Convolutional Attention Model for Retinal Edema Segmentation

Phuong Le Thi[1], Tuan Pham[2,] and Jia Ching Wang[1,3]
[1]National Central University, Taoyuan, Taiwan
E-mail: 106522620@cc.ncu.edu.tw
[2]University of Technology and Education – The University of Danang, Vietnam
E-mail: ptuan@ute.udn.vn
[1] National Central University, Taoyuan, Taiwan
[3] Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan
E-mail: jcw@csie.ncu.edu.tw

***Abstract*—Deep learning and computer vision that become popular in recent years are advantage techniques in medical diagnosis. A large database of Optical Coherence Tomography (OCT) images can be used to train a deep learning model which can support and suggest effectively illnesses and status of a patient. Therefore, semantic image segmentation is used to detect and categorize anomaly regions in OCT images. However, numerous existing approaches ignored spatial structure as well as contextual information in a given image. To overcome existing problems, this work proposes a novel method which takes advantage of the deep convolutional neural network, attention block, pyramid pooling module and auxiliary connections between layers. Attention block helps to detect the spatial structure of a given image. Beside, pyramid pooling module has a responsibility to identify the shape and margin of the anomaly region. In additional, auxiliary connections support to enrich useful information pass through one layer as well as reduce overfitting problem. Our work produces higher accuracy than state-of-the-art methods with 78.19% comparing to Deeplab_v3 76.19% and Bisenet 76.85% in term of dice coefficient. Additionally, a number of parameters in our work is smaller than the previous approaches.**

I. INTRODUCTION

Deep learning attracts numerous attentions in the academy and even in industrial fields recently. Its performance not only overcomes human-level in many applications but also processing time is reduced significantly. One type of deep learning is known popularly as convolution neural networks that are applied widely in most computer vision applications. Actually, deep learning has already researched for a long time ago, but there are many restrictions on the size of data and limited-computational power are the reason why deep learning was employed recently in image processing [1, 2], document clustering [3], audio processing [4], data analysis [5].

Image processing using deep learning is the most important and necessary technique in several applications e.g. object detection, image segmentation, or face recognition which applied potentially in fields of robotics, medical image analysis and many more. In this paper, image segmentation bases on convolutional neural networks is implemented to detect unusual regions and classify those regions at pixel-level. Moreover, there are two implementations in image segmentation e.g. instance segmentation and semantic segmentation. Instance segmentation aims to determinate regions and class of every object, regardless they are the same type. On the other hand, semantic segmentation is automatically to label each pixel in an image with a corresponding class of what is represented. In this work, we concentrate on semantic retinal edema segmentation base on optical coherence tomography dataset using convolutional neural networks. Processed images by semantic segmentation are used as efficient suggestions to diagnose retinal edema disease.

To enhance effectiveness in semantic segmentation, some researchers have mainly investigated the scaling of three dimensions of convolutional neural networks such as depth, width, and height. For example, Bisenet model [6] is focused on scaling of model regarding to width and depth of model architecture to enhance the performance of a model for semantic segmentation. Deeplab V3 model [7] attained a significant result on semantic image segmentation through going deeper with atrous convolution which allows capturing long-range information easily. Deeplab V3 shows that deeper architecture is the important aspect to improve the result of the model. They empirically determinate that scaling of width, height, and depth is an important factor that can help to increase the performance of any model. However, there is a trade-off between performance and processing time i.e bigger model yields higher results and consume more time.

To overcome shortcomings of the previous works, this work proposed an attention-spatial convolution (ASC) model which takes advantage of the deep convolutional neural network, attention block, pyramid pooling module and auxiliary connections between layers. Actually, The ASC model includes two sections. The first one is called a down-sampling section. The second section is defined as an up-sampling section. In detail, the attention block is placed in the down-sampling section to concentrate on outstanding features

because of speckle noise. In addition, pyramid pooling module with different pyramid scales is used into the down-sampling section to avoid evanescence of feature with low-dimension and enhance the representation of the features for different locations.

The rest of this paper is organized by following: section II describes attention block, pyramid pooling module and our model ASC. In section III, experimental results on semantic segmentation are presented. Finally, conclusion is on section IV.

II. PROPOSED METHOD

A. Attention block

Attention block, proposed by [8], is used to concentrate on important features and ignore unimportant ones. In detail, the attention block includes two modules: 1, channel attention module (M_c), and 2, spatial attention module (M_s). Each module performs a different responsibility. More clearly, the channel attention module points out the key features of an input. On the other hand, the spatial attention module utilizes an inter-structure of the input feature. Diagram of convolutional block attention module is illustrated by Fig. 1.

