé, la thèse conduit aux contributions suivantes :

- Obtention des meilleurs résultats au banc d'essai de mémorisation MediaEval pendant 3 années consécutives, en exploitant des modèles profonds pré-entraînés et en

combinant différentes représentations de contenu (texte, caractéristiques visuelles, intégration multimodale, audio). Étudier la robustesse de nos modèles de prédiction de la mémorisation en les testant sur 5 jeux de données.

- Nous avons mené une étude qui isole les éléments textuels des scénarios en fonction de la nature de l'information qu'ils véhiculent (dialogue versus information scénique) et nous avons testé différentes méthodes de pré-entraînement sur deux modèles visio-linguistiques pour la tâche de résumé de séries télévisées. Nous avons montré que l'utilisation d'une architecture visio-linguistique sans données appariées et sans pré-entraînement dans le domaine permet d'obtenir des résultats proches de l'état de l'art.

- Démonstration, notamment grâce à notre participation aux défis TrecVID VSUM, de la manière dont la composante textuelle des médias, qui peut être facilement obtenue automatiquement, peut nous aider à aborder la tâche de résumé basé sur les personnages : soit en exploitant les synopsis créés par les fans, soit en utilisant la classification " zero-shot " pour capturer les principaux événements de la vie des personnages.

- A développé un modèle qui s'appuie sur des connaissances externes (connaissances de sens commun à partir de CONCEPTNET et connaissances linguistiques à partir de modèles de langage pré-entraînés) pour effectuer une classification de texte en mode zéro-shot, c'est-à-dire, à partir d'une simple liste d'étiquettes. Nous avons montré que les capacités d'adaptation au domaine de notre modèle sont bénéfiques pour le domaine de la compréhension des histoires dans le contenu multimédia.

- A commencé à explorer le potentiel des modèles de langage pour générer automatiquement des questions à partir de transcriptions de séries télévisées comme moyen de créer des ensembles de données non coûteux pour le résumé basé sur les requêtes.

C.9.2 Travaux futurs

Classification multimodale à zéro Nous avons abordé à la fois le résumé et l'extraction des aspects narratifs à travers l'objectif de la classification de texte par prise de vue zéro. Nous nous sommes limités à exploiter la représentation textuelle des séries télévisées. À l'avenir, il serait bon d'intégrer des indices visuels et sonores dans ce processus. En général, une meilleure intégration des modalités pour la représentation des contenus multimédias. Suivre les progrès réalisés dans le sens d'une représentation sans modalité. Au-delà de la simple amélioration de chaque modalité, les architectures basées sur les transformateurs semblent approcher du point de maturité où elles peuvent être utilisées sur toutes les modalités et être aussi performantes que celles qui sont spécifiques à une modalité (par exemple, CNN pour la vision)[12].

[12] Le premier algorithme auto-supervisé à haute performance qui fonctionne pour la parole, la vision et le texte

Représentation des histoires sous forme de graphes Dans cette thèse, nous avons commencé à pointer du doigt le fait que les représentations profondes offrent des options limitées d'explications. Nous avons vu qu'une solution consistait à travailler avec le graphe de connaissances ConceptNet. De même, nous pensons qu'une direction passionnante serait d'utiliser des modèles profonds sota pour peupler des graphes de connaissances qui représentent explicitement les histoires dans le contenu multimédia. Le défi TRECVid *Deep Video Understanding (DVU) Challenge* [62] explore cette voie en formulant la tâche de *Compréhension de vidéos* comme une extraction de connaissances à partir de toutes les modalités disponibles (parole, image/vidéo et texte) des vidéos pour résoudre différents types de requêtes liées à la compréhension des histoires.

Explorer davantage le lien entre ce qui est dit et ce qui est fait Dans le chapitre 4, nous avons introduit dans le contexte des scénarios de séries télévisées, le sujet de la complémentarité entre ce qui est dit et ce qui se passe visuellement. Nous avons également commencé à étudier la puissance des modèles de génération de texte comme le Text-to-Text Transfer Transformer (T5) [220] sur des tâches telles que la génération de questions. Il serait intéressant de se demander si les modèles de langue permettent de générer des indications scéniques à partir de dialogues uniquement ? Nous pensons que les scénarios offrent une occasion unique d'explorer plus avant le lien complémentaire entre les dialogues et les indications scéniques, dans le but d'aider la génération d'animations mais aussi d'autres tâches liées à la compréhension du lien entre la parole et le visuel dans les histoires [117]. Le lien entre les dialogues et les indications scéniques dans les scénarios a été utilisé pour obtenir des étiquettes faibles pour la reconnaissance des actions [189]. Cependant, la richesse des informations véhiculées par les indications scéniques va au-delà des verbes d'action. Elles contiennent, entre autres, des informations telles que des indications de localisation ou des sentiments. Ces expériences pourraient être particulièrement pertinentes pour la ligne de recherche qui vise à inclure une dimension narrative au sous-titrage vidéo [38, 113].

