Automatic Polyp Segmentation Using Channel-Spatial Attention with Deep Supervision

Sahadev Poudel[1], Sang-Woong Lee[2]
[1] Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea
[2] Department of Software, Gachon University, Seongnam 13120, South Korea

ABSTRACT

This paper introduces our approach for the automatic segmentation of polyp images in the Gastro-Intestinal (GI) tract. We employ an EfficientNet as an encoder backbone with UNet decoder and leverage the concept of UNet++ of redesigning the skip connections to use multi-scale semantic details. Further, the addition of deep supervision and channel-spatial attention module on the network results in good segmentation performance. The experimental results show the efficiency of the proposed method, which achieves an accuracy of 95.46 %, a recall of 90.31 %, a precision of 86.07 %, and F2-score of 87.4%.

1 INTRODUCTION

The aim of *Medico automatic polyp segmentation challenge* is to segment irregular, small or flat polyps automatically applying different robust and efficient algorithms [1]. A training set of 1000 polyp images with their corresponding masks labeled by medical experts is provided for the participating team. Each team is expected to develop a powerful architecture that can predict the region of interest (ROI) on the testing set. Organizers compare and evaluate all the submitted approaches based on two primary measures: (1) better polyp segmentation task, and (2) efficiency task. In this paper, an encoder-decoder based convolutional neural network (CNN) is introduced to facilitate good segmentation results and efficient polyp detection using the provided data.

2 METHODS

In our method, we utilize the pre-trained weight of variants EfficientNet [2] for the encoder path. Even though medical images are different from the natural images, it is often beneficial to use the pre-trained weights of state-of-the-art CNN architectures for a small datasets [3, 4]. Considering the presence of polyps of varying scales, we utilize the redesigned skip connections from the UNet++ [5]. The densely connected skip connections to the decoder side enable flexible multi-scale feature fusion both horizontally and vertically at the same resolution. Besides, the proposed method powered by deep supervision and channel-spatial attention [6] enables significantly better performance and fast convergence. Integrating channel and spatial attention modules restrain irrelevant features and allow only useful spatial details.

Figure 1 shows a broad overview of our proposed method. EfficientNet uses the MobileNet inverted block (MB) [7] with squeeze and excitation network [8], and a combination of these components works as the good feature extractor module. We keep a network level of s=1 to s=5 depending upon the size of feature map. We reduce the size of feature maps by 2 at each level. The size of the spatial feature map at the last layer (s=5) is $7x7$, which indicates that the feature maps are down-sampled by five times and is halved according to the previous level. At different levels, each node concatenates the feature maps from its previous node of the same level and the upsampled feature maps of the next level, enabling aggregation of multi-scale features. Next, the concatenated features are passed through the channel-spatial network at each node. On the decoder side, a transposed convolution is used for upsampling the feature maps. Similarly, we upscale the outputs of the decoder block at level s=2 to s=5 and apply a 1x1 convolution with 1 kernel and a sigmoid function. Then, all the outputs (after deep supervision) are averaged and a final result is generated. We performed experiments on five different settings as explained below:

- Method 1 uses the UNet++ [5] as a baseline model (skipping deep supervision) and with the EfficientNetB0 as an encoder backbone.
- Method 2 extends Method 1 with the addition of Deep supervision horizontally. The feature maps are upsampled to the input image size at each decoder level, and the sigmoid activation function was applied accordingly.
- Method 3 employs EfficientNetB1 with the same settings as Method 2.
- Method 4 employs EfficientNetB2 with the same settings as Method 3.
- Method 5 employs EfficientNetB3 with the addition of channel-spatial block at each node to restrict the irrelevant features.

For task 2, we use the same architectural design as Method 5. However, we utilize the compound scaling method proposed by EfficientNet [2] in decreasing order to find the optimum scaling dimension of the network. We decrease the network's depth and width and keep the fixed image size of 224x224 to prevent loss of spatial details.

3 DATASET

The dataset includes a total of 1000 polyp images with their corresponding ground truth [9]. The images have a resolution of 332x487 to 1920x1072 pixels. The images are split into training, validation set at a ratio of 80:20. Both training and validation were conducted using images with a pixel resolution of 224x224. We perform a heavy augmentation using albumentation library [10] which includes rotation, vertical and horizontal flipping, cutout, shearing, scaling, zooming.

