HCMUS at MediaEval2021: Polyps Segmentation using TransFuse with Focal Tversky Loss

Nhat-Khang Ngo[1,2], Tuan-Luc Huynh [1,2], Thanh-Danh Le [1,2]
Hai-Dang Nguyen [1,2], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM
[2]Vietnam National University, Ho Chi Minh city, Vietnam
[3] John von Neumann Institute, VNU-HCM
{nnkhang19,htluc19,ltdanh19}@apcs.fitus.edu.vn,nhdang@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

The Medico task, MediaEval 2021, aims at developing accurate and high-performance techniques for automatic medical image segmentation. In this work, we describe an approach for tackling Tasks 1 and 2 of the challenge. We retrain TransFuse, a state-of-the-art model in medical image segmentation, along with focal Tversky loss function to segment the polyp regions in endoscopic images. The approach focuses on computation efficiency while also producing high-quality segmented results. In evaluation, our method achieves appropriate results for both efficiency and accuracy.

1 INTRODUCTION

Medical image segmentation has become more common in recent years, thanks to important advances in artificial intelligence. The work mainly focuses on helping experts diagnose life-threatening cancers by early detecting and segmenting polyps in medical images. However, automatic polyp segmentation is challenging due to the diversity of polyp shapes and positions. Numerous studies leverage the representation power of deep learning to capture numerous variations of polyps in endoscopic images. The MediaEval Task 2021 Transparency in Medical Image Segmentation calls for researchers to investigate a method for polyps segmentation. [5]

This paper presents an approach that can efficiently segment the polyp regions in the endoscopic images. We train from scratch TransFuse [9], a state-of-the-art model in medical image segmentation, along with a generalized focal Tversky loss function [1]. TransFuse is a combination of vision transformers [4] and convolutional neural networks in a parallel manner [9]. While the former learn to model the relations between regions in the images, the latter extracts the local details of these regions. The two processes execute in parallel. Hence, TransFuse boosts the time efficiency in the inference phase. To combine both information, Zhang et al. [9] propose the BiFusion module consisting of several attention modules and convolution blocks. In addition, Kvasir-seg [2], the given dataset, is a small dataset with only 1360 samples. The dataset also consists of many hard samples in which the polyps are large and have unusual locations and shapes. To address this problem, we train TransFuse with focal Tversky loss function. We train the models with various hyperparameter settings to assess the efficacy and failures of this approach.

MediaEval'21, December 13-15 2021, Bergen, Norway and Online

2 RELATED WORK

Self-attention is a critical phenomenon in deep learning. The mechanism enables models to capture the global context between objects in data. Self-attention is used in medical image segmentation to manage the relationships between regions in the images. Oktay et al. [8] integrate Attention Gates into U-net to suppress inessential areas and emphasize salient characteristics. To further handle the global context, Chen et al. [3] proposed TransUnet in which the encoders of U-net are replaced by the encoders of Vision Transformers [4]. Petit et al. [7] proposed a U-net architecture featuring self-attention and cross-attention between the encoder and decoder. While the preceding methods combine self-attention and CNNs sequentially, Zhang et al. [9] combine them in a parallel manner. This kind of incorporation can mitigate the loss of local details in deep CNNs and reduce the inference time.

3 APPROACH

3.1 TransFuse

As illustrated in Figure 1, TransFuse includes three branches; Transformer, CNN, and BiFusion. The Transformer branch makes use of the Vision Transformers architecture, in which an image is embedded into patches before being transmitted to many multi-head self-attention and multi-layer perceptron modules. The result is molded into several feature maps, which are kept for later fusion. Simultaneously, the CNN branch downsamples the image into feature maps with the same size as the corresponding ones in the Transformer branch. The outputs of the two parallel branches are fused in the BiFusion module. The module contains spatial attention, channel attention, and residual blocks to perform multi-modal fusion and self-attention [9]. Finally, the fused output is upsampled to get the segmented result. In addition, deep supervision is provided at the output of the transformer branch and the final BiFusion module. In our experiments, we use TransFuse-S proposed by Zhang et al. [9].

3.2 Focal Tversky Loss

Tversky Score is extended from Dice Score that flexibly adjusts the scores of false positive and false negative cases among the classes [1]. Equation 1 shows how to calculate the Tversky score. In the equation, α is a hyperparameter that we can fine-tune during training. High values of α enhance the recall rate in highly imbalanced datasets [1]. Wider polyp regions, consequently, can be detected in

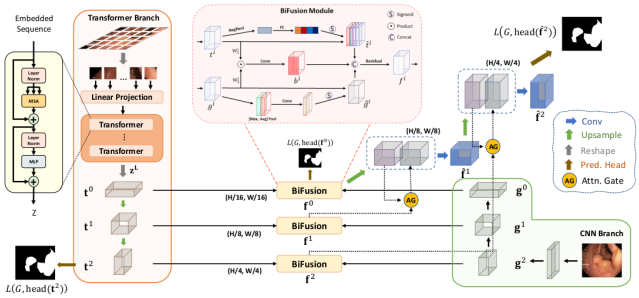

Figure 1: Architecture of TransFuse [9]

images. Additionally, ϵ is a constant that stabilizes the score.

