Mask Classification-based method for Polyps Segmentation and Detection

Mariia Kokshaikyna , Yurii Yelisieiev and Mariia Dobko

The Machine Learning Lab, Ukrainian Catholic University, Lviv, Ukraine

Abstract

We introduce a mask classification model with a transformer decoder for polyps segmentation in endoscopy images. Our novel approach combines custom data pre-processing, a modified mask classification network, test time augmentations, and connected-component analysis. We show the successful performance for polyp semantic segmentation and detection tasks in EndoCV 2022 challenge.

1. Introduction

Endoscopy is a widely used procedure for detecting and diagnosing multiple diseases. Computer-aided endoscopic image analysis and decision support systems can help doctors with diagnosis and increase its effectiveness. Such systems are mainly used to detect, localize, and segment cancer precursor lesions, also called "polyps." EndoCV challenge [1, 2, 3, 4] aims to tackle the generalizability aspect of such methods. In 2022, it has two sub-challenges (Endoscopy artefact detection) EAD 2.0 and (Polyp generalization) PolypGen 2.0. Both tracks set detection and segmentation tasks on a diverse population dataset. This work describes our solution to the EndoCV 2022 challenge on the polyp segmentation and detection tracks.

The dataset of EndoCV 2022 challenge [1, 2, 3, 4] is diverse and comprises images from various endoscope types. This presents an additional difficulty to any computer-aided system. We decided to simplify the input by cropping out the uninformative part and generalizing the input image at pre-processing step.

Standardly, the semantic segmentation task is solved as a per-pixel classification problem, applying a classification loss to each output pixel. An alternative approach is mask classification which, instead of classifying each pixel, predicts a set of binary masks, each associated with a single class prediction. Authors of MaskFormer [5] proposed a modern approach last year by using mask classification to solve both semantic- and instance-level segmentation tasks in a unified manner. This model predicts a set of binary masks corresponding to a single global class label. MaskFormer [5] outperforms per-pixel classification baselines on natural scenes.

We propose to use a mask classification-based method for polyp segmentation in endoscopy data. We are the first to test this model on endoscopic images to our best knowledge. We also customize parts of MaskFormer [5] architecture and show its successful performance for polyp detection.

To increase the robustness of our solution, we add test time augmentations (TTA) and perform connected-component analysis (CCA).

Our contribution can be summed up as follows:

- Evaluated and showed the performance of mask classification method - MaskFormer on endoscopy data. Added custom modifications that improve results of MaskFormer for polyp segmentation.
- Presented a step-by-step pre-processing mechanism for training and inference
- Tested the impact of different loss functions
- Added custom post-processing using test time augmentations and connected-component analysis

2. Data Pre-processing

The PolypGen2.0 subchallenge dataset consists of 46 sequences with 3348 images with polyp labels. Different endoscopes produced these images with various sizes and artifacts - black section located at the left part of the image, blue rectangle with endoscope position, text artifacts, and others. Overall, we can distinguish 15 types of images among these sequences. Statistics about different types is shown on Fig 2.

For train and validation set, we divided sequences into groups using mannually labeled endoscope image types. For validation set we selected sequences *seq1, seq1_endocv22, seq2_endocv22, seq3, seq3_endocv22, seq5_endocv22,*
seq7_endocv22, seq10, seq13_endocv22, seq14_endocv22,

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

mariia.kokshaikyna@ucu.edu.ua (M. K.); yelisieiev@ucu.edu.ua (Y. Yelisieiev); dobkom@ucu.edu.ua (M. D.)

