A MULTI-SCALED RECEPTIVE FIELD LEARNING APPROACH FOR MEDICAL IMAGE SEGMENTATION

Pengcheng Guo *Xiangdong Su** *Haoran Zhang* *Meng Wang* *Feilong*

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
College of Computer Science, Inner Mongolia University, Hohhot, China

ABSTRACT

Biomedical image segmentation has been widely studied, and lots of methods have been proposed. Among these methods, attention U-Net has achieved a promising performance. However, it has drawbacks of extracting the multi-scaled receptive field features at the high-level feature maps, resulting in the degeneration when dealing with the lesions with apparent scale variations. To solve this problem, this paper integrates an atrous spatial pyramid pooling (ASPP) module in the contracting path of attention U-Net. This module employs multiple dilation rates for the purpose of obtaining several multi-scale receptive fields, which significantly improves the networks' ability of handling both large and small lesions. Evaluation experimental result shows that our approach significantly improves the performance of medical image segmentation and substantially outperforms the representative deep learning models on public datasets.

***Index Terms*—** biomedical image segmentation, atrous spatial pyramid pooling, attention U-Net, feature maps, receptive field

1. INTRODUCTION

Biomedical image segmentation aims to perform pixel-level annotations on regions of interest of biomedical images, laying the foundation of biomedical image analysis [1–3]. It has been widely studied and a large number of methods were proposed [4–6]. Among these methods, deep learning-based approaches show very competitive results. Ciresan et al. [7] propose a well-known network, which uses the sliding-window to predict the category label of each pixel by providing a local area (patch) around each pixel. Although the strategy is effective in work [7], there are two disadvantages [8]. First of all, the training speed of this method is relatively slow, as each patch must work separately in the network, where patch overlapping brings numerous redundancy. Secondly, different sizes of patches make it difficult for the network to balance between localization accuracy and context usage. In recent years, researchers attempt to resolve these problems. Olaf Ronneberger et al. [8] propose U-Net for biomedical image segmentation. This network consists of a contracting path with successive layers and a symmetric expanding path with a large number of feature channels. With upsampled output combined, the successive layers, with which high-resolution features in the contracting path were enhanced, achieves more accurate positioning; while in the symmetric expanding path, large numbers of feature channels make contextual information available.

Although U-Net has the above advantages, it also has some disadvantages. For example, computing resources and model parameters are used redundantly. All cascaded network layers repeatedly extract similar low-level features. Hence, Ozan Oktay et al. [9] proposed a more effective method, namely attention U-Net. This network integrates attention gates (AGs) into a standard U-Net model and dispenses with train multiply models and extra model parameters. Moreover, AGs learn to highlight the salient features for the specific task while suppressing irrelevant regions in the input images so as to improve the segmentation performance.

Although attention U-net is an effective model in medical image segmentation, it has drawbacks of extracting the multi-scaled receptive field features at the high-level feature maps, resulting in the degeneration when dealing with the lesions with obvious scale variations. To solve this problem, we integrate an atrous spatial pyramid pooling (ASPP) [10] module in the contracting path of Attention U-Net. This module employs multiple dilation rates to obtains several multi-scale receptive fields so that they improve the networks' ability to handle both large and small lesions. It was proved that explicitly accounting for object scale can improve the network to handle scale variability in semantic segmentation [11]. The experiment shows that our approach outperforms attention U-Net.

The advantages of the proposed approach are as follows. At first, we integrate the ASPP module in the attention U-Net to extract multi-scale features from the high-level feature maps. And the ASSP improves the model in handling biomedical image segmentation with scale variability. Secondly, ASPP improves the receptive field using dilated convolution while keeping the computation cost. Finally, the network in the proposed approach has one more downsampling layer and upsampling layer than the attention U-Net [9]. It is beneficial to mining deeper semantics and predicting the