$$T = \frac{TP + \epsilon}{TP + \alpha FN + (1-\alpha)FP + \epsilon} \quad (1)$$

The Tversky Loss L equals $1-T$. To tackle hard samples, Abraham et al. [1] adapt the loss function to a focal version. The loss is written as $FL = (1-T)^{\frac{1}{\gamma}}$, where $\gamma \in [1,3]$ is a hyperparameter. When a high Tversky score has a high number of erroneous predictions, i.e., FN and FP, the loss decreases dramatically. By using $\alpha > 0.5$ and $\gamma > 1$, the function focuses on merely misclassified samples. As a result, the model can widen the segmented polyp regions.

4 EXPERIMENTS AND RESULTS

4.1 Experiments

We train TransFuse-S with the focal Tversky loss by varying α in five Runs. In the first four Runs, we split the dataset into training and validation sets with the ratio of 8:2, whereas we train the model with all samples in Run 5. We use four values of α, including 0.3, 0.4, 0.6, and 0.7. In Run 1 and Run 5, α equals to 0.7, while α equals to 0.6, 0.4, and 0.3 in Runs 2,3,4, respectively. It is worth noting that when $\alpha = 0.5$, the Tversky score becomes Dice score. Thus, we do not use 0.5 in our experiments. In addition, we fix the value of γ to $\frac{4}{3}$ which is proved to be the most effective in [1]. We use Adam [6] to optimize the loss function with a learning rate of $1e-4$, and the batch size of data is 16. Additionally, because we use deep supervsion, there are three losses $L_1, L_2,$ and L_3 with the corresponding scales $\beta_1 = 0.5, \beta_2 = 0.2,$ and $\beta_3 = 0.3$. And thus, the final loss L equals $0.5L_1 + 0.2L_2 + 0.3L_3$.

4.2 Results

Table 1 displays the outcomes of our submissions from Run 1 to Run 5 in the challenge's Task 1. Accuracy, Jaccard score, Dice Score, F1-score, Recall, and Precision are the six metrics used to assess predictions. In Run 2, when $alpha = 0.6$, we attain the highest Jaccard score of 0.6780. This run also produces the highest Dice Score of 0.7756. All runs have a greater recall than a higher precision. This demonstrates our approach's responsibility for false negative predictions. We achieve the greatest recall and accuracy of 0.8584 and 0.8208, respectively. Table 1 further shows that the accuracy ratings for the five runs are almost comparable. In this section, we additionally present the inference time for Task 2. Table 2 shows the

Run ID	Acc	Jacc	DSC	F1	Rec	Prec
Run 1	**0.9484**	0.6684	0.7672	0.7672	0.8430	0.7628
Run 2	0.9462	**0.6780**	**0.7756**	0.7756	0.8413	0.7748
Run 3	0.9406	0.6596	0.7583	0.7583	0.8427	0.7656
Run 4	0.9441	0.6700	0.7644	0.7644	0.7814	**0.8208**
Run 5	0.9407	0.6689	0.7659	**0.7659**	**0.8584**	0.7569

Table 1: Results in Task 1

average inference time and frame rate, as well as the Jaccard Score, Recall, and Precision of Task 2's Run 1. On average, the model makes one prediction in 0.0132 seconds. Besides fast inference, our technique produces accurate findings, with a Jaccard score of 0.6692, a high Recall of 0.8586, and a high Precision of 0.7572. Furthermore, Figure 2 depicts the efficacy and failure of focusing on enhancing the recall rate in the dataset. We paint the polyp regions green based on the projections to see if the borders of these regions are suitable. The first image demonstrates that strong recall is acceptable, whereas the green hue in the second image surpasses the polyp regions.

Run ID	Avg-time	Avg-fps	Jacc	Rec	Prec
Run 1	0.0132	75.7629	0.6692	0.8586	0.7572

Table 2: Results in Task 2

Image **Prediction Map** **Overlay**

Figure 2: Visualization

5 CONCLUSION

We present an approach to automatically segment polyp regions in endoscopic images. Our work is to train from scratch TransFuse along with focal Tversky Loss to tackle hard samples in an imbalanced dataset. We plan to investigate this approach more thoroughly in the future.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] Nabila Abraham and Naimul Mefraz Khan. 2019. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 683–687.

[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

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[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 P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. 2021. U-net transformer: self and cross attention for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*. Springer, 267–276.

[8] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* 53 (2019), 197–207.

[9] Yundong Zhang, Huiye Liu, and Qiang Hu. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. *arXiv preprint arXiv:2102.08005* (2021).