Depth-Wise Separable Atrous Convolution for Polyps Segmentation in Gastro-Intestinal Tract

Syed Muhammad Faraz Ali, Muhammad Taha Khan, Syed Unaiz Haider, Talha Ahmed, Zeshan Khan, Muhammad Atif Tahir
{k190861,k173656,k173667,k173721,zeshan.khan,atif.tahir}@nu.edu.pk
National University of Computer and Emerging Sciences, Karachi Campus, Pakistan

ABSTRACT

Identification of polyps in endoscopic images is critical for the diagnosis of colon cancer. Finding the exact shape and size of polyps requires the segmentation of endoscopic images. This research explores the advantage of using depth-wise separable convolution in the atrous convolution of the ResUNet++ architecture. Deep atrous spatial pyramid pooling was also implemented on the ResUNet++ architecture. The results show that architecture with separable convolution has a smaller size and fewer Giga-Floating Point Operations (GFLOPs) without degrading the performance too much.

1 INTRODUCTION

Wireless capsule endoscopy (WSE) has been used for diagnosis for nearly 10 years now. WSE images provide diagnosis capability for many diseases such as colon cancer, ulcer, polyps detection, etc. With the advent of deep learning in computer vision, this diagnosis task can be automated.

2 RELATED WORK

The gastrointestinal tract has been an active area of research. The benefit that can be achieved through computer-aided diagnosis is significant. Jha et al. [9] studied the semantic segmentation of polyps in the GI tract. This research utilizes the well-accepted U-net architecture and modified U-net architecture also called ResUNet for segmentation. Further research was conducted to introduce a novel architecture named ResUNet++.

3 APPROACH

The approach follows the method used by Jha et al. [9]. The ResUNet++ architecture was employed which uses the encoder and decoder structure for semantic segmentation. Pyramid pooling was used as a bridge between the encoder and decoder block. The encoder block contains residual units that take advantage of skip connection in a neural network. The skip connection allows training a deep neural network without degrading the performance. Squeeze and excitation blocks were used which ensure that the channel output features are weighted equally [5]. The attention mechanism is used in the decoder block. The attention mechanism is useful in making a pixel-wise prediction. This approach is popular in natural language processing (NLP) where attention is given to each word of a sentence. In semantic segmentation, an attention mechanism is used to give attention to each pixel of an image which can then be used to make a prediction at pixel level [6].

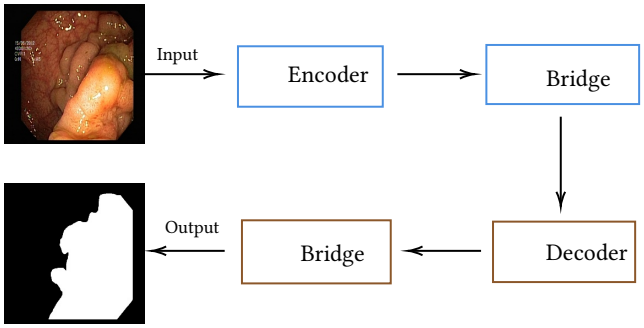

Figure 1: Process Flow

A bridge of pyramid pooling is used between encoder and decoder block [2] [1]. The atrous convolution is used in this bridge through which the output of the encoder is viewed at the various respective fields. This block convolves the features with the kernel of different dilation rates and the final output is the concatenation of all the convolutions. This way the contextual information in features is captured at various scales.

This Atrous Spatial Pyramid Pooling (ASPP) block in ResUNet++ was implemented using depth-wise separable convolution as well as replaced with Deep Atrous Spatial Pyramid Pooling (DASPP) module from [4] in separate experiment. The implementation of depth-wise separable convolution is done by applying kernel on input at channel level. The output from here is passed through the pointwise convolution with 1x1 kernel [3]. The application of depth-wise convolution results in fewer GFLOPs and parameters.DASPP was implemented to see if going deep in network improves performance on polyps segmentation. Three modified architecture are:

(1) sepv_conv_resunet++ : ASPP module from ResUNet++ [9] replaced with depth-wise separable convolution.
(2) dsapp_resunet++ : ASPP module replaced with DASPP module from [4].
(3) dsapp_relu_resunet++ : 2 implemented with ReLu activation.

Semantic segmentation, unlike object detection, can be treated as a pixel wise classification problem. The output of semantic segmentation for a pixel is a mask identifying the class to which the pixel belongs. For the polyps segmentation problem [7], this mask

This research work was funded by Higher Education Commission (HEC) Pakistan under NRPU Project 10225/2017.
Copyright 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'20, December 14-15 2020, Online

CEUR-WS.org/Vol-2882/paper31.pdf

is either 0 or 1. The evaluation metrics used in semantic segmentation are accuracy, precision, recall, mean Intersection over Union (mIoU), and dice co-efficient. All these except for accuracy were used to identify model performance. The custom loss function for mIoU was implemented and all model architectures were trained on this custom loss.

4 DATASET

The experiments were performed on Kvasir-SEG dataset [8]. This data consists of thousand polyps images. The ground truth values against each of these images were provided as image masks in a separate folder.