Bibliography

[1] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. Used: a large-scale social event detection dataset. In *Proceedings of the 7th International Conference on Multimedia Systems*, pages 1–6, 2016.

[2] Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah. Video summarization: techniques and classification. In *International Conference on Computer Vision and Graphics*, pages 1–13. Springer, 2012.

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, 2019.

[5] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing*, pages 3673–3676. IEEE, 2011.

[6] Nouf Ibrahim Altmami and Mohamed El Bachir Menai. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*, 2020.

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015. IEEE.

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE international conference on computer vision*, pages 2425–2433, 2015.

[10] Marta Aparício, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, Ricardo Ribeiro, and Luís Marujo. Summarization of films and documentaries based on subtitles and scripts. *Pattern Recognition Letters*, 73:7–12, 2016.

[11] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Performance over random: A robust evaluation protocol for video summarization methods. In *28th ACM International Conference on Multimedia*, pages 1056–1064, 2020.

[12] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863, 2021.

[13] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, Luis Enrique Erro No, Sta Ma Tonantzintla, and Fabio A González. Gated multimodal units for information fu. *Stat*, 1050:7, 2017.

[14] Jürgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, and Walter Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer vision and image understanding*, 92(2-3):285–305, 2003.

[15] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, et al. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *TREC Video Retrieval Evaluation: TRECVID*, 2019.

[16] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *TRECVID 2020 Workshop*, Gaithersburg, MD, USA, 2020. NIST.

[17] David Azcona, Enric Moreu, Feiyan Hu, Tomás E Ward, and Alan F Smeaton. Predicting media memorability using ensemble models. In *MediaEval 2019: Multimedia Benchmark Workshop*, Sophia Antipolis, France, 2019.

[18] Maria Larsson Lars Backman. Modality memory across the adult life span: evidence for selective age-related olfactory deficits. *Experimental Aging Research*, 24(1):63–82, 1998.

[19] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language, 2022.

[20] Wilma A Bainbridge. Shared memories driven by the intrinsic memorability of items. *Human Perception of Visual Information: Psychological and Computational Perspectives*, 2021.

[21] Wilma A Bainbridge. Shared memories driven by the intrinsic memorability of items. In *Human Perception of Visual Information*, pages 183–206. Springer, 2022.

[22] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *ACL'96)*, pages 86–90, 1998.

[23] Mauro Barbieri, Lalitha Agnihotri, and Nevenka Dimitrova. Video summarization: methods and landscape. In *Internet Multimedia Management Systems IV*, volume 5242, pages 1–13. International Society for Optics and Photonics, 2003.

[24] Madhushree Basavarajaiah and Priyanka Sharma. Survey of compressed domain video summarization techniques. *ACM Computing Surveys (CSUR)*, 52(6):1–29, 2019.

[25] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: The emotional bias. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 491–495, 2016.

[26] André Bazin. What is cinema? vol. i, 2004.

[27] Luca Belli, Sofia Ira Ktena, Alykhan Tejani, Alexandre Lung-Yut-Fon, Frank Portman, Xiao Zhu, Yuanpu Xie, Akshay Gupta, Michael Bronstein, Amra Delić, Gabriele Sottocornola, Walter Anelli, Nazareno Andrade, Jessie Smith, and Wenzhe Shi. Privacy-preserving recommender systems challenge on twitter's home timeline, 2020.

[28] Luca Belli, Sofia Ira Ktena, Alykhan Tejani, Alexandre Lung-Yut-Fon, Frank Portman, Xiao Zhu, Yuanpu Xie, Akshay Gupta, Michael Bronstein, Amra Delić, Gabriele Sottocornola, Walter Anelli, Nazareno Andrade, Jessie Smith, and Wenzhe Shi. Recsys challenge 2020, 2020.

[29] Olfa Ben-Ahmed and Benoit Huet. Deep multimodal features for movie genre and interestingness prediction. In *International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2018.

