Multimodal and Multilingual Understanding of Smells using VilBERT and mUNITER

Kiymet Akdemir[1], Ali Hürriyetoğlu[1,*], Raphaël Troncy[2], Teresa Paccosi[3], Stefano Menini[3], Mathias Zinnen[4] and Vincent Christlein[4]

[1]*KNAW Humanities Cluster DHLab, the Netherlands*
[2]*EURECOM, France*
[3]*Fondazione Bruno Kessler, Italy*
[4]*Pattern Recognition Lab, Friedrich-Alexander-Universität, Germany*

Abstract
We evaluate state-of-the-art multimodal models to detect common olfactory references in multilingual text and images in the scope of the Multimodal Understanding of Smells in Texts and Images (MUSTI) at Mediaeval'22. The goal of the MUSTI Subtask 1 is to classify paired text and images as to whether they refer to the same smell source or not. We approach this task as a Visual Entailment problem and evaluate the performance of the English model ViLBERT and the multilingual model mUNITER on MUSTI Subtask 1. Although base VilBERT and mUNITER models perform worse than a dummy baseline, fine-tuning these models improve performance significantly in almost all scenarios. We find that fine-tuning mUNITER with SNLI-VE and MUSTI train data performs better than other configurations we implemented. Our experiments demonstrate that the task presents some challenges, but it is by no means impossible. Our code is available on https://github.com/Odeuropa/musti-eval-baselines.

1. Introduction

Olfactory information is considered difficult to identify in texts or images. This is mainly due to the relatively rare linguistic evidence documented about its occurrence in texts and its implicit representation in images. Consequently, automating olfactory information extraction in text or images has been attracting considerably less attention [1, 2, 3]. Although novel approaches for multimodal analysis of texts and images have recently been developed, to the best of our knowledge, the olfactory information has not been the focus of any academic work in a multimodal setting.

The Multimodal Understanding of Smells in Texts and Images (MUSTI)[1] task that is organized in the scope of MediaEval 2022[2] fills this gap [4]. The text-image pairs provided by the MUSTI organizers are multilingual – English, German, French, and Italian – and gathered from historical data spanning a period between the 17th and 20th centuries.

We evaluate the performance of two state-of-the-art models, VilBERT [5] and mUNITER [6], on the MUSTI challenge test data and present the performances of base and fine-tuned versions of these models.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
kiymet.akdemir@dh.huc.knaw.nl (K. Akdemir); ali.hurriyetoglu@dh.huc.knaw.nl (A. Hürriyetoğlu); raphael.troncy@eurecom.fr (R. Troncy); tpaccosi@fbk.eu (T. Paccosi); menini@fbk.eu (S. Menini); mathias.zinnen@fau.de (M. Zinnen); vincent.christlein@fau.de (V. Christlein)

CEUR Workshop Proceedings (CEUR-WS.org)
[1]https://multimediaeval.github.io/editions/2022/tasks/musti/, visited on 2022/12/12.
[2]https://multimediaeval.github.io/editions/2022/, visited on 2022/12/12.

Table 1
MUSTI train and test set data statistics. The class distribution of the test data is kept confidential.

	Train		Test	
	Positive pairs	**Total pairs**	**Positive pairs**	**Total pairs**
English	198 (24.90 %)	795	-	200
German	95 (19.79 %)	480	-	213
French	102 (34.00 %)	300	-	200
Italian	198 (24.78 %)	799	-	201

We provide a brief description of the data and a detailed explanation of our methods in Sections 2 and 3, respectively. Next, the results of the models in various configurations are reported in Section 4. Finally, a summary of our evaluation and an outlook concludes this paper in Section 5.

2. Data

The MUSTI data consists of pairs of texts and images that can evoke smell, extracted from historical archives. Text passages cover four languages, namely English (EN), German (DE), French (FR), and Italian (IT). The goal of MUSTI Subtask 1 is to detect whether a text and a paired image evoke the same smell source or not [4]. For instance, a pair consisting of the text *"Some dead acacias were etched against a rose-brown sky. A curious smell of lemons, fur, and flowers filled the room."* and an image containing a lemon or a flower should be classified as YES. The train set consists of 2 374 image-text pairs, out of which 593 are classified as positive, while the test set consists of 814 pairs. Table 1 lists the number of pairs for each language in train and test set, and the number of positive samples in the training set.