MediaEval'20, December 14-15 2020, Online

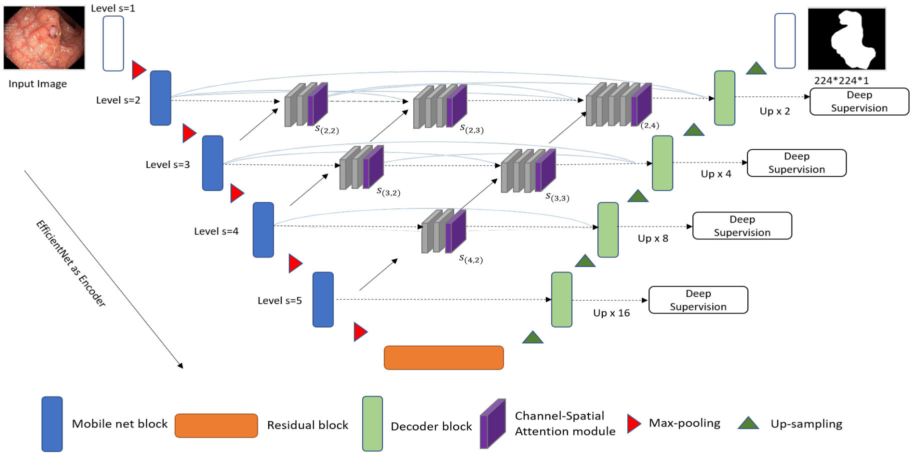

Figure 1: Overview of proposed method.

Table 1: Jaccard index, Dice Similarity Coefficient (DSC), Recall, Precision, Accuracy, F2-score for the segmentation of Task 1.

Model	Jaccard	DSC	Recall	Precision	Accuracy	F2	Params (in million)
Model 1	0.7446	0.8230	0.8835	0.8323	0.94702	0.8469	6.8241
Model 2	0.7688	0.8434	0.8476	0.8954	0.9603	0.8402	6.8338
Model 3	0.7538	0.8305	0.8931	0.8316	0.9448	0.8548	9.3595
Model 4	0.7552	0.8364	0.8901	0.8366	0.9470	0.8563	10.7474
Model 5	0.7897	0.8607	0.9031	0.8673	0.9546	0.8748	14.0491

Table 2: Frame per second (FPS), Mean time taken, Jaccard index, Dice Similarity Coefficient (DSC), Recall, Precision, Accuracy, F2-score for the segmentation of Task 2.

Model	FPS	Mean time taken	Jaccard	DSC	Recall	Precision	Accuracy	F2	Params (in million)
Model 1	2.2514	0.4441	0.5083	0.6265	0.6003	0.7870	0.9149	0.6029	2.5

4 IMPLEMENTATION DETAILS

The implementation is based on Keras, and the backend is TensorFlow. We use a stochastic gradient descent with a batch size of 16 and use a weight decay of 0.0001 with a momentum of 0.9 without an accelerated gradient. The experiments were conducted using an Intel® Core™ i7-7700 CPU @ 3.60GHz × 8 with a GeForce GTX 1080 Ti with 36 GB of RAM.

5 RESULTS AND DISCUSSION

We submitted the predictions of five methods for the testing set (160 images) to the organizers for the evaluation. Table 1 and Table 2 report the experimental results achieved by different models on the segmentation dataset for task1 and task2. Table 1 shows that the addition of deep supervision in model 2 enables better segmentation performance. The model achieves a 2% improvement in performance in terms of the Dice coefficient score. However, under the same settings, applying EfficientB1 and EfficientNetB2 on the encoder path gives a similar performance and a small marginal gain in F2-score. The channel-spatial attention module's addition in model 5 turns out to be the best model achieving 86.07 dice coefficient score and 78.97 Jaccard index. This suggests that the attention module contributes more in comparison to other modules. Similarly, for task 2, the efficient model achieved an accuracy of 91.49 % with F2-score of 0.60, with just 2.5 million parameters. Further, the frame rate in Frames per Second (FPS) while testing in CPU is 2.25142.

6 CONCLUSION

This paper presented five different methods for the accurate segmentation of polyps in GI tract diseases. The proposed methods use an encoder-decoder based architecture where the variants of EfficientNet are applied as an encoder backbone with the UNet decoder. Further, the combination of deep supervision and channel-spatial attention module with an additionally redesigned skip connections achieved the best performance on the test set. We plan to continue researching efficient tasks and further improve the results.

ACKNOWLEDGMENTS

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01907, Development of Smart Signage Technology for Automatic Classification of Untact Examination and Patient Status Based on AI).

REFERENCES

[1] Debesh Jha, Steven A. Hicks, Krister Emanuelsen, Håvard D. Johansen, Dag Johansen, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. Medico Multimedia Task at MediaEval 2020:Automatic Polyp Segmentation.

[2] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[3] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[4] Sahadev Poudel, Yoon Kim, Duc Vo, and Sang-Woong Lee. Colorectal disease classification using efficiently scaled dilation in convolutional neural network. *IEEE Access*, PP:1–1, 05 2020.

[5] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-SEG: A segmented polyp dataset. In *Proc. of International Conference on Multimedia Modeling*, pages 451–462, 2020.

[10] E. Khvedchenya V. I. Iglovikov A. Buslaev, A. Parinov and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*, 2018.