[30] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[31] Eloïse Berson, Claire-Hélène Demarty, and Ngoc Duong. Multimodality and deep learning when predicting media interestingness. In *Proc. MediaEval 2017 Workshop*, 2017.

[32] James Bigelow and Amy Poremba. Achilles' ear? inferior human short-term and recognition memory in the auditory modality. *PloS one*, 9(2):e89914, 2014.

[33] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993—-1022, 2003.

[34] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[35] Yan-Ying Chen Bor-Chun Chen and Francine Chen. Video to text summary: Joint video summarization and captioning with recurrent neural networks. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 118.1–118.14. BMVA Press, September 2017.

[36] Xavier Bost. *A storytelling machine?: automatic video summarization: the case of TV series*. PhD thesis, Université d'Avignon, 2016.

[37] Xavier Bost, Serigne Gueye, Vincent Labatut, Martha Larson, Georges Linarès, Damien Malinas, and Raphaël Roth. Remembering winter was coming. *Multimedia Tools and Applications*, 78(24):35373–35399, September 2019.

[38] Sabine Braun, Kim Starr, Jorma Laaksonen, et al. Comparing human and automated approaches to visual storytelling. *Innovation in audio description research. London: Routledge*, 2020.

[39] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 09 2021.

[40] Zoya Bylinskii, Michelle A Borkin, Nam Wook Kim, Hanspeter Pfister, and Aude Oliva. Eye fixation metrics for large scale evaluation and comparison of information visualizations. In *Workshop on Eye Tracking and Visualization*, pages 235–255. Springer, 2015.

[41] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[42] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

[43] Ying-Hong Chan and Yao-Chung Fan. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, 2019.

[44] Buru Chang, Hyunjae Kim, Raehyun Kim, Deahan Kim, and Jaewoo Kang. A deep neural spoiler detection model using a genre-aware attention mechanism. In *22nd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 183–195, 2018.

[45] Buru Chang, Inggeol Lee, Hyunjae Kim, and Jaewoo Kang. "killing me" is not a spoiler: Spoiler detection model using graph neural networks with dependency relation-aware attention mechanism. In *EACL*, pages 3613–3617, 2021.

[46] Peng Chang, Mei Han, and Yihong Gong. Extract highlights from baseball game video with hidden markov models. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002.

[47] Jiaao Chen and Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, 2020.

[48] Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, pages 1–1, 2021.

[49] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *22^{nd} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785—794, San Francisco, California, USA, 2016.

[50] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325, 2015.

[51] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Transformers:"the end of history" for natural language processing? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 677–693. Springer, 2021.

[52] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[53] Angelo V Ciardiello. Did you ask a good question today? alternative cognitive and metacognitive strategies. *Journal of adolescent & adult literacy*, 42(3):210–219, 1998.

[54] Michael A Cohen, Todd S Horowitz, and Jeremy M Wolfe. Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, 106(14):6008–6010, 2009.

[55] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2531–2540, 2019.

[56] Romain Cohendet, Claire Hélène Demarty, Ngoc QK Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability. In *Working Notes Proceedings of the MediaEval Workshop 2018*. CEUR-WS, 2018.

[57] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 178–186, 2018.

[58] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[59] Mihai Gabriel Constantin and Bogdan Ionescu. Using vision transformers and memorable moments for the prediction of video memorability. In *MediaEval 2021 workshop*, 2021.

[60] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Helene Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjoberg. The predicting media memorability task at mediaeval 2019. *Proc. MediaEval workshop*, 2019.

[61] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)*, 52(2):1–37, 2019.

[62] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361, 2020.

[63] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.

[64] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[65] Francis Danny and Benoit Huet. L-stap : Learned spatio-temporal adaptive pooling for video captioning. In *First International Workshop on AI for Smart TV Content Production (AI4TV)*, 2019.

[66] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[67] Guoquan Sun Dazhan Xu, Xiaoyu Wu. Media memorability prediction based on machine learning. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[68] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016.

[69] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Duong. Mediaeval 2017 predicting media interestingness task. In *MediaEval Workshop*, 2017.

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171—-4186, Minneapolis, Minnesota, USA, 2019. ACL.

[71] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota, June 2019. ACL.

[72] Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[73] Samuel Felipe Dos Santos and Jurandy Almeida. Gibis at mediaeval 2019: Predicting media memorability task. In *MediaEval*, 2019.

[74] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models, 2022.

[75] Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. On the hidden treasure of dialog in video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2064–2073, 2021.