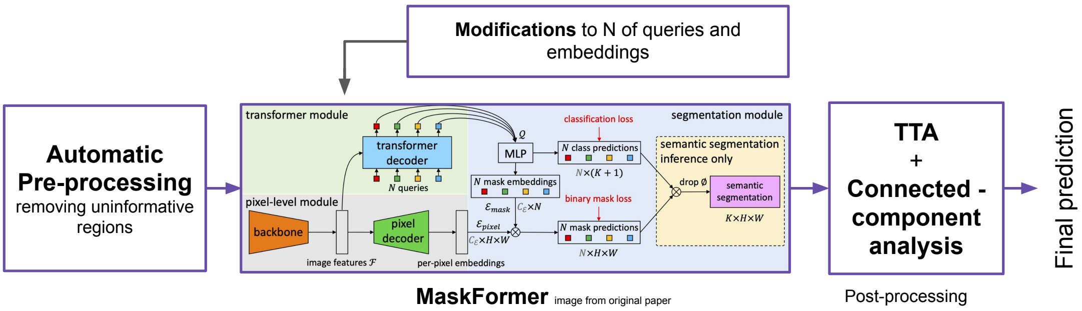

Figure 1: Our Pipeline. The main stages include: pre-processing, MaskFormer with modified queries, post-processing via test time augmentation and connected-component analysis

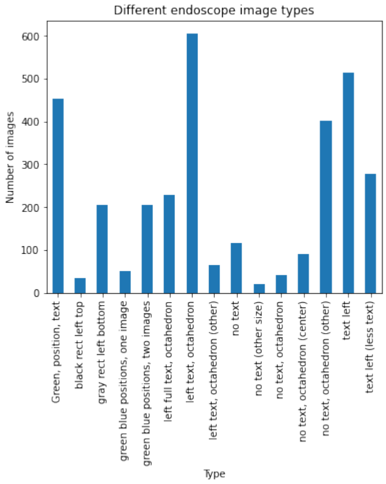

Figure 2: Different endoscope image types

seq15, seq17, seq19_endocv22, seq21_endocv22, and *seq24_endocv22.* Other sequences were used in training set. Overall, our train set contains 3306 images and validation set contains 649 images, which is 19,63% of total image number.

To bring all images to the same view and use the most informative regions during training, we make simple preprocessing and automatically crop images cutting black areas on the left and right sides of the input. To do that, we take the center row of the image, sum up values of RGB channels in this row and use a threshold equal to 48. Continuous left and right parts under this threshold are considered redundant and cut. Examples of cropped images are shown in Fig 3. This cropping improves the informativeness of images and model generalization.

3. Method

We chose MaskFormer [5] as the primary model for our approach. MaskFormer approaches the problem of semantic segmentation as a classification of masks. This approach is an alternative to the per-pixel classification, which predominates in semantic segmentation problems. Instead of classifying each pixel separately, mask classification approaches disjoins the process of semantic segmentation into a division of the image into regions and classification of these regions. Such an approach is general enough to solve semantic and instance segmentation problems. MaskFormer is divided into three modules: pixel-level, transformer, and segmentation.

3.1. Pixel-level module

This module is an encoder-decoder architecture typically used for the semantic segmentation task. The encoder part (a backbone) generates a high-level feature representation of the image. Further, we obtain pixel-level embeddings by iteratively upsampling feature representation from the encoder. Since this is a typical problem setting for a per-pixel classification semantic segmentation task, any model of this type can be plugged into this module.

3.2. Transformer module

Transformer module generates N learnable positional embeddings (i.e., queries) as in DETR[6], which encodes global information about each segment of MaskFormer prediction. This module architecture is adapted from transformers[7], popular for sequence data. In contrast to the standard transformer architecture, each object is decoded in parallel. In Transformer module each output is predicted in an autoregressive manner. The attention mechanism encodes information about the relation of

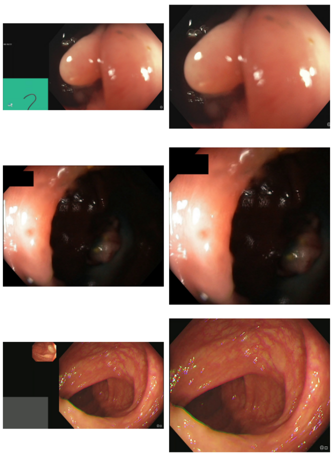

Figure 3: Examples of cropped images. First column - before, second - after our pre-processing procedure

these segments and enhances them with the image context.

3.3. Segmentation module

The segmentation module utilizes a linear classifier and a softmax activation function to acquire class probabilities from each query. Note that we have only two distinct categories of object and no object in the case of the EndoCV PolypGen subtask. An MLP with two hidden dimensions converts queries into mask embeddings for further conversion. The dot product between mask embeddings and per-pixel embeddings is used to calculate mask predictions.