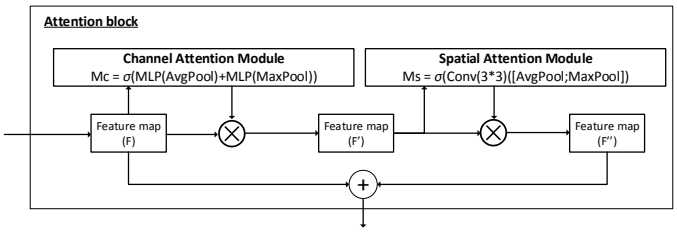

Fig. 1 Diagram of Attention block

The attention block receives the input feature (F) from the previous convolution layers. Mathematically, an output of corresponding F through attention block can be calculated as:

$$F' = M_c(F) \times F \quad (1)$$

$$F'' = M_s(F') \times F' \quad (2)$$

The channel attention module includes: 1, multi-layer perceptron (MLP) [9] which receive average-pooled [10] and max-pooled [11] features; 2, activation function. Two types of pooling layers are used to highlight different important features. To reduce the number of parameters, MLP is a shared-weight network. Activation function can be linear or non-linear. After an element-wise summation is applied which can calculate total the features at the output vectors. Overall, the formulation of the channel attention module is described directly by:

$$M_c = \sigma(MLP(F^c_{avg}) + MLP(F^c_{max})) \quad (3)$$

Where σ indicates the activation function.

Secondly, the spatial attention module figures out relationships among the input features. The spatial attention block consists of: 1, convolution layer which receives concatenation of average-pooled (F^s_{avg}) and max-pooled features (F^s_{max}); 2, activation layer. The convolution layer is used to detect inner-structures of the feature map (F'). To augment the feature map, the max-pooled features and the average-pooled features are stacked together. Shortly, spatial attention is described as:

$$M_s = \sigma(Conv^{(3\times3)}([F^s_{avg}, F^s_{max}])) \quad (4)$$

Where σ indicates the sigmoid function, $Conv^{(3\times3)}$ denotes a convolution with the filter size is 3x3.

B. Pyramid Pooling Module

Pyramid Pooling Module (PPM) [12] exploits efficiently contextual information under multi-aspects with various filter sizes. PPM includes multi-pooling layers, up-sampling layers, and convolution layers. PPM observes features under different filter sizes, therefore, PPM can help to recognize low-dimension objects as well as low inner-structure.

The PPM feed input into multi-pooling layers and then following by convolution layers. After that up-sampling layer scales up feature maps from convolution layers. In detail, sizes of pooling layers are 1x1, 2x2, 4x4, 8x8 respectively. The different pooling size levels in the PPM abstract the feature maps with varied sizes. However, the difference in filter sizes set at a reasonable gap to pyramid features maintain certain correlations.

C. Attention-Spatial Convolution Model

ASC model includes 2 sub-sections: down-sampling section and up-sampling section. There are auxiliary connections between these two sub-sections. Fig. 2 shows a diagram of convolutional attention model. The down-sampling section comprises a sequence of convolution layers and attention blocks. The up-sampling section contains a pyramid pooling module, convolution layers, concatenation. Incorporation of the attention block and PPM allows ASC model deeper recognizing the spatial structure of low or high dimension objects. This combination outcomes disadvantages of the previous models in exploiting the inner structure of feature maps deeply. Besides, by arranging the attention blocks in the down-sampling section and the PPM in the up-sampling section can help ASC model avoid missing contextual information.

Auxiliary connections help to enlarge the size of feature maps as well as to restrict overfitting. The down-sampling decrease the size of the feature map along with layers; however, meaningful features increase. The auxiliary connections assist ASC model to segment tiny anomaly regions and augment contextual information in the up-sampling section. In addition, the auxiliary connections can help ASC model achieve higher performance than the previous models with a lower number of parameters.