[76] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[77] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[78] Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference (ESWC)*, pages 177–190, Heraklion, Greece, 2018.

[79] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6363–6372, 2018.

[80] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer, 2018.

[81] Mengjuan Fei, Wei Jiang, and Weijie Mao. Creating memorable video summaries that satisfy the user's intention for taking the videos. *Neurocomputing*, 275:1911–1920, 2018.

[82] Walter R Fisher. Narration as a human communication paradigm: The case of public moral argument. *Communications Monographs*, 51(1):1–22, 1984.

[83] Lea Frermann, Shay B Cohen, and Mirella Lapata. Whodunnit? crime drama as a case for natural language understanding. *ACL*, 6:1–15, 2018.

[84] Junyu Gao, Xiaoshan Yang, Yingying Zhang, and Changsheng Xu. Unsupervised video summarization via relation-aware assignment learning. *IEEE Transactions on Multimedia*, 23:3203–3214, 2020.

[85] Alba García Seco de Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. Overview of MediaEval 2020 predicting media memorability task: What makes

a video memorable? In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[86] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, and Shweta Yadav. Knowledge-infused deep learning. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 309–310, 2020.

[87] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017.

[88] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70, 2019.

[89] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[90] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *23rd International Conference on Machine learning (ICML)*, pages 377–384, 2006.

[91] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 855—-864, San Francisco, CA, USA, 2016.

[92] Antonio Gulli. *AG's corpus of news articles*, 2005.

[93] Himanshu Gupta, Amogh Badugu, Tamanna Agrawal, and Himanshu Sharad Bhatt. Zero-shot open information extraction using question generation and reading comprehension, 2021.

[94] Rohit Gupta and Kush Motwani. Linear models for video memorability prediction using visual and semantic features. In *MediaEval*, 2018.

[95] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520, 2014.

[96] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, et al. Can we measure beauty? computational evaluation of coral reef aesthetics. *PeerJ*, 3:e1390, 2015.

[97] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. Learning computational models of video memorability from fmri brain imaging. *IEEE transactions on cybernetics*, 45(8):1692–1703, 2014.

[98] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE transactions on multimedia*, 7(1):143–154, 2005.

[99] Ijaz Ul Haq, Khan Muhammad, Tanveer Hussain, Javier Del Ser, Muhammad Sajjad, and Sung Wook Baik. Quicklook: Movie summarization using scene-based leading characters with psychological cues fusion. *Information Fusion*, 76:24–35, 2021.

[100] Ismail Harrando, Alison Reboud, Pasquale Lisena, Raphaël Troncy, Jorma Laaksonen, Anja Virkkunen, Mikko Kurimo, et al. Using fan-made content, subtitles and face recognition for character-centric video summarization. In *TRECVID 2020 Workshop*, 2020.

[101] Ismail Harrando and Raphael Troncy. Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph. In *3rd Conference on Language, Data and Knowledge (LDK)*, Zaragoza, Spain, 2021.

[102] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[104] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, Nevada, USA, 2016. IEEE.

[105] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017.

[106] Mohammad Hesham, Bishoy Hani, Nour Fouad, and Eslam Amer. Smart trailer: Automatic generation of movie trailer using only subtitles. In *2018 First International Workshop on Deep and Representation Learning (IWDRL)*, pages 26–30. IEEE, 2018.

[107] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[108] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), feb 2019.

[109] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819, 2011.

[110] Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 580–589, 2021.

[111] Jia-Hong Huang and Marcel Worring. *Query-Controllable Video Summarization*, page 242–250. Association for Computing Machinery, New York, NY, USA, 2020.

[112] Jia-Hong Huang and Marcel Worring. Query-controllable video summarization. In *ICMR*, pages 242–250, 2020.

[113] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.

[114] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 839–842, 2010.

[115] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013.

[116] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011.

[117] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195, 2017.

[118] Janadhip Jacutprakart, Rukiye Savran Kiziltepe, John Q Gan, Giorgos Papanastasiou, and Alba G Seco de Herrera. Essex-nlip at mediaeval predicting media memorability 2020 task. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[119] Sungho Jeon, Sungchul Kim, and Hwanjo Yu. Spoiler detection in tv program tweets. *Information Sciences*, 329:220–235, 2016.

[120] Pin Jiang and Yahong Han. Hierarchical variational network for user-diversified & query-focused video summarization. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 202–206, 2019.