5 RESULTS AND ANALYSIS

All the experiments were performed on google Colab which provides a session for up to 12 hours. This 12 hours session is not enough to train a deep learning model. So to make a fair comparison, the number of epochs for all the experiments was kept the same. The data was split into training, validation, and testing set with the ratio of 80, 10, and 10 percent respectively. With this split, 800 images were selected for model training. These 800 images are not enough to train a deep learning model. To increase the training set data augmentation technique was applied to the training set. The validation set and testing set were not modified and thus the size of validation and test sets were 100 images each. 30 different augmentations were applied to the training set after that the size of it grew to 24800 images. The augmentations were also applied to the provided mask so that the target variable is transformed in the same way as the input image.

The optimizer used from training was NAdams optimizer with a learning rate of 0.0001 and a batch size of 8. The learning curve for training and validation loss was recorded for each epoch. The learning curve provides insights into the model convergence.

Figure 2 shows the learning curve for each architecture. The architecture with the DASPP bridge shows that it may have converged within 10 epochs as the validation error started increasing. However, the ResUNet++ and separable convolved ResUNet show that the model can be trained for few more epochs as both training and validation error are still decreasing. For UNet, the learning curve is also decreasing at the 10th epoch. However, the value of the loss is higher than the loss of ResUNet++ architecture.

Model	Recall	Precision	Dice	mIoU
Unet	**75.23%**	84.52%	71.91%	59.53%
resunet++	64.97%	89.81%	**78.35%**	**69.48%**
sepv_conv_resunet++	60.55%	**93.31%**	77.25%	67.56%
dsapp_resunet++	69.72%	82.62%	76.66%	66.71%
dsapp_relu_resunet++	61.54%	92.33%	74.63%	66.03%

Table 1: Test Data Results

Table 1 gives the performance of each model on testing data. The performance of ResUNet++ on dice coefficient and mIoU is better than other models. The performance of the model with separable convolution has comparable results on Dice and mIoU metrics. However, the model with the DASPP bridge did not perform well.

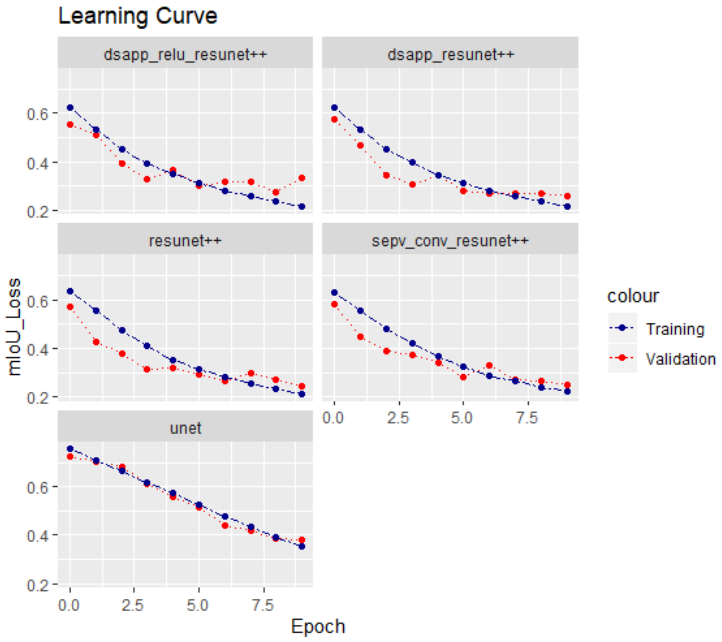

Figure 2: Learning Curve

This shows that increasing depth any further did not improve performance. The size of the model which is measure by the number of parameters and Giga-Floating Point Operations (GFLOPs) is best for the model with separable convolution. The results are compiled in table 2. The less number of parameters means that the model size is smaller and it may be easy to move this in a production environment.

Model	Params	GFLOPs
Unet	3,588,997	7,165,148
resunet++	4,371,265	8,718,068
sepv_conv_resunet++	**3,047,265**	**6,070,057**
dsapp_resunet++	5,024,705	10,024,898
dsapp_relu_resunet++	5,024,705	10,024,898

Table 2: Model Size

6 CONCLUSION AND FUTURE WORK

The research gives empirical results of the advantage of using depth-wise separable convolution which resulted in smaller model size without significantly affecting the performance. It has also been shown that increasing the depth further may not improve performance and can result in overfitting of the model. It has been observed that the implementation of depth-wise separable convolution results in a smaller model without much degradation in overall performance. The tuning of hyper-parameter and a larger number of epochs will give a better understanding of the performance.

REFERENCES

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[3] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[4] Taha Emara, Hossam E Abd El Munim, and Hazem M Abbas. 2019. LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–7.

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

[6] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 603–612.

[7] Debesh Jha, Steven A. Hicks, Krister Emanuelsen, Håvard 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 the MediaEval 2020 Workshop*.

[8] 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 *International Conference on Multimedia Modeling*. Springer, 451–462.

[9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. 2019. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 225–2255.