Enhancing Endoscopic Image Classification with Symptom Localization and Data Augmentation

Trung-Hieu Hoang
hthieu@selab.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

Hai-Dang Nguyen
nguyenhd@eurecom.fr
Eurecom
France

Viet-Anh Nguyen
nvanh160013@ump.edu.vn
University of Medicine and Pharmacy, Ho Chi Minh city, Vietnam

Thanh-An Nguyen
ntan@selab.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

Vinh-Tiep Nguyen
tiepnv@uit.edu.vn
University of Information Technology, VNU-HCM, Vietnam

Minh-Triet Tran*
tmtriet@fit.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

ABSTRACT

Inspired by recent advances in computer vision and deep learning, we propose new enhancements to tackle problems appearing in endoscopic image analysis, especially abnormality finding and anatomical landmark detection. In details, a combination of **Residual Neural Network** and **Faster R-CNN** are jointly applied in order to take all of their advantages and improve the overall performance. Nevertheless, novel data augmentation is designed and adapted to corresponding domains. Our approaches prove their competitive results in term of not only the accuracy but also the inference time in Medico: The 2018 Multimedia for Medicine Task and The Biomedia ACM MM Grand Challenge 2019. These results show the great potential of the collaborating between deep learning models and data augmentation in medical image analysis applications. Especially, more than 4900 bounding boxes localizing the symptom of some classes from KVASIR dataset that we annotated and used in this project are shared online for future research.

CCS CONCEPTS

• **Computing methodologies → Object detection**; *Transfer learning*; **Object recognition**.

KEYWORDS

endoscopic image, symptoms localization, anatomical landmarks detection, object detection, image classification

ACM Reference Format:

Trung-Hieu Hoang, Hai-Dang Nguyen, Viet-Anh Nguyen, Thanh-An Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Enhancing Endoscopic Image Classification with Symptom Localization and Data Augmentation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343031.3356073

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3356073

1 INTRODUCTION

In order to encourage the multimedia research community to tackle the problems of abnormality finding and anatomical landmark detection in human gastrointestinal tract, besides proposing new endoscopic images datasets [10, 11], several challenges are conducted, such as the Medico: Multimedia Task at MediaEval 2018 [9] and The Biomedia ACM MM Grand Challenge 2019 [16]. The goal of these challenges is building efficient endoscopic image classifiers with less training data as possible.

In our approach, we adapted our first prototype [4] that consists of two deep networks, a Residual Neural Network (ResNet) [3] followed by a Faster Region-based Convolutional Neural Network (Faster R-CNN) [13]. As our observation, ResNet image classification module usually fails to classify images that abnormal symptoms appear as small objects on diversity backgrounds. Therefore, we aim to bring the classification problem to symptom localization problem. The system can also propose the location of every abnormal symptom appeared in an input image. Besides improving the overall performance, the information provided makes the classification result more explainable. Additional annotations for symptoms localization training are released for future research.

Additionally, due to the limitation and the imbalance between classes in the training samples and using extra source of endoscopic data is infeasible, we proposed a novel data augmentation mechanism which can enhance the performance of both models. Nevertheless, a multi-tasks classifiers is introduced to reduce the confusion level of classifier module in cases that multiple symptoms of diseases appeared in the same image.

The content of this paper is as follows. In Section 2, related work on endoscopic image analysis are briefly introduced. Details about our proposed method is described in Section 3. Our method is evaluated during two challenges mentioned in Section 4 and the corresponding results are reported in Section 5. Finally, Section 6 contains our conclusions and future work in order to enhance the performance of the proposed system.

2 RELATED WORK

Early approaches that tackle with endoscopic image analysis usually work on the bleeding detection problem. By using color threshold and applying some handcrafted features, Shah et al.[8] and Jung et al.[7] proposed several solutions using color domain and region segmentation. Later, supervised learning machine techniques are used jointly with superpixel saliency to identify bleeding regions of by recognizing blood color patterns [5].

Since introduced, deep neural networks have been used in order to solve several problems in the field of analyzing endoscopic images of the gastrointestinal (GI) tract. Particularly, to localize and identify polyps within real-time constraint, deep CNNs has recently shown an impressive potential when achieving up to 96.4% accuracy - published in 2018 by Urban G et al. [17]. Another interesting article of Satoki Shichijo et al. [15] also applied multiple deep CNNs to diagnose Helicobacter pylori gastritis based on endoscopic images. Further, gastrointestinal bleeding detection using deep CNNs on endoscopic images has been successfully done and published by Xiao Jia et al. [6].