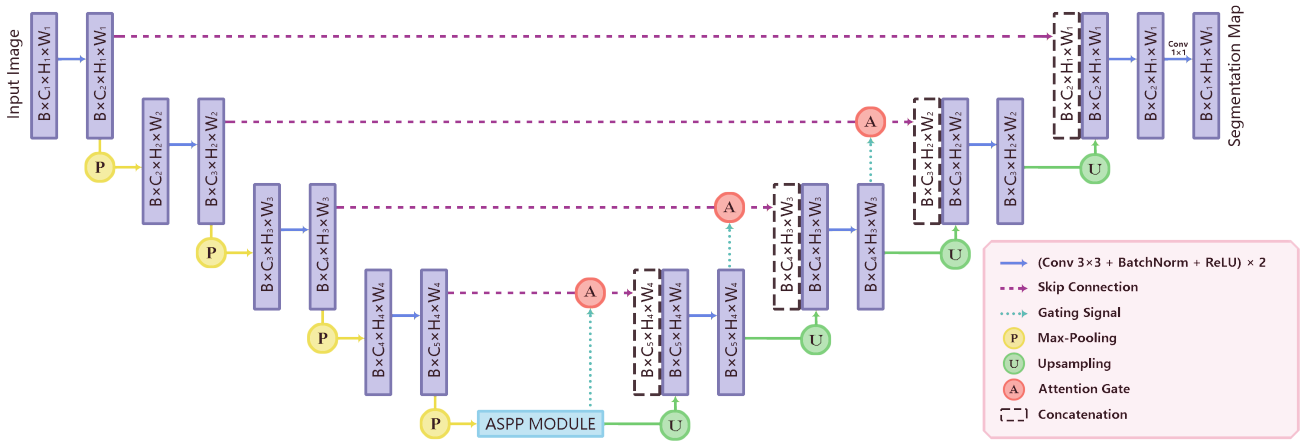

Fig. 1. Proposed Attention U-Net with ASPP module architecture.

lesions more accurately.

2. METHODOLOGY

2.1. Framework

This paper proposes a novel biomedical image segmentation network, which embeds the ASPP module into the attention U-Net architecture. The ASPP module utilizes multiple parallel atrous convolutions to extract multi-scaled receptive field features, which improves the segmentation accuracy of the model for the lesions with obvious scale variations. The proposed ASPP Attention U-Net contains two parts, encoder and decoder, as shown in Figure 1. There is an essential part of the network structure called skip connection. It concatenates the resulted feature maps from the decoder block with that from the corresponding encoder block.

2.1.1. Encoder

Encoder contains eight downsampling blocks and an ASSP module. Each downsampling block contains a 3×3 convolution layer, a batch normalization operation and a ReLU activation operation. Two downsampling blocks are concatenated together, followed by a 2×2 max pooling operation. After all downsampling blocks, ASPP modules with different dilation rates ($r = \{1, 6, 12, 18\}$) are used to extract the multi-scaled receptive field features at the high-level feature maps, and these features help the model to deal with the lesions with obvious scale variations. The ASPP module is described in Section 2.2 and Section 2.3.

2.1.2. Decoder

The decoder section is similar to Attention U-net [9]. It contains eight upsampling blocks. The eight upsampling blocks are divided into four parts, and the first three parts have the same structure. Each part includes a gating signal(g) operation, an attention gate (AG) operation, a concatenation operation and two 3×3 convolution layers. The fourth part does not use the gating function because it does not represent input data in high-dimensional space [9]. To ensure the semantic information of the intermediate feature map is different at each image scale, the deep supervision strategy [12] was introduced into the network structure. Deep supervision applies 1×1 convolution and sigmoid activation to each output mapping layer.

2.2. Atrous Convolution

In the ASPP module, atrous convolution was developed initially for wavelet transform [13] and called dilated convolution in [14]. It inflates the kernel by increasing the kernel interval so as to increase the reception field size to incorporate a larger context, as shown in Figure 2. Dilation rate represents the kernel interval, compared to the standard convolution. By changing the dilation rate value, the filter's receptive field can be modified correspondingly. For a one-dimensional(1-D) input signal $x[i]$, the output $y[i]$ of dilated convolution on it with a filter $w[k]$ of length K is defined as:

$$y[i] = \sum_{i=1}^{K} x[i + r \cdot k] \, w[k] \tag{1}$$

where r represents the dilated rate. If $r = 1$, the dilated convolution is the same as the conventional convolution. A visual description is shown in Figure 2.

Given an image, the first step is to reduce the resolution to half by using downsampling operation, and then perform a standard convolution operation. After the convolution, the feature map is put into the coordinate system of the original image. The result shows that only 1/4 of the response in the original image. If the standard convolution operation

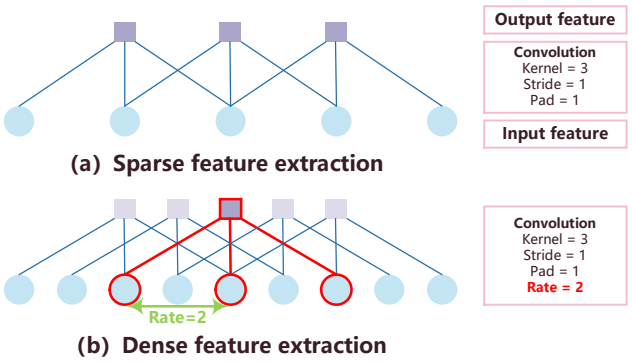

Fig. 2. Illustration of atrous convolution in 1-D. (a)traditional convolution extracts sparse feature, (b)atrous convolution ($r = 2$) intensive extracts feature.