3. Method

To evaluate the MUSTI data using state-of-the-art models, we use the VOLTA framework (Visiolinguistic Transformer Architectures) [7], which unifies several BERT-based Vision-Language (V&L) Models built on top of ViLBERT-MT (Vision & Language BERT Multi-Task) [8]. The VOLTA repository[3] contains models pre-trained on their original setup given in their papers and several models pre-trained in a controlled setup. We use ViLBERT [5], pre-trained in its original setup to evaluate English data, and the multilingual model mUNITER [6], pre-trained in the controlled setup to evaluate data in all languages (English, Italian, French, and German).

These V&L Models are pre-trained on Conceptual Captions [9] to perform several V&L tasks such as Visual Question Answering, Visual Entailment, Grounding Referring Expressions, Caption-Based Image Retrieval, etc. It is a standard approach to fine-tune these models on a specific task [7].

In the Visual Entailment task, the goal is to determine, given an image as a premise and text as a hypothesis, whether the premise implies the hypothesis [10]. Models output one of the three labels: entailment, neutral, or contradiction. For the MUSTI Subtask 1, to evaluate if an image and a text pair refer to the same smell object, we evaluate them on the Visual Entailment task. Then, we consider the output as YES if the model outputs entailment and NO if neutral or contradiction since Subtask 1 is a binary classification problem.

[3]https://github.com/e-bug/volta

Table 2
ViLBERT results on the MUSTI English test set, given as F1-macro score.

ViLBERT	ViLBERT -SNLI	ViLBERT -MUSTI	ViLBERT-SNLI-MUSTI
0.4609	0.4373	0.7834	**0.8024**

Table 3
Multilingual models results on the MUSTI test set, given as F1-macro score. The *Overall* score is the F1-macro on all predictions on all test data.

	English	German	French	Italian	Overall
dummy-baseline	0.4285	0.4289	0.3333	0.4273	0.4075
mUNITER	0.4269	0.4289	0.3551	0.4398	0.4177
mUNITER-SNLI	0.4474	0.4644	0.3605	0.5020	0.4473
mUNITER-MUSTI	0.6965	0.4579	0.5022	0.6535	0.6011
mUNITER-SNLI-MUSTI	0.7482	**0.5014**	**0.5053**	0.6850	**0.6176**
Shao et al. [13]	**0.7867**	0.4568	0.3743	**0.7501**	0.6033

To extract features of the images, we use Faster R-CNN [11] with a ResNet-101 [12] backbone that outputs 36 boxes per image following VOLTA. First, we fine-tune ViLBERT and mUNITER on the Visual Entailment dataset SNLI-VE [10] for 20 epochs with a learning rate of 2e-5 and batch size 128. Afterward, we train these fine-tuned models for 10 epochs using the MUSTI train data, splitting 20 % of it as a validation set with a learning rate of 2e-5 and batch size 64. We train ViLBERT only on English train data and the multilingual model mUNITER on the complete MUSTI train data. The parameter sets that yield the best validation score during training are used for inference. From the pre-trained models, we obtain the following models: fine-tuned on SNLI-VE, fine-tuned on MUSTI, and fine-tuned on MUSTI after the SNLI-VE. [4]

4. Results

We present the F1-macro scores calculated for predictions of ViLBERT in Table 2 and dummy baseline, mUNITER models, and proposed method of Shao et al. [13] in Table 3. The pre-trained model mUNITER differs from the dummy baseline at most at 1 point with a very low YES output rate. In particular, it does not predict YES for any DE data and it yields 1, 2, and 5 YES for EN, FR, and IT, respectively. Fine-tuning models on SNLI-VE does not improve scores significantly. The result is not surprising since MUSTI Subtask 1 is not directly a visual entailment task, as the text does not need to describe the image. It is sufficient to have the same smell object in the text and image pair to be classified as YES.