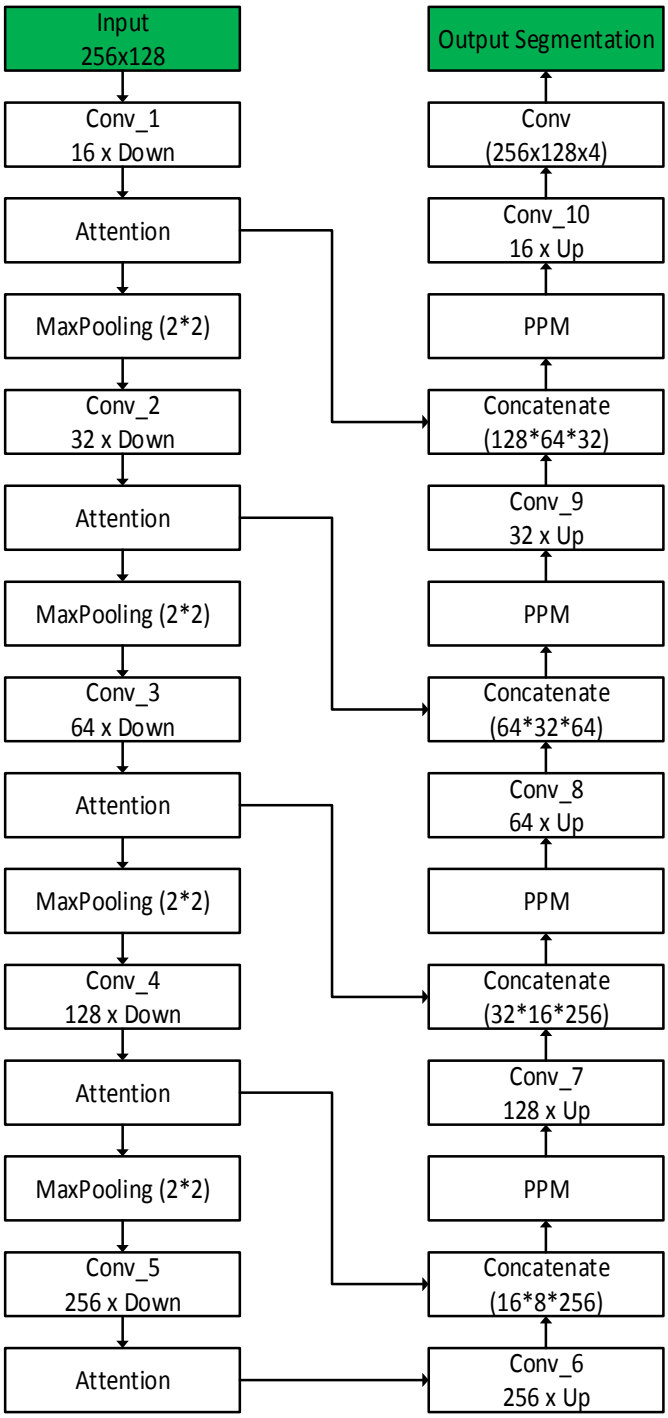

Fig. 2 Diagram of our model

Overall, the attention block can learn features many aspects by using two kinds of pooling layer simultaneously. The incorporation of PPM and the attention block can enhance ASC model deeply to identify anomaly regions which imply an association of nearby pixels. Another advantage of our work is the auxiliary connection recognizing small unusual areas as well as restrict overfitting, increase performance with a lower number of parameters.

III. Experimental Results

A. *Dataset, baselines, and experimental setting*

The Retinal Edema Dataset is the first medical image detection competition for Fundus lesions in China with the largest fundus lesion dataset currently, which combines AI and medical technology. With the data of 100 OCT and a volume of 128 images have been labeled by a professional ophthalmologist. Data is separated into training data and testing data with 85%, 15% respectively. So total training data is 7616 and testing data is 1344.

Bisenet [6] and Deeplab V3[7] are selected to compare with our method. Actually, Deeplab V3 is the famous model in 2017 for semantic segmentation. Moreover, In 2018, Bisenet was evaluated significantly for semantic segmentation.

Our model is set up with the following configuration: input size is 128x256, 5 convolution layers in the down-sampling section, the up-sampling section. Kernel size is 3x3 during the

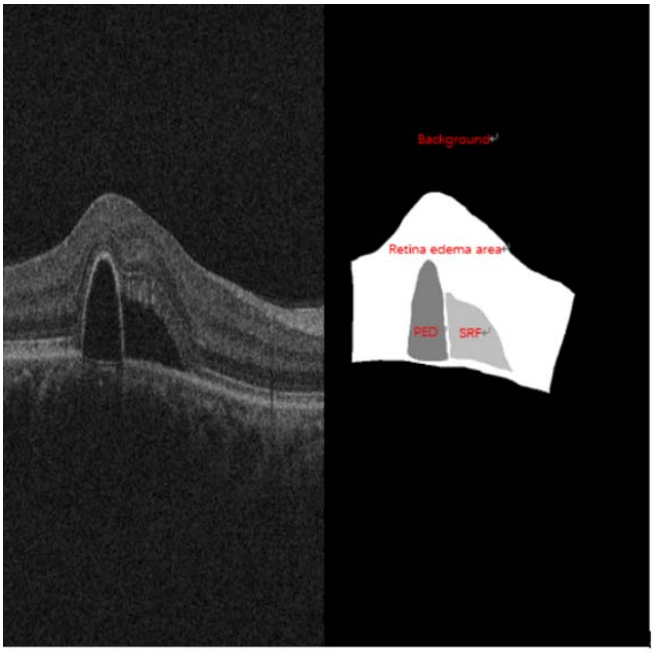

Fig. 3 Original image (left) and target image (right)