[121] Yudong Jiang, Kaixu Cui, Bo Peng, and Changliang Xu. Comprehensive video understanding: Video summarization with content-based video recommender design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[122] Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom, Dimitri Kartsaklis, Nal Kalchbrenner, Mehrnoosh Sadrzadeh, Nal Kalchbrenner, Phil Blunsom, Nal Kalchbrenner, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 212–217. Association for Computational Linguistics, 2014.

[123] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing, 2021.

[124] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[125] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[126] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015.

[127] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[128] Yoon Kim. Convolutional neural networks for sentence classification. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[129] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

[130] Edwin A Kirkpatrick. An experimental study of memory. *Psychological Review*, 1(6):602, 1894.

[131] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Hélène Demarty, Graham Healy, Camilo Fosco, Alba García Seco de Herrera, Sebastian Halder, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Lorin Sweeney. Overview of The MediaEval 2021 Predicting Media Memorability Task. In *Multimedia Benchmark Workshop (MediaEval)*, 2021.

[132] Ricardo Kleinlein, Cristina Luna-Jiménez, Zoraida Callejas, and Fernando Fernández-Martínez. Predicting media memorability from a multimodal late fusion of self-attention and lstm models. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[133] Ricardo Kleinlein, Cristina Luna-Jiménez, David Arias-Cuadrado, Javier Ferreiros, and Fernando Fernández-Martínez. Topic-oriented text features can match visual deep models of video memorability. *Applied Sciences*, 11(16), 2021.

[134] Ricardo Kleinlein, Cristina Luna-Jiménez, and Fernando Fernández-Martínez. Thaupm at mediaeval 2021: From video semantics to memorability using pretrained transformers. In *MediaEval 2021 workshop*, 2021.

[135] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[136] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. Face recognition systems: A survey. *Sensors*, 20(2):342, 2020.

[137] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[138] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[139] Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *EMNLP-IJCNLP*, pages 540–551, 2019.

[140] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[141] Ken Lang. Newsweeder: Learning to filter netnews. In *12^{th} International Conference on Machine Learning (ICML)*, pages 331–339, 1995.

[142] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231, 2007.

[143] D Le, H Vo, T Nguyen, T Do, T Pham, T Vo, T Nguyen, V Nguyen, and T Ngo. Nii_uit at trecvid 2020. In *TRECVID 2020 Workshop*, 2020.

[144] Phuc H Le-Khac, Ayush K Rai, Graham Healy, Alan F Smeaton, and Noel E O'Connor. Investigating memorability of dynamic media. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[145] Myungji Lee, Hongseok Kwon, Jaehun Shin, WonKee Lee, Baikjin Jung, and Jong-Hyeok Lee. Transformer-based screenplay summarization using augmented learning representation with dialogue information. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 56–61, Virtual, June 2021. ACL.

[146] Jie Lei, Qiao Luan, Xinhui Song, Xiao Liu, Dapeng Tao, and Mingli Song. Action parsing-driven video summarization based on reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2126–2137, 2018.

[147] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965.

[148] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880, 2020.

[149] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[150] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[151] Roberto Leyva, Faiyaz Doctor, Alba Garcia Seco De Herrera, and Sohail Sahab. Multimodal deep features fusion for video memorability prediction. In *CEUR Workshop Proceedings*, volume 2670, 2019.

[152] Baoxin Li and M Ibrahim Sezan. Event detection and summarization in sports video. In *Proceedings IEEE workshop on content-based access of image and video libraries (CBAIVL 2001)*, pages 132–138. IEEE, 2001.

[153] L. H. Li, M. Yatskar, D. Yin, C-J. Hsieh, and K-W. Chang. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[154] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020.

[155] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[156] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing*, 26(8):3652–3664, 2017.

[157] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. Key frame extraction in the summary space. *IEEE transactions on cybernetics*, 48(6):1923–1934, 2017.

[158] Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C-CJ Kuo. Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. *IEEE signal processing magazine*, 23(2):79–89, 2006.

[159] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. *CoRR*, abs/1604.02748, 2016.

[160] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[161] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[162] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[163] Pasquale Lisena, Jorma Laaksonen, and Raphaël Troncy. Facerec: An interactive framework for face recognition in video archives. In ACM, editor, *DataTV-2021, 2nd In-*

ternational Workshop on Data-driven Personalisation of Television, 21-23 June 2021, New-York, USA (Virtual Conference), New-York, 2021.

