Unsupervised Image Segmentation via Self-Supervised Learning Image Classification

Andrea M. Storås[1]
[1]SimulaMet, Norway
andrea@simula.no

ABSTRACT

This paper presents the submission of team Medical-XAI for the Medico: Transparency in Medical Image Segmentation task held at MediaEval 2021. We propose an unsupervised method that utilizes tools from the field of explainable artificial intelligence to create segmentation masks. We extract heat maps, which are useful in order to explain how the 'black box' model predicts the category of a certain image, and the segmentation masks are directly derived from the heat maps. Our results show that the created masks can capture the relevant findings to a certain extent using only a small amount of image-level labeled data for the classification model and no segmentation masks at all for the training. This is promising for addressing different challenges within the intersection of artificial intelligence for medicine such as availability of data, cost of labeling and interpretable and explainable results.

1 INTRODUCTION

Medical image segmentation is one of the focus areas for researchers working on artificial intelligence (AI) and medicine. Especially since the release of U-Net [9], the field has somehow exploded, leading to a myriad of publications of different segmentation approaches within different medical specialisations. One of the areas that get most of the attention is the segmentation of polyps in the colon. An important motivation is that colon cancer is one of the most prominent cancers worldwide and early detection by finding polyps is an efficient method to reduce mortality. The MediaEval Medico challenge 2021 [2] is using this important medical challenge as a task in addition to adding an extra challenge by asking participants to provide as transparent solutions as possible. Transparency within AI is a rather new concept and has several sub-parts that contribute to it ranging from open data, over explainable and interpretable results to open source code. Our solution for solving this year's task is going a step further by also addressing the problem of dealing with a low amount of labeled segmentation data because obtaining accurate labels for the medical data is often difficult due to the availability of medical experts. In the following, we provide a detailed description of our approach, followed by experimental results and a detailed discussion of the advantages and disadvantages of our method.

2 APPROACH

Our method consists of several steps, as illustrated in Figure 1. First, global features are extracted from the images, and clustering is applied for labeling unlabeled medical images. A few labeled images

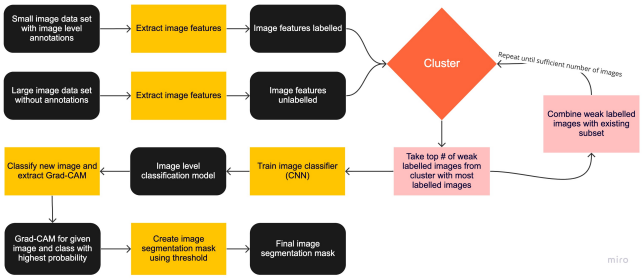

Figure 1: Overview of the complete pipeline of the proposed solution. We start with a small number of labeled and large number of unlabeled images, which are clustered using global image features. The clusters are used to label unlabeled images. This is repeated until we have a sufficient number of labeled images for training a deep neural network. From the resulting model, we extract the Grad-CAM representation from the layers and use a threshold to obtain the segmentation mask.

are clustered together with a high number of unlabeled images from the HyperKvasir data set [1]. Unlabeled images that fall into the cluster with the highest number of labeled images get the same label as the labeled images in the cluster. The process is repeated until a sufficient number of labeled images is reached. The k-means algorithm from scikit-learn [6] is used for clustering, and different numbers of clusters are tested. The k-means algorithm is always initialized with init = 'k-means++', random_state = 0, max_iter = 300 and algorithm = 'auto'. The rest of the hyperparameters are set to default values. An EfficientNet-b1 classifier [4] implemented in Pytorch [5] is trained on the labeled images to predict the correct category.

The Adam optimizer and the BCEWithLogitsLoss from PyTorch with default hyperparameter settings are applied during model training. The classifier is evaluated on images with known labels to measure model performance as well as evaluating the clustering technique for labeling images. Grad-CAM [10] is applied to create heat maps highlighting which pixels the classifier focuses on during classification. Since the model is trained to detect polyps, we expect the heat maps to highlight these as segments. Segmentation masks are constructed from the heat maps. Several thresholds are tested and the most promising values are selected based on visual inspection. All created models and source code can be found publicly online[1].

For each task we submitted five different runs with different configurations of our system. Run 1 applied the development data set provided in the challenge. In order to get images without polyps,

MediaEval'21, December 13-15 2021, Online

[1]https://github.com/kelkalot/Medico-2021-Team-Medical-XAI

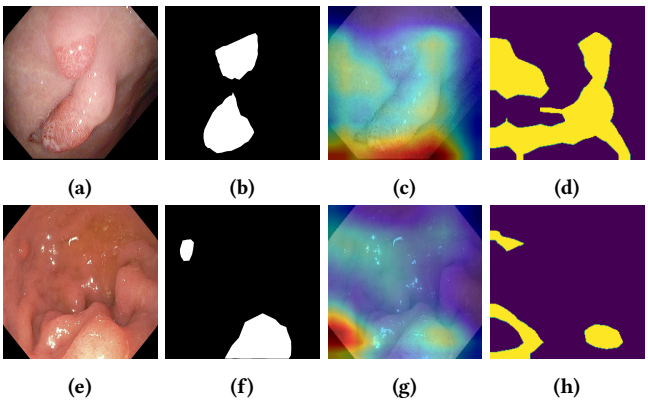

Figure 2: (a, e) Original images, (b,f) ground truth segmentation masks, (c, g) heat maps and (d, h) generated segmentation masks for Run 1 and Run 3.

each image was split into 4 tiles. The corresponding segmentation masks were used to label 100 tiles as 'polyps' and 100 as 'non-polyps'. The rest of the tiles were unlabeled. For Runs 2 - 5 we used 100 images labeled as 'polyps' and 100 images labeled as 'non-polyps' from the Kvasir data set [7]. Unlabeled images from the HyperKvasir data set [1] were included for labeling using the clustering technique described above.