3 APPROACH

3.1 Overview of Proposed Methods

As mentioned before, we aim for using a stacked model consists of a Residual Neural Network (ResNet) and a Faster Region-based Convolutional Neural Network (Faster R-CNN). In the training phase of the Faster R-CNN model, addition information regarding to the location of abnormal symptoms need to be annotated which is described in details in section 3.2.

Additionally, unbalancing between classes makes deep models bias to some classes than others, which reduce the overall accuracy. In the context of these challenges that require using as less as training data as possible, therefore, using extra data is infeasible. In section 3.3, we also provide some augmentation mechanism on the given training dataset in order to provide a better version for the training step of both modules.

Nevertheless, inference time must be taken into account, while the Faster R-CNN needs longer time to process than that of ResNet. Instead of feeding all images through the ResNet module, we should have different strategies as described in section 3.5 that can reduce the number of images need to go through the Faster R-CNN module.

Last but not least, an other improvement of our method is using a multi-task classification architecture that can predict multiple classes at the same time, which can reduce the confusing level of the image classifier model in these difficult cases. Further information about this architecture is presented in Section 3.6.

3.2 From Classification to Symptoms Region Localization

An object detection module can be really useful to detect small abnormal symptoms and diseases, such as *polyps, instruments, dyed-lifted-polyps* and *dyed-resection-margins*. Besides, in some cases that multi-symptoms appear in the same image, identify all of them is necessary to draw the final conclusion. Nevertheless, system that can not only predict the abnormalities but also propose the corresponding positions of that is more reliable and more convenient for endoscopists.

Due to the limitation of the given training dataset, there is no extra information about positions of abnormalities corresponding to each endoscopic image. Therefore, we decide to annotate all the abnormal symptoms in every images of the following classes: *dyed-resection-margins, dyed-lifted-polyps, instruments* and *polyps*. Examples of our proposed bounding box can be seen in Figure 1.

Totally, 4470 images from the KVASIR dataset [11] and the official development dataset were provided by the challenges belonged to the mentioned classes are annotated with 4944 bounding boxes. Although, the number of training samples we annotated is much larger than the number of actual training samples used in training phase, it is still useful for future works. Noticeably, we also publish these annotations online for research community at https://endoscopy.selab.hcmus.edu.vn .

3.3 Data Augmentation

Instruments - the second highest priority class has only 36 training samples. In order to maintain the balancing between all of these classes and also improve the diversity of the *instruments* images, we aim to augment the given dataset by generate more images for the *instruments* based on the current given development set.

From a given image, we carefully crop the region that contains symptom of diseases or *instruments* along their edges. Then, we randomly select images from other classes and use them as the background of the cropped objects. Random affine transformation are also applied. With this strategy, we are able to generate as much as samples we want for the target class. As can be seen in Figure 2, there is not significant visual differences between original and augmented samples.

Impressively, the augmented dataset can be used to enhance the robustness of classification and detection models. Apparently, we can easily get the position of the foreground object on generated images, which can be utilized to augment the training samples of the object detection module (Faster R-CNN). Meanwhile, by adding more positive examples to the training phase, classification models can learn better discriminative features to represent input images.

3.4 Fine-tuning Deep Neural Network on Endoscopic Images

Besides high computational cost, one of the main drawback of deep learning architecture is that it requires a large amount of training data. Moreover, labeled medical data for supervised learning is limited and manual labelling of medical images is a difficult task. Therefore, it requires a lot of effort and time to train the network, which would depend on the size of training data used. However, there is a possible solution to deal with these limitations is using transfer learning, where a pre-trained network on a large dataset (such as ImageNet [14]) is used.

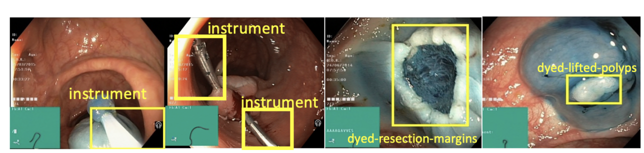

Figure 1: Our proposed symptoms region localization, annotated images with bounding box and class name.