[164] Chan Liu, Lun Li, Xiaolu Yao, and Lin Tang. A survey of recommendation algorithms based on knowledge graph embedding. In *IEEE International Conference on Computer Science and Educational Informatization (CSEI)*, pages 168–171, Xinxiang, China, 2019.

[165] Chang Liu, Armin Shmilovici, and Mark Last. Towards story-based classification of movie scenes. *PloS one*, 15(2):e0228579, 2020.

[166] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879, 2016.

[167] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[168] Peter J Liu, Yu-an Chung, and Jie Jessie Ren. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders, 2019.

[169] Tengfei Liu, Yongli Hu, Junbin Gao, Yanfeng Sun, and Baocai Yin. Zero-shot text classification with semantically extended graph convolutional network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8352–8359, 2021.

[170] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[171] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. Simplifying paragraph-level question generation via transformer language models. In Duc Nghia Pham, Thanaruk Theeramunkong, Guido Governatori, and Fenrong Liu, editors, *PRICAI 2021: Trends in Artificial Intelligence*, pages 323–334, Cham, 2021. Springer International Publishing.

[172] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[173] Youwei Lu and Xiaoyu Wuez. Cross-modal interaction for video memorability predictions. In *MediaEval 2021 workshop*, 2021.

[174] Jiří Lukavskỳ and Filip Děchtěrenko. Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics*, 79(7):2044–2054, 2017.

[175] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020.

[176] Nianzu Ma, Alexander Politowicz, Sahisnu Mazumder, Jiahua Chen, Bing Liu, Eric Robertson, and Scott Grigsby. Semantic Novelty Detection in Natural Language Descriptions. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 866–882, 2021.

[177] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, 2002.

[178] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *49^{th} Annual Meeting of the Association for Computational Linguistics: Human language technologies*, pages 142–150, Portland, OR, USA, 2011.

[179] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [08.26.2020].

[180] Stephen Mayhew, Tatiana Tsygankova, Francesca Marini, Zihan Wang, Jane Lee, Xiaodong Yu, Xingyu Fu, Weijia Shi, Zian Zhao, and Wenpeng Yin. Karthikeyan k, jamaal hay, michael shur, jennifer sheffield, and dan roth. 2019b. university of pennsylvania lorehlt 2019 submission. Technical report, Technical report, 2019.

[181] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

[182] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[183] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[184] Arthur G Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, 19(2):121–143, 2008.

[185] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[186] Diego Monti, Enrico Palumbo, Giuseppe Rizzo, Pasquale Lisena, Raphaël Troncy, Michael Fell, Elena Cabrio, and Maurizio Morisio. An Ensemble Approach of Recurrent Neural Networks using Pre-Trained Embeddings for Playlist Completion. In *12^{th} ACM Conference on Recommender Systems (RecSys), Challenge Track*, Vancouver, BC, Canada, 2018.

[187] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline, 2019.

[188] Jaap MJ Murre and Joeri Dros. Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, 10(7):e0120644, 2015.

[189] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020.

[190] Judith S Nappi. The importance of questioning in developing critical thinking skills. *Delta Kappa Gamma Bulletin*, 84(1):30, 2017.

[191] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34, 2021.

[192] Nihal Nayak and Stephen Bach. Zero-shot learning with common sense knowledge graphs, 2021.

[193] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *European Conference on Computer Vision*, pages 223–240. Springer, 2020.

[194] E-Ro Nguyen, Hai-Dang Huynh-Lam, Hai-Dang Nguyen, and Minh-Triet Tran. Hcmus at mediaeval2021: Attention-based hierarchical fusion network for predicting media memorability. In *MediaEval 2021 workshop*, 2021.

[195] Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. In *2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2242. Association for Computational Linguistics, 2017.

[196] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7596–7604, 2019.

[197] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Video summarization using deep semantic features. In *Asian Conference on Computer Vision*, pages 361–377. Springer, 2016.

[198] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021.

[199] Enrico Palumbo, Diego Monti, Giuseppe Rizzo, Raphaël Troncy, and Elena Baralis. entity2rec: Property-specific knowledge graph embeddings for item recommendation. *Expert Systems with Applications*, 151, 2020.

[200] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay Summarization Using Latent Narrative Structure. In *ACL*, pages 1920–1933, 2020.

[201] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. In *EMNLP*, pages 1707–1717, 2019.

[202] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie summarization via sparse graph construction. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

[203] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[204] Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online, August 2021. Association for Computational Linguistics.

[205] Letitia Parcalabescu, Nils Trost, and A. Frank. What is Multimodality? In *1st Workshop on Multimodal Semantic Representations (MMSR)*. ACL, 2021.