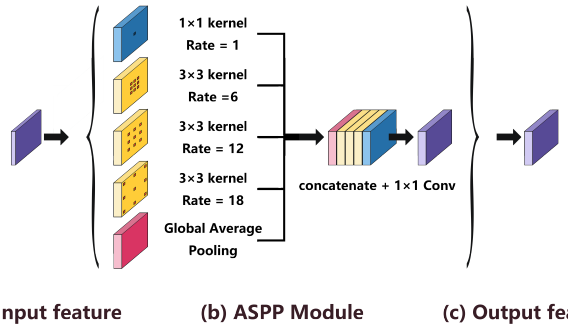

Fig. 3. Illustration of ASPP module structure. (a) Input feature, (b)ASPP module employs four parallel atrous convolutions and a global average pooling, all atrous convolution followed by batch normalization and ReLu, (c)Output feature.

is replaced by atrous convolution operation, we can calculate responses at all image positions. The reason is that atrous convolution with rate r inserts $r-1$ zeros in the middle of the filter. For example, a filter (kernel size $k \times k$, dilate rate r), which is enlarged to $k_e \times k_e$ where $k_e = k + (k-1)(r-1)$. The advantage of this way is that it increases the receptive field but does not add new parameters and calculations.

2.3. Atrous Spatial Pyramid Pooling(ASPP)

In our approach, the ASPP module is used in the bottleneck part to extract multi-scale features from the high-level feature maps. This module combines multiple receptive field features by using atrous convolutions of different dilation rates as the final prediction.

As shown in Figure 3, the mentioned ASPP module employs four parallel atrous convolution and global average pooling. Each atrous convolution contains a convolution operation, a batch normalization operation and ReLU. Then, four parallel atrous convolution and global average pooling are concatenated together, followed by a 1×1 convolution operation. Here the dilation rates are $1, 6, 12, 18$, respectively.

2.4. Loss Function

We use dice scores as a loss function to optimize the proposed model.

$$L_{DSC} = 1 - \frac{2\sum_i^N s_i r_i}{\sum_i^N s_i + \sum_i^N r_i} \tag{2}$$

where s_i and r_i represent the continuous values of the softmax prediction graph $\in [0, \ldots, 1]$ and the ground truth at each voxel i, respectively. The parameter ϵ represents Laplace smoothing, and the value of ϵ is 1. Using formulation 2 to measure the error between the results of the segmentation and the label. This formulation 2 can be differentiated yielding the gradient computed with respect to the j-th voxel of the prediction.

$$\frac{\partial L_{DSC}}{\partial s_j} = -2\left[\frac{r_j\left(\sum_i^N s_i + \sum_i^N r_i + \epsilon\right) - \left(\sum_i^N s_i r_i + \epsilon\right)}{\left(\sum_i^N s_i + \sum_i^N r_i + \epsilon\right)^2}\right] \tag{3}$$

Using formulation 3, we do not need to assign weights to samples of different classes to establish the right balance between foreground and background voxels.

3. EXPERIMENT

3.1. Datasets and Metrics

We evaluate our model on two public medical image datasets.

ISIC 2018 skin lesion dataset [15, 16]: This dataset includes 2594 RGB images of skin lesions with an average image size of 2166×3188 pixels. The dataset were resampled to 192×256 pixels and divided into training sets and test sets according to $75 - 25$.

LGG segmentation dataset [17]: This dataset was created for the task of brain glioma segmentation. It contains 3929 brain MR images and corresponding manual FLAIR anomaly segmentation masks. The average image size is 256×256 pixels. They correspond to 110 patients. In our experiments, the dataset was split into the training dataset (90%) and the test dataset (10%).

We use three standard evaluation metrics in biomedical image segmentation. They are dice similarity coefficient (DSC), prediction and recall, respectively. The three metrics are computed as follows:

$$DSC = \frac{2\,|I_1 \cap I_2|}{|I_1| + |I_2|} \tag{4}$$

The area parameter I_1, I_2, is the area of pixel-level segmentation area, ground truth, respectively.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

where TP, TN, FP, FN are the number of pixel-level true positives, true negatives, false positives, and false negatives, respectively.

3.2. Baselines

In this paper, we compared ASPP Atttention U-Net with the following methods, including U-Net [8], Attention U-Net [9] and Attention U-Net with Multi-Input [18] on two standard biomedical image datasets.