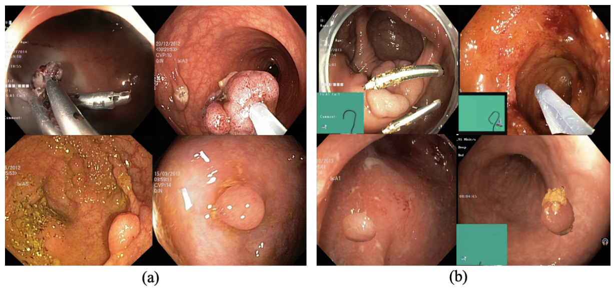

Figure 2: Dataset augmentation result. (a) original samples (b) augmented samples generated by our method.

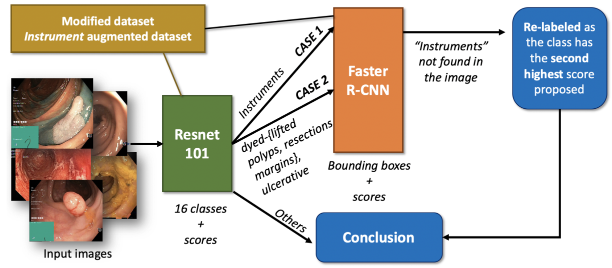

Figure 3: Overview of the the best scenario (3$^{\text{th}}$ scenario).

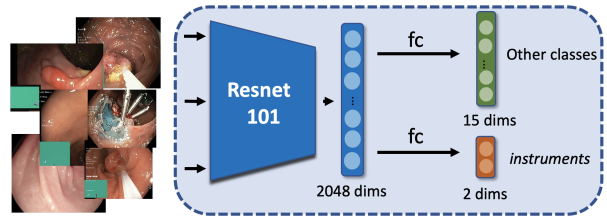

Figure 4: Overview of the multi-classes classifier network.

In our approach, both Residual Network with 101 layers and Faster R-CNN [1] are both share a same features encoder. Therefore, it is necessary to propose an encoder that is specialized on endoscopic images. This is the reason that we fine-tuned our deep neural network models (pre-trained on ImageNet [2]) by using our modified development dataset. After training the whole neural network and then we freeze several first layers in its architecture and fine-tune the remains with small learning-rate. We also tried to train the network from scratch and all of our experiments point out that in term of using convolution neural network for medical images, knowledge transferring from natural images to medical images is possible, even though there is a large difference between the source and the target databases.

This idea is also mentioned in [12], it is especially useful in the case of small dataset of images provided. Fine-tuning on the ImageNet pre-trained model significantly improves the efficient of deep learning model on medical domains.

3.5 Configuration of Conditional Scenarios

There is a trade-off between the inference time and the accuracy. Our goal is finding the most balanced configuration in term of the accuracy and inference time by conducting experiments. The following section briefly describes the pipeline for each of our scenarios that we proposed in [4].

First scenario: instruments, polyps double-checked. After passed through ResNet101, output images classified as special classes become the input of Faster R-CNN network, which is trained for detecting *instruments* in images. In case *instruments* are not detected by Faster R-CNN, they are re-labeled as the class having second highest score proposed by ResNet101.

Second scenario: instruments double-checked. In this second scenario, we limited the images passed through Faster R-CNN by only performing the first case of the first scenario.

Third scenario: instruments double-checked and data augmentation. The configuration is as same as the second scenario and illustrated in Figure 3. In this scenario, we train our model on a combination between re-labeled version and augmented *instruments* set.

Forth scenario: 75% of training set. We reduce the number of training samples by selecting randomly 75% images of each class. Other processing steps are configured in the same as the third scenario.

Fifth scenario: esophagitis priority. Normal-z-line and *esophagitis* are the top most confusing even for human to distinguish them. In this scenario, by adding a small bias 0.3 to the probability of the *esophagitis*, the model is forced to prefer *esophagitis* to *normal-z-line* when it is confused between these classes.

3.6 Solving Multi-classes Problem with Multi-tasks Classifier

After Medico 2018, another improvement is to reduce the confusion level of the classification model in cases that various type of abnormalities appeared in a same image. Instead of using only one classifier and forcing the classification model to follow the priority list in these cases by feeding a number of positive and negative samples, we narrow down this job to multiple classifiers. For instance, given an image that both *esophagitis* and *instruments* appear simultaneously, the job to determine whether or not *instruments* appear inside the image is left for a 2-classes classifier. This classifier can output the probability that the given image contains *instruments*. Meanwhile, the second classifier works independently, which can output the probability for other classes, except *instruments*.