[206] Anil Singh Parihar, Priyansh Verma, Prashansa Bhattacharyya, and Rohit Goyal. A comparative analysis of query-constrained video summarization techniques. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1165–1170. IEEE, 2021.

[207] Jungin Park, Jiyoung Lee, Sangryul Jeon, and Kwanghoon Sohn. Video summarization by learning relationships between action and scene. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[208] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[209] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[210] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[211] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[212] Maxime Peyrard. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, 2019.

[213] Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 2751–2767. Association for Computational Linguistics, 2020.

[214] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5781–5789, 2017.

[215] Georges Polti. *The thirty-six dramatic situations*. Editor Company, 1917.

[216] A Potyagalova and G. J. F. Jones. Dcu adapt at trecvid 2021: Video summarization - keeping it simple. In *TRECVID 2021 Workshop*, 2021.

[217] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[218] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

[219] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[220] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[221] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[222] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[223] Prajit Ramachandran, Peter J Liu, and Quoc Le. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, 2017.

[224] David Ramsay, Ishwarya Ananthabhotla, and Joseph Paradiso. The intrinsic memorability of everyday sounds. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[225] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphael Troncy, and Hector Laria Mantecon. Combining Textual and Visual Modeling for Predicting Media Memorability. In *Multimedia Benchmark Workshop (MediaEval)*, volume 2670 of *CEUR Workshop Proceedings*, 2019.

[226] Alison Reboud, Ismail Harrando, Jorma Laaksonen, and Raphael Troncy. Predicting Media Memorability with Audio, Video, and Text representations. In *Multimedia Benchmark Workshop (MediaEval)*, volume 2882 of *CEUR Workshop Proceedings*, 2020.

[227] Alison Reboud, Ismail Harrando, Pasquale Lisena, and Raphaël Troncy. Zero-shot classification of events for character-centric video summarization. In *TRECVID 2021 Workshop*, 2021.

[228] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982—-10, Hong Kong, China, 2019. ACL.

[229] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *CVPR)*, June 2019.

[230] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

[231] Dongyu Ru, Zhenghui Wang, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. *QuAChIE: Question Answering Based Chinese Information Extraction System*, page 2177–2180. Association for Computing Machinery, New York, NY, USA, 2020.

[232] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *AI Open*, 2022.

[233] Vasile Rus, Zhiqiang Cai, and Art Graesser. Question generation: Example of a multi-year evaluation campaign. *Proc WS on the QGSTEC*, 2008.

[234] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[235] Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. A deep architecture for multimodal summarization of soccer games. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 16–24, 2019.

[236] Jitao Sang and Changsheng Xu. Character-based movie summarization. In *18th ACM International Conference on Multimedia*, pages 855–858, 2010.

[237] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[238] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, 2021.

[239] Thomas Schleider, Thibault Ehrhart, Pasquale Lisena, and Raphaël Troncy. Silknow knowledge graph, November 2021.

[240] Thomas Schleider and Raphael Troncy. Zero-shot information extraction to enhance a knowledge graph describing silk textiles. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities*

and Literature, pages 138–146, Punta Cana, Dominican Republic (online), November 2021. Association for Computational Linguistics.

[241] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 9:82146–82168, 2021.

[242] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[243] Gayle Seese. Soap opera viewers' perceptions of the real world. Master's thesis, University of Central Florida, 1987.

[244] Debashis Sen, Balasubramanian Raman, et al. Video skimming: Taxonomy and comprehensive survey. *arXiv preprint arXiv:1909.12948*, 2019.

[245] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. ACL.

[246] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018.

[247] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739, 2017.

[248] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[249] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020.

[250] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. arXiv:2004.08744, 2020.

[251] Mats Sjöberg, Hamed R. Tavakoli, Zhicun Xu, Héctor Laria Mantecón, and Jorma Laaksonen. PicSOM experiments in TRECVID 2018. In *Proceedings of the TRECVID 2018 Workshop*, Gaithersburg, MD, USA, 2018.

[252] Alan F Smeaton, Bart Lehane, Noel E O'Connor, Conor Brady, and Gary Craig. Automatically selecting shots for action movie trailers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 231–238, 2006.

[253] Xinhui Song, Ke Chen, Jie Lei, Li Sun, Zhiyuan Wang, Lei Xie, and Mingli Song. Category driven deep recurrent neural network for video summarization. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[254] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.

