Predictive Uncertainty Masks from Deep Ensembles in Automated Polyp Segmentation

Felicia Ly Jacobsen[1]
[1]SimulaMet, Norway
f.l.jacobsen@fys.uio.no

ABSTRACT

This paper presents the submission of team F-HOST for the Medico: Transparency in Medical Image Segmentation task held at MediaEval 2021. We propose a U-Net-based ensemble model for solving the automatic polyp segmentation task and interpret the predictions using a specific method for obtaining uncertainty. Our predicted segmentation masks show a mean Dice score of 45.01% based on the test data. The corresponding uncertainties show systematic errors towards the training data, which indicates overfitting.

I INTRODUCTION

Polyps are abnormal growths inside the lining of the colon or rectum. They can potentially develop into being malignant, leading to colorectal cancer, and thereby act as a precursor for cancer. Detecting and removing polyps with colonoscopic polypectomy during or before further development, will allow for more treatment options and overall improved prognosis [11].

Currently, the gold standard of finding and removing polyps is through a procedure called colonoscopy. This procedure is dependent upon differences in skill, experience, and technique of the endoscopists. However, studies show that up to 28% remain undetected [8]. Automated semantic segmentation based on deep learning frameworks can be used as a tool to detect polyps based on images from colonoscopy examinations. Deep Ensembles can provide an uncertainty quality of the predicted segmentation, even for ensembles with five trained models [7]. This method is known as being easy to implement and being scalable to different deep learning (DL) frameworks and can additionally improve classification error and robustness in terms of dataset shift. In this paper, the results based on the challenge test data are presented and discussed, including their corresponding uncertainty mask estimated from a Deep Ensemble model consisting of five U-Net networks.

II APPROACH

In this section, the approach to the Medico task "Transparency in Medical Image Segmentation" of the MediaEval 2021 challenge is presented. All models were trained using the PyTorch framework [9] on an Nvidia Tesla V100 32GB General-Purpose Graphics Processing Unit (GPGPU).

II.1 Datasets

There is a total of $1,362$ images in the development dataset [5]. We randomly select 272 for validation and the rest for training. The test data only consist of a total of 200 images, excluding the ground truth masks. The dataset is based on the HyperKvasir dataset [2], but includes additional images and masks.

II.2 Experimental Setup

We used the U-Net architecture as the base model for the Deep Ensemble, with a total of five U-Nets. The development data was resized into 256×256 pixels before training, due to memory constraints and to reduce training time. The training data was split into batches of 32 images in order to obtain greater training efficiency as opposed to a larger batch size of, e.g., 64. Data augmentation was performed on the fly for each training iteration in order to obtain improved generalization. We use techniques such as blurring, color jitter, horizontal flip, random rotate 90°, and vertical flip. Instead of using transposed convolution in the decoder part of the network as proposed in the original U-Net paper [10], two-dimensional bilinear upsampling is used in order to avoid potential checkerboard artifacts. All models in the ensemble were trained using an initial learning rate of $1 \cdot 10^{-4}$, with a learning rate scheduler with a minimum learning rate of $1 \cdot 10^{-7}$. Each model had a total of 150 training iterations, using the Adam optimizer [6] and the Dice coefficient loss. After the last training iteration, the model weights for each model in the deep ensemble was saved in a *.pt* format. Hyperparameter tuning was done manually by observing the dice loss on the validation data as a function of training iterations, and evaluating the Dice Coefficient (DE), Jaccard Index (JI) and Accuracy.

When performing prediction with the deep ensemble, each individual model is loaded, and each predict on the input image from the test dataset. The element-wise mean is calculated from the output from each of the models in the ensemble. They are later pushed through a Sigmoid activation and thresholded into binary pixel values. The variance provided by the ensemble is used as an approximation for the uncertainty of each prediction mask. This is calculated by taking the squared sum of each probability prediction (Sigmoid output) minus the mean probability prediction from the ensemble. This squared sum is later divided by the number of models in the ensemble, five in this case.

For subtask 2: "Algorithm Efficiency", the time in seconds was calculated for the ensemble to make its overall mean prediction for each of the test images in order to measure the model efficiency of the ensemble. A Docker image is made, and using this image will make a *.csv* file with the image name and its corresponding prediction time in seconds. The Deep Ensemble will be run on the challenge organizers' hardware, and they provide us with the frames per second (FPS), which is the average number of masks from the test dataset the ensemble is able to make per second.

MediaEval'21, December 13-15 2021, Online

For subtask 3: "Transparent Machine Learning Systems", all source code is made publicly available on GitHub[1], which also includes the uncertainty images for the prediction masks.

III RESULTS AND ANALYSIS

Table 1 summarizes the results for the Medico subtask 1, including the mean DC, mean JI and mean Accuracy for the prediction masks on the validation data and the official task test data. These results show that the Deep Ensemble generalize poorly onto the test data, with a decrease of approximately 55% in the DC score and 46% decrease in the mean JI when comparing the results from the validation data on the test data. There is a high variance of DC score in the individual images from the test images, some get a DC as high as 0.8935, whereas some images get as low as 0.0000. Higher performance can be increased by performing more hyperparameter tuning, training the Deep Ensemble on more training examples including similar datasets such as for example the CVC-ClinicDB dataset [4] and the CVC-ColonDB dataset [1]. Additionally, decreasing the number of training iterations can also contribute to a more generalized ensemble model. Also, as proposed in the original paper [7], adding adversarial training and increasing the number of models in the ensemble from 5 to 15, may potentially decrease the prediction error significantly.

