

Automatic Polyp Segmentation using U-Net-ResNet50

Saruar Alam¹, Nikhil Kumar Tomar², Aarati Thakur³, Debesh Jha^{2,4}, Ashish Rauniyar^{5,6}

¹University of Bergen, Norway

²SimulaMet, Norway

³Nepal Medical College, Kathmandu University, Nepal

⁴UiT The Arctic University of Norway

⁵University of Oslo, Norway

⁶Oslo Metropolitan University, Norway

saruar.alam@uib.no

ABSTRACT

Polyps are the predecessors to colorectal cancer which is considered as one of the leading causes of cancer-related deaths worldwide. Colonoscopy is the standard procedure for the identification, localization, and removal of colorectal polyps. Due to variability in shape, size, and surrounding tissue similarity, colorectal polyps are often missed by the clinicians during colonoscopy. With the use of an automatic, accurate, and fast polyp segmentation method during the colonoscopy, many colorectal polyps can be easily detected and removed. The “Medico automatic polyp segmentation challenge” provides an opportunity to study polyp segmentation and build an efficient and accurate segmentation algorithm. We use the U-Net with pre-trained ResNet50 as the encoder for the polyp segmentation. The model is trained on Kvasir-SEG dataset provided for the challenge and tested on the organizer’s dataset and achieves a dice coefficient of 0.8154, Jaccard of 0.7396, recall of 0.8533, precision of 0.8532, accuracy of 0.9506, and F2 score of 0.8272, demonstrating the generalization ability of our model.

1 INTRODUCTION

Identification and removal of polyps during colonoscopy have become a standard procedure. It is often challenging to detect polyps, as they are often hard to differentiate from surrounding normal tissue. These polyps are usually covered with stool, mucosa, and other materials that can obscure the correct diagnosis. This is especially true for the small, flat, and sessile polyps that are typically not visible during colonoscopy. Moreover, this increases the miss-rate of polyps up-to 25% [8] and increases the risk of colorectal cancer in the affected patient. An increase in the 1% adenoma detection rate leads to a 3% decrease in the risk of colorectal cancer [3]. Recently, deep learning techniques have been developed to overcome these challenges and improve polyp detection accuracy during colonoscopy. Polyp segmentation based deep learning methods has been successfully applied for automatic polyp detection in a real-time.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MediaEval'20, December 14-15 2020, Online

The automatic polyp segmentation plays an important role in the identification and localization of the polyps in the affected regions. It helps in analyzing the images or even video frames and classify each pixel into polyp or non-polyp class instances. This allows the clinician in easy, fast, and more accurate identification of the polyp in the affected region. The automated polyp segmentation can help in the development of a Computer-Aided Diagnosis (CADx) system, which is specially designed for colonoscopy procedures.

The “Medico Automatic Polyp Segmentation Challenge” [6] consists of two tasks. The first task is “Polyp segmentation task” and the second is “Algorithm efficiency task”. We have submitted our model in task 1 only.

2 RELATED WORKS

For semantic segmentation task, encoder-decoder networks like FCN [9], U-Net [10], etc are mostly preferred over other approaches. U-Net and its variants are used for both natural image segmentation and biomedical image segmentation. In general, the encoder uses multiple convolutions to learn and capture the essential semantic features ranging from low-level to high-level. These upsampled features are then concatenated with the features from the encoder using the skip connections and then followed by convolution layers to generate the final output in the form of a binary mask.

The encoder acts as a feature extractor, where the decoder uses features extracted from the input to produce to desired segmentation mask. The encoder can be replaced by a pre-trained network such as VGG16 [12], VGG19 [12], etc. These pre-trained networks are already trained on the ImageNet [11] dataset and have the necessary feature extraction capabilities. Architectures like SegNet [2] and U-Net [10] use pre-trained VGG16 and VGG11 respectively for segmentation task.

With the success of the residual network [4], ResNet50 is one of the commonly used architecture for any transfer learning task. The residual network uses two 3×3 convolutional layers and an identity mapping. Each convolution layer is followed by a batch normalization layer and a Rectified Linear Unit (ReLU) activation function. The identity mapping is the shortcut connection connecting the input and output of the convolutional layer. The identity mapping helps in building a deeper neural network by eliminating the problem of vanishing gradients and exploding gradients.

3 APPROACH

Figure 1 shows an overview of the proposed U-Net-ResNet50 architecture. It is an encoder-decoder based architecture, where ResNet50