Architecture. In order to reduce the inference time, we decide to share the weights of backbone ResNet 101 for both classifiers. The overview of this module can be seen in Figure 4. In general, the proposed multi-task classification model consists of a ResNet 101 architecture except the last fully connected layer, working as a features extractor module that can output a 2048 dimensions vector for each input image. There are two fully connected branches on top of the output of that features extractor in order to get the prediction of *instruments* class and other 15 *classes*. The number of classifiers is extendable in the future.

Table 1: Official evaluation results of Medico: The 2018 Multimedia for Medicine Task for both sub-tasks (provided by the organizers) and speed (fps) on Tesla K80 GPU.

RunID	PREC	REC	ACC	F1	MCC	RK	FPS
Run01	94.245	94.245	99.281	94.245	93.861	93.590	6.589
Run02	93.959	93.959	99.245	93.959	93.556	93.273	**23.191**
Run03	**94.600**	**94.600**	**99.325**	**94.600**	**94.240**	**93.987**	23.148
Run04	93.043	93.043	99.130	93.043	92.579	92.257	22.654
Run05	94.508	94.508	99.314	94.508	94.142	93.884	21.413

Loss function. Two classifiers are trained simultaneously with the overall loss function is a weighted sum of each of their loss, given as follow

$$\mathcal{L}(p_i, p_o, p_i^*, p_o^*) = \lambda \sum_i \mathcal{L}_{instr}(p_i, p_i^*) + (1-\lambda) \sum_i \mathcal{L}_{others}(p_o, p_o^*) \tag{1}$$

where (p_i, p_i^*) and (p_o, p_o^*) denote the prediction and the ground-truth of *instr* class and other classes, respectively. $\mathcal{L}$ is Cross Entropy Loss function. λ is a combination weight.

Final prediction. Since, there are two output vectors from the model, the final prediction can be determined by

$$y_f = \begin{cases} \text{argmax}(p_o) & p_i < 0.5 \\ y_{instr} & p_i \geq 0.5 \end{cases} \tag{2}$$

where y_f stands for the final prediction of input image, p_o is a 15 dimensions vector indicates the probability which class that image is likely to belong. p_i is the probability that *instruments* appear inside the image and y_{instr} is the label of *instruments* class.

4 EXPERIMENTAL SETUP

Medico: The 2018 Multimedia for Medicine Task

The configuration of *Run*01 to *Run*05 corresponding to five scenarios descried in Section 3.5.

Biomedia ACM MM Grand Challenge 2019

We continue to tackle existing problems that we did not solve efficiently in Medico 2018, which are the confusing between *normal-z-line* and *esoiphagitis*; multi-classes problem with *instruments* class. Besides increasing the size of training data with our augmentation strategy, there are two major improvements in this challenge.

(1) With the multi-classes problem, **we applied the Multi-tasks Classifier** (*MUL*) which is introduced in Section 3.6.
(2) With the *normal-z-line* and *esophagitis*, with a help of a medical expert, we re-annotate the labels of the original dataset and train our models on the modified version of the dataset (*RE_LBL*).

5 RESULTS

As illustrated in Table 1, there is a trade-off between speed and accuracy when comparing the result of *Run*01 and *Run*02. When reducing a large number of images passing through Faster R-CNN, so its performance seems to be relatively worse than *Run*01's. As we mentioned earlier, training samples takes an important role in building a deep-neural network model. Through our experiments, in the case of less training data, the augmented dataset helps us improve the performance of deep-neural network model. *Run*03 and *Run*05 show impressive results comparing to the first two runs. Especially, the *Run*03 has the highest *MCC* score among all team submissions in the Medico 2018 challenge. On the other hand, using the Residual Neural Network cannot classify efficiently the two classes *esophagitis* and *normal-z-line*. The same problem also occurs between the *dyed-resection-margins* and *dyed-lifted-polyps* classes. Additionally, the configuration of *Run*05 intuitively prefers *esophagitis* to *normal-z-line*, which may leads to an increasing of the false-positive cases in the result. By comparison to the others, *Run*04 has the lowest precision since it uses 75% of training data. Decreasing the amount of training samples of course affects the performance in deep-learning models. Nevertheless, the result is still acceptable when it decreases only a few percentages and its configuration is as same as *Run*03.