[255] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[256] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *31st AAAI Conference on Artificial Intelligence*, 2017.

[257] MU Sreeja and Binsu C Kovoor. Towards genre-specific frameworks for video summarisation: A survey. *Journal of Visual Communication and Image Representation*, 62:340–358, 2019.

[258] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations, 2019.

[259] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[260] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[261] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[262] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[263] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[264] Lorin Sweeney, Graham Healy, and Alan Smeaton. Predicting media memorability: Comparing visual, textual, and auditory features. In *MediaEval 2021 workshop*, 2021.

[265] Lorin Sweeney, Graham Healy, and Alan F Smeaton. Leveraging audio gestalt to predict media memorability. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[266] Lorin Sweeney, Graham Healy, and Alan F Smeaton. The influence of audio on video memorability with an audio gestalt regulated video memorability system. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2021.

[267] Lorin Sweeney, Ana Matran-Fernandez, Sebastian Halder, Alba G Seco de Herrera, Alan Smeaton, and Graham Healy. Overview of the eeg pilot subtask at mediaeval 2021: Predicting media memorability. *arXiv preprint arXiv:2201.00620*, 2021.

[268] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[269] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[270] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[271] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[272] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, 2021.

[273] Ronald B Tobias. Master plots: and how to build them. *Writer's Digest Books*, 20.

[274] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE international conference on computer vision (ICCV)*, pages 4489–4497, 2015.

[275] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *International Conference on Computer Vision (ICCV)*, pages 4489–4497, Santiago, Chile, 2015. IEEE.

[276] K.D. Tran, N. P. L. Quang, Do.T, Mai.T, and A.P.N. Truong. Nii_uit at trecvid 2021: Video summarization task. In *TRECVID 2021 Workshop*, 2021.

[277] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. Predicting media memorability using deep features with attention and recurrent network. In *MediaEval*, 2019.

[278] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. Predicting media memorability using deep features and recurrent network. In *MediaEval*, 2018.

[279] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3–es, 2007.

[280] Shubhnkar Upadhyay, Avadhesh Singh, Kumar Abhishek, and M. P. Singh. Deploying a Social Web Graph Over a Semantic Web Framework. In *Computational Intelligence in Data Mining (CIDM)*, pages 73–83, Cap Town, South Africa, 2016.

[281] Patrizia Varini, Giuseppe Serra, and Rita Cucchiara. Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Transactions on Multimedia*, 19(12):2832–2845, 2017.

[282] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590, 2017.

[283] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[284] Alexander Viola and Sejong Yoon. A hybrid approach for video memorability prediction. In *MediaEval*, 2019.

[285] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, 2019.

[286] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. Ruc at mediaeval 2019: Video memorability prediction based on visual textual and concept related features. In *MediaEval*, 2019.

[287] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. Video summarization via semantic attended networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[288] Kenneth Wilhelmsson. Automatic question generation from Swedish documents as a tool for information extraction. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 323–326, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT).

[289] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, pages 1112–1122, New Orleans, Louisiana, June 2018. ACL.

[290] Jianxiong Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010.

[291] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Ziyu Guan, and Deng Cai. Query-biased self-attentive network for query-focused video summarization. *IEEE Transactions on Image Processing*, 29:5889–5899, 2020.

[292] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, 2021.

[293] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016.

[294] Min Xu, Jesse S Jin, Suhuai Luo, and Lingyu Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 677–680, 2008.

[295] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *EMNLP*, pages 3914–3923, 2019.

[296] Junyong You, Guizhong Liu, Li Sun, and Hongliang Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):273–285, 2007.

[297] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[298] Tiezheng Yu, Zihan Liu, and Pascale Fung. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. In *NAACL*, pages 5892–5904, 2021.

[299] Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):226–237, 2017.

[300] Hongxin Zhang and Haitao Liu. Visualizing structural "inverted pyramids" in english news discourse across levels. *Text & Talk*, 36(1):89–110, 2016.

[301] Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[302] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[303] Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. A unified multi-task learning framework for joint extraction of entities and relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14524–14531, May 2021.

[304] Tony Zhao, Irving Fang, Jeffrey Kim, and Gerald Friedland. Multi-modal ensemble models for predicting video memorability. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[305] Kaiyang Zhou, Tao Xiang, and Andrea Cavallaro. Video summarisation by classification with deep reinforcement learning. *British Machine Vision Conf. (BMVC*, 2018.

[306] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3485–3495, 2016.

[307] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.

[308] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015.