- U-Net: U-Net is a CNN-based image segmentation network, which mainly used for medical image segmentation. This network consists of a contracting path and a symmetric expanding path. We used the same preprocessing method and implementation as mentioned in [8].
- Attention U-Net: Attention U-Net was proposed by Jo Schlemper et al. [9], which is an effective method for medical image segmentation. This network integrates attention gates (AGs) into a standard U-Net model and contributes to highlight salient features.
- Attn U-Net with Multi-Input: Attn U-Net with Multi-Input is an improvement on Attention U-Net. It integrates the encoder with an input image pyramid before each maxpooling layer. We reimplemented the model and use the same hyperparameter as that in [18].

3.3. Implementation Details

Our method was compared with three baselines on ISIC 2018 datasets and LGG segmentation datasets. We tested metrics through 5-fold cross-validation experiments. However, the two datasets are different in the experimental details, which are described in details below.

The ISIC 2018 experiment has been trained in 50 epochs with a batch size of 8. The SGD optimizer is adopted to update network parameters with the learning rate set to 0.01 initially, momentum value is 0.9 and the learning decay rate is $1\mathrm{e}{-6}$.

The LGG segmentation experiment was also trained for 100 epochs with a batch size of 16. The optimization method is Adam with learning rate at 0.0001, decay rate is 0.9.

3.4. Quantitative Results

Table 1 lists the performances of our approach and the baseline approaches on ISIC 2018. It is clear that out approach achieves the hightest DSC 86.5% and recall 88.4%. Attn U-Net with multi-input gets the highest precision 89.6%. Among the three metrics, DSC is widely recognized as the main reference. Out approach obtains a 3.8% improvement on DSC than Attn U-Net with multi-input. Therefore, our approach performs better than the baseline models comprehensively. This is because the ASPP models can extract multi-scale receptive fields feature on high-level feature maps.

Table 1. Performance on ISIC 2018

Method	DSC	Precision	Recall
U-Net [8]	0.820±0.013	0.849±0.038	0.867±0.048
Attention U-Net [9]	0.806±0.033	0.874±0.080	0.827±0.055
Attn U-Net Multi-input [18]	0.827±0.055	**0.896±0.019**	0.829±0.076
Ours	**0.865±0.002**	0.886±0.013	**0.884±0.015**

Table 2. Performance on LGG segmentation

Method	DSC	Precision	Recall
U-Net [8]	0.901±0.007	0.9188±0.013	0.901±0.003
Att U-Net [9]	0.903±0.002	0.915±00.007	0.895±0.011
Attm U-Net [18]	0.906±0.004	0.926±0.005	0.891±0.002
Ours	**0.911±0.012**	**0.932±0.008**	**0.912±0.012**

Table 2 presents the results from our approach and the baseline approaches on LGG. Our approach surpasses the baseline approaches on all the three metrics, indicating that our approach is effective.

From the comparison on ISIC 2018 and LGG, we come to the conclusion that the proposed approach is robust and advances the performance of attention U-Net by integrating ASPP module in it.

4. CONCLUSION

This paper proposes an improvement of attention U-Net (named ASPP Attention U-Net) in biomedical image segmentation by adding a ASPP module in the contracting path of attention U-Net. ASPP employs different multiple dilation rates and can effectively extracts the multi-scaled receptive field features at the high-level feature maps. It improves the methods' ability to handle both large and small lesions. Evaluation experimental result on ISIC 2018 and LGG shows that our approach can significantly improve the segmentation performance and substantially outperforms the baseline models.

5. ACKNOWLEDGMENTS

This work was funded by National Natural Science Foundation of China (Grant No. 61762069, 61773224), Natural Science Foundation of Inner Mongolia Autonomous Region (Grant No. 2017BS0601, 2018MS06025) and Science and Technology Program of Inner Mongolia Autonomous Region (2019).

6. REFERENCES

[1] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al., "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, pp. 65, 2018.

[2] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler, "Multi-label whole heart segmentation using cnns and anatomical label configurations," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 190–198.

[3] Holger R Roth, Le Lu, Nathan Lay, Adam P Harrison, Amal Farag, Andrew Sohn, and Ronald M Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical image analysis*, vol. 45, pp. 94–107, 2018.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[7] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

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

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[12] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*, 2015, pp. 562–570.

[13] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, pp. 286–297. Springer, 1990.

[14] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[15] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[16] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, pp. 180161, 2018.

[17] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha, "Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data," *Journal of neuro-oncology*, vol. 133, no. 1, pp. 27–35, 2017.

[18] Nabila Abraham and Naimul Mefraz Khan, "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, 2019, pp. 683–687.