Table 1: Results from validation data and test data by the ensemble model of five U-Nets. The results on predicted test data were provided by the task organizers.

	Mean Dice	Mean Jaccard Index	Mean Accuracy
Validation data	0.8226	0.7005	0.9242
Official Test data	0.4501	0.3231	0.8831

For the efficiency subtask, a FPS of 82.9496 was obtained. This means that the time of approximately 2.4111 seconds in total was used to generate the masks on the entire test dataset. This result indicates satisfactory model efficiency, but in return the deep ensemble is both memory- and time consuming to train.

A set of three randomly chosen images from the test data and their corresponding prediction masks and uncertainty heatmaps are shown in Figure 1. The brighter areas in the heatmaps illustrate the pixels where the models in the ensemble disagree the most. These results show that the borders of the detected polyps are where they disagree the most. Furthermore, the two uncertainty heatmaps (from the left) shows an outlining of a rectangle in the bottom left corner. Many of the input images in the HyperKvasir dataset show green rectangles located in the same area, this is information important to the medical experts. Thus, it is common to observe several images with green rectangles in the development dataset. However, note that the input images do not contain these green rectangles. These results indicate that the ensemble expected these rectangles, thus showing systematic bias towards the training data. Increasing the number of training examples, as well as performing corrections to training images where these rectangles appear by, e.g., cropping them out may boost model performance.

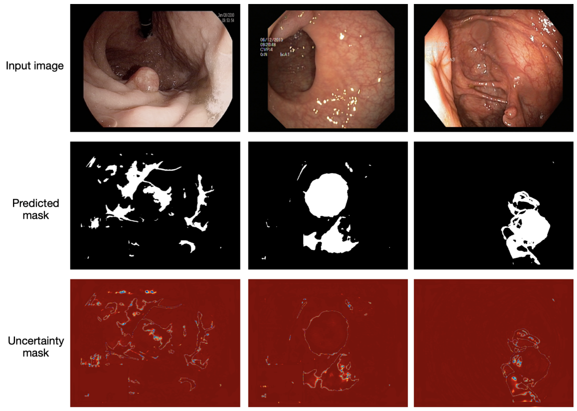

Figure 1: Examples of the input images from the official test dataset are shown on the top row. Their corresponding predicted masks are shown on the middle row, and their uncertainty heatmap representation are on the bottom row. The prediction masks and uncertainty heatmap are calculated using the Deep Ensemble of five trained U-Net networks.

IV CONCLUSION AND FUTURE WORK

In this paper, we presented a method of obtaining the approximate uncertainty values for a set of predicted segmentation masks. The uncertainty masks provide an uncertainty measure of the performance of a U-Net based DL model trained on medical colonoscopy images of polyps.

A mean Dice score of 0.4501 was obtained on the test data, and compared to the Dice score of 0.8226 from the validation data, this indicated that the Deep Ensemble model was being overfitted to the training data, and thus generalizing poorly onto unseen data. Increasing the number of training examples by including similar datasets, decreasing the number of training iterations, increasing the number of models in the ensemble, as well as including adversarial training may improve generalization. A total average FPS of 82.9496 was obtained on the test data, but came at a high computational cost when training the Deep Ensemble. In future work, we will add the aforementioned proposed extensions, as well as experiment and compare to alternative methods such as Masksembles [3] in order to decrease computational cost of obtaining an ensemble model.

REFERENCES

[1] Sánchez J. Vilarino-F Bernal, J. 2012. Towards automatic polyp detection with a polyp appearance model. *Endoscopy* (2012), 3166–3182.

[2] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, 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

[1] https://github.com/feliciajacobsen/MediaEval2021

[3] Nikita Durasov, Timur M. Bagautdinov, Pierre Baqué, and Pascal Fua. 2020. Masksembles for Uncertainty Estimation. *CoRR* abs/2012.08334 (2020). https://arxiv.org/abs/2012.08334

[4] Bernal J. López-Cerón M. Córdova H. Sánchez-Montes C. Rodríguez de Miguel C. Sánchez F. J. Fernández-Esparrach, G. 2016. Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps. *Endoscopy* (6 2016), 837–842.

[5] 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*.

[6] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. (12 2017), 6405–6416 pages. arXiv:stat.ML/1612.01474

[8] Kim NH, Jung YS, Jeong WS, and et al. 2017. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal research* 15, 3 (6 2017), 411–418. https://doi.org/10.5217/ir.2017.15.3.411.

[9] 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*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 234–241.

[11] Winawer SJ, Zauber AG, Ho MN, and et al. 1993. Prevention of colorectal cancer by colonoscopic polypectomy. *The New England Journal of Medicine* 329, 27 (12 1993), 1977–1981. https://doi.org/10.1056/NEJM199312303292701