3.4. Model training

We need one-to-one correspondence between ground truth labels and predictions to calculate losses. This problem is solved as in DETR via bipartite matching. Mask and class predictions are used instead of bounding boxes to calculate costs.

Model training given matching is performed by utilizing mask classification composed of cross-entropy classification loss and a binary mask loss.

$$\mathcal{L}_{\text{mask-cls}}\left(z, z^{\text{gt}}\right) = \sum_{j=1}^{N}\left[-\log p_{\sigma(j)}\left(c_j^{\text{gt}}\right) + \right. \\ \left. 1_{c_j^{\text{gt}} \neq \varnothing} \mathcal{L}_{\text{mask}}\left(m_{\sigma(j)}, m_j^{\text{gt}}\right), \right. \tag{1}$$

where mask loss is a linear combination of dice and focal loss as in MaskFormer.

Since we exploited MaskFormer for binary segmentation, most ground truth classes for each query will be zero, and the cross-entropy loss will rapidly converge to zero. Therefore we changed cross-entropy loss to focal loss to mitigate class imbalance in the classification. We have experimented with Boundary loss, which showed promising results in other medical imaging tasks. For our results with this loss, refer to Section 5.2.

3.5. Our modifications

The MaskFormer's transformer module's has an ability to reason about connections in different localities of the image and make distinct predictions for each segment. The model was designed for such large datasets as ADE20k and COCO-Stuff-10k. Whereas the challenge dataset is small compared to them, some model hyperparameters were changed to increase the performance and generalizability of our model. We decreased the number of queries from 100 to 50, FC layers dimensionality from 2048 to 24, and pixel- from 256 to 64. We use a standard convolution ResNet backbone (R50 with 50 layers) instead of SWIN because transformer backbones have poor performance on datasets with few samples and this was proven in our experiments as well. We use the same pixel decoder as described in [5]. Normalization coefficients were recalculated for the PolypGen dataset.

4. Post-processing

Test time augmentation is widely used to increase the model's robustness in deep learning. This procedure makes the final prediction by averaging the predictions after several separately performed augmentations. Our TTA includes horizontal and vertical flips, rotations for 90 and 180 degrees, scaling the input from original size down to 50% of the original size.

Connected-component analysis We perform connected-component analysis of predicted labels during inference. The algorithm divides the segmentation mask into components according to the given connectivity. CCA can have 4 or 8-connected-neighborhood. We remove all smaller parts from the prediction based on the largest connected component.

5. Experiments

We compare our approach against CaraNet [8] which is one of the state-of-the-art methods for polyp segmentation. This model has proven to be effective on many endoscopy datasets including Kvasir-SEG [9]. On this challenge, however, CaraNet with default parameters shows a good Precision score of 0.6041, but much worse Dice than our proposed solution, refer to Table1. In our experiments MaskFormer is capable of capturing more cases of polyps presence.

Table 1
Metrics on our local validation set. MF stands for MaskFormer.

Method	Dice	Dice std	Type2 error
CaraNet	0.37516	0.31954	0.71444
MF	0.73587	0.30823	0.28758
MF + TTA + CCA	0.75717	0.32518	0.27494

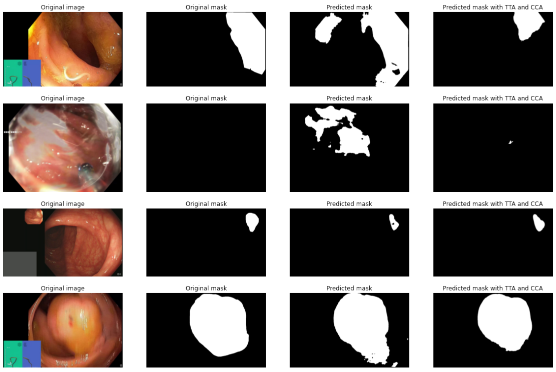

Figure 4: Examples of images where TTA and CCA improved predicted masks

5.1. TTA and CCA impact

TTA and CCA impact on result on our validation set is provided in Table 1. We observe that TTA and CCA in most cases help to decrease false positive regions. For images where TTA and CCA improved predicted masks, see Fig. 4.