Table 2: Official evaluation results of The Biomedia ACM MM Grand Challenge 2019 for detection sub-task (provided by the organizers) and speed (fps) on GTX 1080 Ti GPU.

RunID	PREC	REC	F1	MCC	FPS
SIN	85.654	85.034	84.583	91.259	3.5461
MUL	**87.630**	**87.706**	**87.463**	**94.059**	**3.6101**
SIN+RE_LBL	86.674	85.674	84.825	91.681	3.5842
MUL+RE_LBL	86.976	85.732	84.912	92.016	3.5842

Regarding to the evaluation results in Table 2, the performance of the Multi-tasks classifier has successfully proved to have better performance than the single-task classifier used in previous experiments. Although, having competitive results on our own validation set, re-labeling the development dataset on *esophagitits* and *normal-z-line* approach seem to have worse performance on the official test set. Distinguishing these classes is an challenging problem, especially for computers since there are still disagreements between experts in some cases.

6 CONCLUSION AND FUTURE WORK

Endoscopic image classification is a challenging problem because of the fine-grained images, less training data and require high accuracy. In our approach, we focus on enhancing the performance of image classification model, such as Residual Neural Network by using object detection model, such as Faster R-CNN that can localize small symptoms of diseases, which are useful evidences. Besides, data augmentation strategy can be applied to solve the limitation of training samples which is commonly occurred in medical datasets. Accuracy and inference time that we reach is acceptable and appropriate for real-time constraint. However, for future work, additional medical testing must be taken into account besides visual information to create a more robust approach to exploit the distinction between easy-confused classes.

ACKNOWLEDGMENTS

We would like to express our appreciation to AIOZ Pte Ltd for supporting our research team. This research is partially supported by the research funding for the Honors Program of Faculty of Information Technology, and Software Engineering Laboratory, University of Science, VNU-HCM.

REFERENCES

[1] Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Trung-Hieu Hoang, Hai-Dang Nguyen, Thanh-An Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. An application of residual network and faster - RCNN for medico: Multimedia task at mediaeval 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018.*, 2018.

[5] Dimitris Iakovidis, Dimitris Chatzis, Panos Chrysanthopoulos, and Anastasios Koulaouzidis. Blood detection in wireless capsule endoscope images based on salient superpixels. volume 2015, 08 2015.

[6] Xiao Jia and Max Q.-H. Meng. A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016.

[7] Y. S. Jung, Y. H. Kim, D. H. Lee, and J. H. Kim. Active blood detection in a high resolution capsule endoscopy using color spectrum transformation. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 1, pages 859–862, May 2008.

[8] Subodh K Shah, Pragya P Rajauria, Jeongkyu Lee, and M. Emre Celebi. Classification of bleeding images in wireless capsule endoscopy using hsi color domain and region segmentation. 07 2019.

[9] Pål Halvorsen Thomas de Lange Kristin Ranheim Randel Duc-Tien Dang-Nguyen Mathias Lux-Olga Ostroukhova Konstantin Pogorelov, Michael Riegler. Medico multimedia task at mediaeval 2018. *Media Eval' 2018*, 2018.

[10] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Nerthus: A bowel preparation quality video dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 170–174, New York, NY, USA, 2017. ACM.

[11] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[12] Adnan Qayyum, Syed Anwar, Muhammad Majid, Muhammad Awais, and Majdi Alnowami. Medical image analysis using convolutional neural networks: A review. 42, 09 2017.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[15] Aoyama Kazuharu Nishikawa Yoshitaka Miura Motoi Shinagawa Takahide Takiyama Hirotoshi Tanimoto Tetsuya Ishihara Soichiro-Matsuo Keigo-Tada Tomohiro Shichijo Satoki, Nomura Shuhei. Application of convolutional neural networks in the diagnosis of helicobacter pylori infection based on endoscopic images. *EBioMedicine*, 25:106–111, Nov 2017.

[16] Pia Smedsrud Trine B. Haugen Kristin Randheim Randel Konstantin Pogorelov Håkon Kvale Stensland Duc-Tien Dang-Nguyen Mathias Lux Andreas Petlund Thomas de Lange Peter Thelin Schmidt Pål Halvorsen Steven Hicks, Michael Riegler. Acm mm biomedia 2019 grand challenge overview. 2019.

[17] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4), 2018.