On the other hand, fine-tuning the pre-trained mUNITER on MUSTI data increases the number of YES outputs to 55 for EN, 22 for DE, 85 for FR, and 46 for IT, and the scores increase remarkably. We achieve the best performance when the models are first fine-tuned on SNLI-VE, and then on MUSTI train data. Thus, we got the highest scores on mUNITER fine-tuned on both SNLI-VE and MUSTI, and ViLBERT fine-tuned on SNLI-VE and MUSTI. For the EN data,

[4]Note that reproducing the experiments might lead to different results since the task performance of BERT-based models after fine-tuning heavily depends on the weight initialization seed, such that the minimum and maximum scores differ by 1 or more points across 10 fine-tuning experiments Bugliarello et al. [7]. Furthermore, we fine-tune models using both SNLI-VE and MUSTI, which may increase the variation in the scores.

ViLBERT-SNLI-MUSTI outperforms mUNITER models and the proposed method of Shao et al. [13]. Our best multilingual model mUNITER-SNLI-MUSTI outperforms Shao et al. [13] except for the EN and IT score, while their overall performances are close to each other.

5. Conclusion

We consider the MUSTI Subtask 1, i. e., detecting whether an image-text pair refers to the same smell object, as a Visual Entailment task and propose using the BERT-based multimodal models ViLBERT and mUNITER for this task. We fine-tune models on SNLI-VE to improve the performance on the visual entailment task, and we observe that training further on MUSTI train data boosts performance. ViLBERT-SNLI-MUSTI achieves the highest F1-macro scores on English data, while mUNITER-SNLI-MUSTI achieves the best multilingual performance.

Promising lines of future experiments would be the adaptation of a CLIP [14] model towards the MUSTI task, replacing the visual backbone with more performant architectures such as SWIN [15], trying different ways of fusing visual and textual features, or using more training data for the fine-tuning step.

Acknowledgements

MUSTI is a result of Odeuropa (EU's Horizon 2020 research and innovation programme, grant agreement No. 101004469).

References

[1] S. Menini, T. Paccosi, S. Tonelli, M. Van Erp, I. Leemans, P. Lisena, R. Troncy, W. Tullett, A. Hürriyetoğlu, G. Dijkstra, F. Gordijn, E. Jürgens, J. Koopman, A. Ouwerkerk, S. Steen, I. Novalija, J. Brank, D. Mladenic, A. Zidar, A multilingual benchmark to capture olfactory situations over time, in: Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1–10. URL: https://aclanthology.org/2022.lchange-1.1. doi:10.18653/v1/2022.lchange-1.1.

[2] M. Zinnen, P. Madhu, R. Kosti, P. Bell, A. Maier, V. Christlein, Odor: The icpr2022 odeuropa challenge on olfactory object recognition, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 4989–4994. doi:10.1109/ICPR56361.2022.9956542.

[3] M. Zinnen, P. Madhu, R. Kosti, P. Bell, A. Maier, V. Christlein, Odeuropa dataset of smell-related objects, 2022. URL: https://doi.org/10.5281/zenodo.6367776. doi:10.5281/zenodo.6367776.

[4] A. Hürriyetoğlu, T. Paccosi, S. Menini, M. Zinnen, P. Lisena, K. Akdemir, R. Troncy, M. van Erp, MUSTI - Multimodal Understanding of Smells in Texts and Images at MediaEval 2022, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.

[5] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Curran Associates Inc., Red Hook, NY, USA, 2019.

[6] F. Liu, E. Bugliarello, E. Ponti, S. Reddy, N. Collier, D. Elliott, Visually grounded reasoning across languages and cultures, in: NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future, 2021. URL: https://openreview.net/forum?id=-pKZ0OO-L7l.

[7] E. Bugliarello, R. Cotterell, N. Okazaki, D. Elliott, Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs, Transactions of the Association for Computational Linguistics 9 (2021) 978–994. URL: https://doi.org/10.1162/tacl_a_00408. doi:10.1162/tacl_a_00408.

[8] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, S. Lee, 12-in-1: Multi-task vision and language representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10434–10443. doi:10.1109/CVPR42600.2020.01045.

[9] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565. URL: https://aclanthology.org/P18-1238. doi:10.18653/v1/P18-1238.

[10] N. Xie, F. Lai, D. Doran, A. Kadav, Visual entailment task for visually-grounded language learning, arXiv preprint arXiv:1811.10582 (2018).

[11] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[13] Y. Shao, Y. Zhang, W. Wan, J. Li, J. Sun, Multilingual Text-Image Olfactory Object Matching Based on Object Detection, in: Proceedings of MediaEval 2023 CEUR Workshop, 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.