processing. For attention block, one hidden layer is emplemented for MLP. PPM includes 4 filter sizes such as 1x1, 2x2, 4x4, 8x8.

Dice coefficient [13] is used to estimate performance of the baseline models and our proposed. Dice coefficient aims to measure the overlap of two regions. Dice coefficient is calculated based on two areas X and Y as:

$$\text{Dice coefficient} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (5)$$

Where X is ground-truth area and Y predicted area. |X∩Y| represents the common elements between two areas X and Y.

B. *Results*

The Retinal Edema Dataset is the first medical image detection

Table 1: The result of our model and baselines

	Bisenet	Deeplab V3	ASC model
Parameter(million)	26.26	**2.14**	5.9
Dice coefficient(%)	76.85	76.19	**78.19**

Overall, ASC model produces a best result with 78.19% is higher Bisenet models [6] and Deeplab V3 [7] with margin of 1.34%, 2% respectively. Although the result of our model only is higher than BiseNet model as 1.34%, the number parameters in BiseNet model is fourfold than our model.

IV. CONCLUSIONS

The ASC model achieves a best performance for Retinal Edema Segmentation with reasonable low parameters. Conventional deep convolutional neural networks are integrated by the attention block and the pyramid pooling module to take advantage of contextual information. Our contribution comprises: 1, loss of contextual information is restricted by ordering the attention blocks in the down-sampling section and the PPM in the up-sampling section; 2, anomaly region with different scales is recognized effectively by the auxiliary connection. Besides, the auxiliary connection assists our model overcome overfitting, improve accuracy through gaining features in the up-sampling stage.

ACKNOWLEDGEMENT

This research is partially supported by the Ministry of Science and Technology under Grant Number 108-2634-F-008-004 through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

REFERENCES

[1] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," In *Proc of IEEE International Conference on Computer Vision*, 2017

[2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," In *Proc of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4681-4690, 2017.

[3] J. Kim, J. Janghyeok, P. Eunjeong, C. Sungchul, "Patent document clustering with deep embeddings," DOI: 10.13140/RG.2.2.20820.71042, 2018.

[4] T. Pham, Y. S. Lee, Y. B. Lin, T. C. Tai, and J. C. Wang, "Source separation using dictionary learning and deep recurrent neural network with locality preserving constraint," In *Proc of IEEE International Conference on Multimedia and Expo*, pp. 151-156, 2017.

[5] B. Jan, et al. "Deep learning in big data analytics: a comparative study," *Computers & Electrical Engineering*, vol. 75, pp. 275-287, 2019.

[6] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Lecture notes in Computer Science,* vol. 9531. Springer, pp. 234–241, 2015.

[7] Y. Changqian, W. Jingbo, P.Chao, G. Changxin, Yu, S. Nong, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," *arXiv preprint*, arXiv:1808.00897v1, 2018.

[8] L. Cheng, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation", *arXiv preprint*, arXiv:1706.05587, 2017.

[9] S. Woo , J. Park , J. Lee, I. S. Kweon, "CBAM: Convolutional block attention module," In *Proc of the European Conference on Computer Vision*, pp. 3-19, 2018.

[10] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: Architecture optimization and training," *International Journal. Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, pp. 26-30, 2016.

[11] S. Qiu, "Global weighted average pooling bridges pixel-level localization and image-level classification," *arXiv preprint*, arXiv:1809.08264v1, 2018.

[12] H. Wu, X. Gu, "Max-pooling dropout for regularization of convolutional neural networks," *arXiv preprint*, arXiv:1512.00242, 2015.

[13] H. Kaiming. Z. Xiangyu, R. Shaoqing, S. Jian, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *arXiv preprint*, arXiv:1406.4729, 2014

[14] V. Thada, V. Jaglan, "Comparison of Jaccard, Dice, Cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm," *International Journal of Innovations in Engineering and Technology*, vol. 2, no. 4, pp. 202, 2013.