5.2. Boundary loss

We use a combination of cross-entropy classification loss and a binary mask loss for each predicted segment during training. The binary loss is a linear combination of focal, and dice losses [10]. We also experimented with other losses. Boundary loss [11] was initially proposed for highly unbalanced segmentation, for instance, when the size of the target foreground region is several times less than the background. It works as a distance metric on the space of contours; computing active-contour flows through a non-symmetric L2 distance on the space of contours as a regional integral. This method has shown remarkable results on medical images, for example, in the task of white matter hyperintensities segmentation. However, our experiments didn't show any positive impact of boundary loss for polyp segmentation. It decreased the performance severely, refer to the comparison in Table 2.

Table 2
Metrics on round 2 test data of PolypGen2.0 track in EndoCV2022 challenge. MF stands for MaskFormer.

Method	Dice	Dice std	Type2 error
MF	0.5497	0.4319	0.556
MF + boundary loss	0.3346	0.3631	0.400

6. Discussion

We assume that including sequence information as an input to MaskFormer [5] can potentially improve the results. Since the original MaskFormer architecture starts with a regular convolution, one could combine sequences into a volume and pass it as a separate channel for the convolutional layer. Another option is to use a Mask2Former [12] model, which was created for video segmentation and inspired by MaskFormer. Mask2Former [12] is based on Masked-attention Mask Transformer for universal image and video segmentation. It is possible to incorporate their idea in combining the images from the same sequence into a single input with additional dimension responsible for time frames.

7. Conclusion

We are first to show the mask classification-based model performance on endoscopy data. We use MaskFormer [5] as the main component of our approach, adding modifications to the number of queries, for instance, decreasing the number as polyp segmentation is a binary segmentation task. We also introduce a simple pre-processing technique for endoscopy images, which helps to remove redundant information from the input. This step simplifies the learning of meaningful features for the model. Moreover, we add test time augmentation and connected-component analysis at post-processing. Combining all these components achieves a 54.97 Dice score on round 2 validation in the EndoCV2022 challenge.

In this work, we also experiment with boundary loss for MaskFormer [5] and show that it doesn't bring improvements in the polyp segmentation task.

References

[1] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[2] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002. URL: https://doi.org/10.1016/j.media.2021.102002. doi:10.1016/j.media.2021.102002.

[3] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, et al., An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific Reports 10 (2020). URL: https://doi.org/10.1038/s41598-020-59413-5. doi:10.1038/s41598-020-59413-5.

[4] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, J. Rittscher, A deep learning framework for quality assessment and restoration in video endoscopy, Medical Image Analysis 68 (2021) 101900. URL: https://doi.org/10.1016/j.media.2020.101900. doi:10.1016/j.media.2020.101900.

[5] B. Cheng, et al., Per-pixel classification is not all you need for semantic segmentation, 2021. URL: https://arxiv.org/abs/2107.06278. doi:10.48550/ARXIV.2107.06278.

[6] N. Carion, et al., End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.

[7] A. Vaswani, et al., Attention is all you need, Advances in neural information processing systems 30 (2017).

[8] A. Lou, et al., Caranet: Context axial reverse attention network for segmentation of small medical objects, arXiv preprint arXiv:2108.07368 (2021).

[9] D. Jha, et al., Kvasir-SEG: A segmented polyp dataset, in: MultiMedia Modeling, Springer International Publishing, 2019, pp. 451–462. URL: https://doi.org/10.1007/978-3-030-37734-2_37. doi:10.1007/978-3-030-37734-2_37.

[10] T.-Y. Lin, et al., Focal loss for dense object detection, 2017. URL: https://arxiv.org/abs/1708.02002. doi:10.48550/ARXIV.1708.02002.

[11] H. Kervadec, et al., Boundary loss for highly unbalanced segmentation, Medical Image Analysis 67 (2021) 101851. URL: http://dx.doi.org/10.1016/j.media.2020.101851. doi:10.1016/j.media.2020.101851.

[12] B. Cheng, et al., Masked-attention mask transformer for universal image segmentation, arXiv (2021).