For all runs, the following global features were extracted using the LIRE [3, 8] library: EdgeHistogram, Tamura, LuminanceLayout and SimpleColorHistogram. The internal evaluations of the classifiers were performed on 60 images not used for training. All models were trained for 25 epochs on 1,000 labeled images for a fair comparison. Heat maps were generated by passing the images through the final model. We explored extracting heat maps from several layers in order to identify the most appropriate model. The different configurations are the following: Run 1: 50 clusters, layer 14; Run 2: 200 clusters, layer 13; Run 3: 200 clusters, layer 14; Run 4: 250 clusters; layer 20; Run 5: 250 clusters, layer 22. 'Layer' corresponds to the layer in the model that the heat maps were extracted from. The segmentation masks were constructed from the heat maps. The thresholds were set as > 0.4 and < 0.7 for Runs 1 - 3, while the threshold was > 0.28 for Runs 4 and 5.

3 RESULTS AND ANALYSIS

Segmentation masks for two of the runs are illustrated in Figures 2a - 2h. Table 1 shows the results for Subtask 1, which is polyp segmentation. Overall, we observe that our method does not reach perfect scores, but taking into account that the segmentation is performed unsupervised, it still achieves acceptable results. We also observe that the number of clusters is connected to how good the performance is. This is most probably due to the fact that a higher number of clusters leads to more specialized clusters, which then have a positive effect on the performance of the following model. However, the choice of layer for the heat maps and thresholds for the segmentations could also affect the performance.

Table 1: Results for Subtask 1: polyp segmentation.

Run	Accuracy	Jaccard	Dice	F1	Recall	Precision
#1	0.6762	0.1115	0.1812	0.1812	0.3971	0.1465
#2	0.6009	0.0991	0.1696	0.1696	0.4105	0.1316
#3	0.6018	0.1075	0.1816	0.1816	0.4497	0.1391
#4	0.6402	0.1175	0.1861	0.1861	0.3728	0.1505
#5	0.5214	0.1388	0.2211	0.2211	0.6337	0.1572

Table 2: Results for Subtask 2: inference speed. Abbreviations: Av: average, Mi: minimum, Ma: maximum.

Run	Av-time	Mi-time	Ma-time	Av-fps	Mi-fps	Ma-fps
#1	0.11	0.09	0.12	9.34	10.72	8.05
#2	0.11	0.10	0.13	8.98	9.97	7.75
#3	0.11	0.10	0.13	8.95	10.10	7.94
#4	0.09	0.09	0.09	11.00	11.19	10.54
#5	0.09	0.09	0.09	10.97	11.14	10.71

In terms of how fast the inference is performed, we actually observe that the model based on the larger number of clusters is faster than the others. The reason for this is not clear, although it seems that higher performance is connected to faster inference time. This needs to be investigated more in future work.

4 DISCUSSION AND CONCLUSION

To cluster the medical images, we used global features, which are easy to interpret and increase the transparency of our system. The heat maps are useful in order to explain how the 'black box' model predicts the category of a certain image, and the segmentation masks are directly derived from the heat maps. We believe the level of transparency of our system is quite high. The overall performance of the approach is for sure on the lower scale, but taking into account that it is completely unsupervised segmentation, it can still be considered as good. Overall the presented method seems promising for medical applications and opens up several directions of future work.

5 FUTURE WORK

1,000 images were used to train the classifiers. For future work, we will test how the performance changes with an increasing number of images. Moreover, we want to explore other global features for clustering the images as well as deep features. Other clustering algorithms should also be tested. By connecting the knowledge we have about the global features with the deep features, we might get an interpretation about what features (color, texture) are important for a certain disease. We will also look into other techniques for generating segmentation masks from heat maps. We plan to test the system on other types of medical data sets, including non-image data.

6 ACKNOWLEDGEMENTS

I thank Michael A. Riegler for the assistance with experiments, methodology and writing.

REFERENCES

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Eskeland Sigrun L, Kristin Ranheim Rand l, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Stensland Håkon K, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* 7, 1 (2020), 283. https://doi.org/10.1038/s41597-020-00622-y

[2] Steven Hicks, Debesh Jha, Vajira Thambawita, Hugo Hammer, Thomas de Lange, Sravanthi Parasa, Michael Riegler, and Pål Halvorsen. 2021. Medico Multimedia Task at MediaEval 2021: Transparency in Medical Image Segmentation. In *Proceedings of MediaEval 2021 CEUR Workshop*.

[3] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. Lire: open source visual information retrieval. In *Proceedings of the 7th International Conference on Multimedia Systems*. 1–4.

[4] Luke Melas-Kyriazi. 2019. EfficientNet-Pytorch. https://github.com/lukemelas/EfficientNet-PyTorch. (2019).

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[7] 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. 2017. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, New York, NY, USA, 164–169. https://doi.org/10.1145/3083187.3083212

[8] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'how' reflects what's what: content-based exploitation of how users frame social images. In *Proceedings of the 22nd ACM international conference on Multimedia*. 397–406.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.