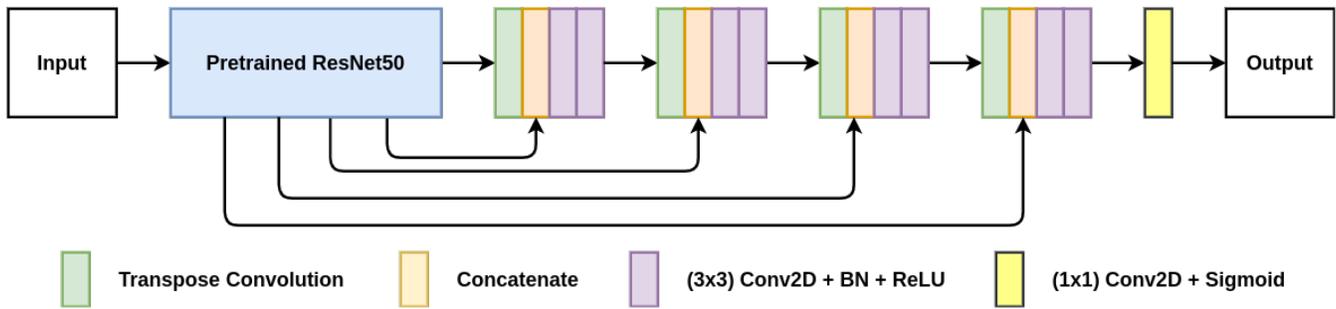


Figure 1: The proposed U-Net-ResNet50 architecture

trained on ImageNet dataset [11] is used. The use of a pre-trained encoder helps the model to converge easily. The input image is fed into the pre-trained ResNet50 encoder, consisting of a series of residual blocks as their main component. These residual blocks help the encoder extract the important features from the input image, which are then passed to the decoder. The decoder starts a transpose convolution that upscales the incoming feature maps into the desired shape. Next, these upscaled feature maps are concatenated with the specific shape feature maps from the pre-trained encoder via skip connections. These skip connections help the model to get all the low-level semantic information from the encoder, which allows the decoder to generate the desired feature maps. After that, it is followed by the two 3×3 convolution layer, where each layer is followed by a batch normalization layer and a ReLU non-linearity. The last decoder block's output is passed to a 1×1 convolution layer, which is further passed to a sigmoid activation function, finally generating the desired binary mask.

The FastAI (version 2.0) library [1] is used to train and evaluate our model. We have employed resizing, flipping, rotating, zooming, lightning, warping, and normalizing intensity based on the ImageNet dataset to augment the input images for training. The model uses Adam optimizer with an initial learning rate of 10^{-2} , and cross-entropy loss as its loss function. We have employed the one-cycle policy where the learning rate changes during training and achieves super-convergence [13]. We have run just 50 epochs for training, and the model has converged.

4 RESULTS AND ANALYSIS

The Medico Automatic Polyp Segmentation challenge [6] provides an opportunity to study the potential and challenges of automated polyp segmentation. This study aims at building a model that performs well on the organizer's dataset while training on a separate Kvasir-SEG dataset [7].

Table 1 shows the overall results of the U-Net-ResNet50 architecture on the Kvasir-SEG test dataset and the organizer's test dataset provided for the final evaluation of the model. For the evaluation of the model, the Jaccard index, Sørensen-Dice coefficient (DSC), recall, precision (Prec.), accuracy (Acc.), and the F2 are used as the evaluation metrics. Our trained U-Net-ResNet50 model achieved a dice coefficient of 0.8154, Jaccard of 0.7396, recall of 0.8533, precision of 0.8532, accuracy of 0.9506 and F2 score of 0.827 on the organizer's test dataset which can be seen from the table 1. These

Table 1: Quantitative Results on Kvasir-SEG and Test Set (Challenge) Dataset for Task 1.

Dataset	Jaccard	DSC	Recall	Prec.	Acc.	F2
Kvasir-SEG	0.7871	0.8926	0.8433	0.9207	0.9639	0.8585
Test Set	0.7396	0.8154	0.8533	0.8532	0.9506	0.8272

results demonstrate the generalization ability of our model. Moreover Table 1 also shows that the recall value of the organizer's test dataset is 1.00% higher than the Kvasir-SEG test dataset. This shows that the model is not overfitting.

5 CONCLUSION & FUTURE WORK

With our U-Net-ResNet50, we achieved competitive performance on the organizer's dataset with a dice coefficient of 0.8154. By replacing the U-Net encoder with a pre-trained ResNet50 and employing a one-cycle policy during training, we are able to converge the model in a short time. Thus, it helps in reducing the training time as the encoder weights are not initialized from scratch. This is an important step towards faster convergence, which would be useful when the availability of high-performance computing resources is limited.

In the future, we would like to experiment with more than one pre-trained encoder by fusing their feature maps and using them for training our model.

ACKNOWLEDGMENTS

The computations in this paper were performed on the equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under the contract 270053.

The authors would also like to thank the machine learning group of Mohn Medical Imaging and Visualization (MMIV) Centre, Norway, for providing the computing infrastructure for the experiments.

REFERENCES

- [1] 2020. FastAI Library. (2020). <https://docs.fast.ai/>.
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.
- [3] Douglas A Corley, Christopher D Jensen, Amy R Marks, Wei K Zhao, Jeffrey K Lee, Chyke A Doubeni, Ann G Zauber, Jolanda de Boer, Bruce H Fireman, Joanne E Schottinger, and others. 2014. Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine* 370, 14 (2014), 1298–1306.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [5] Vladimir Iglovikov and Alexey Shvets. 2018. TeraNet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746* (2018).
- [6] Debesh Jha, Steven A. Hicks, Krister Emanuelsen, Håvard D. Johansen, Dag Johansen, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. 2020. Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation. In *Proc. of MediaEval 2020 CEUR Workshop*.
- [7] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-SEG: A Segmented Polyp Dataset. In *Proc. of International Conference on Multimedia Modeling (MMM)*. 451–462.
- [8] Sheila Kumar, Nirav Thosani, Uri Ladabaum, Shai Friedland, Ann M Chen, Rajan Kochar, and Subhas Banerjee. 2017. Adenoma miss rates associated with a 3-minute versus 6-minute colonoscopy withdrawal time: a prospective, randomized trial. *Gastrointestinal endoscopy* 85, 6 (2017), 1273–1280.
- [9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.
- [10] 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.
- [11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- [12] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [13] Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Vol. 11006. International Society for Optics and Photonics, 1100612.