Encoder-Decoder Attention Network for Lesion Segmentation of Diabetic Retinopathy

Shuanglang Feng[1], Weifang Zhu[1,2], Heming Zhao[1], Fei Shi[1], Zuoyong Li[2], and Xinjian Chen[1,3](✉)

[1] School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China
xjchen@suda.edu.cn
[2] Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University, Fuzhou 350108, China
[3] State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou 215123, China

Abstract. The segmentation of lesions such as retina edema, sub-retinal fluid and pigment epithelial detachment in optical coherence tomography (OCT) images is a crucial task for automated diagnosis of diabetic retinopathy. However, the multi-class lesion joint segmentation is very challenging due to the blurred boundary, complex structure, influence of noise, and the imbalanced class. In this paper, we propose a novel convolutional neural network with an encoder-decoder structure to perform joint segmentation of these three lesions. Unlike the common skip-connection employed in U-shape network for obtaining rich information from encoder feature map, we explore an encoder-decoder attention module (EDAM) via low-complexity non-local operation to capture more useful spatial dependency information between encoder feature and decoder feature. In this way, the network will take full advantage of the correlation information of the same stage feature and pay more attention to lesion areas. In order to capture large receptive fields and accurately segment small lesion, the modified light-weight residual network with dilated convolution is employed in encoding path. Besides, a hybrid loss, consisting of cross-entropy loss and multi-class Dice loss, is used to optimize our network. The proposed method was evaluated on a public database: AI-challenger 2018 for automated segmentation of retinal edema lesions, and achieved a compelling performance with less parameters compared to state-of-the-art networks.

1 Introduction

Diabetic Retinopathy (DR) is one of the main blinding diseases, affecting the normal life of approximately 34% of diabetic patients. DR may cause many symptoms that appear on the retina such as retina edema (RE), sub-retinal fluid (SRF), pigment

S. Feng, and W. Zhu—These authors contributed equally to this work.

H. Fu et al. (Eds.): OMIA 2019, LNCS 11855, pp. 139–147, 2019.
https://doi.org/10.1007/978-3-030-32956-3_17

epithelial detachment (PED). Optical coherence tomography (OCT) images are widely used in ophthalmology clinic for the diagnosis of retinal diseases. Therefore, the automatic segmentation of lesions in OCT images plays a key role in the diagnosis and treatment of DR. However, the main challenge for this task lies in the following factors: (1) Joint segmentation of multiple types of lesions is difficult due to the extreme imbalance of the data distribution between different lesions. (2) The boundary of the retina edema area (REA) is blurred and difficult to determine. (3) The influence of speckle noise and vascular artifacts is severe.

In recent years, many segmentation studies on DR lesion have been proposed. Most of these methods such as graph search based methods [1, 2], kernel regression based methods [3] have two stages: retinal layer segmentation, lesion delineation. The computational bottleneck caused by algorithm optimization makes it urgent to develop an end-to-end solution. Recently, many deep learning methods based on convolutional neural networks (CNN) [4] have been applied to medical image analysis. Guha et al. [5] proposed a ReLayNet with position indices pooling for retinal layer and fluid pocket segmentation. Freerk et al. [6] utilized typical U-shape neural network for segmentation of macular edema. Most of these CNN-based approaches only focus on single type lesion. To the best of our current knowledge, there are no methods based on CNN for joint lesion segmentation in DR OCT images, which is always challenging to jointly segment imbalanced medical data for CNN based on encoder-decoder architecture. In this paper, we design a novel and efficient network to address these problems.

The skip-connection of U-Net [4] is an ingenious design, which can combine the encoder feature to make up for the information loss caused by downsampling. However, simple skip-connection ignores contextual information and is an indiscriminate combination of semantic information that will introduce noise of irrelevant clutters. Previous work [7, 8] overlooked this important detail. Although [9, 10] proposed a global convolutional network (GCN) between encoder and decoder, it can’t capture the global information in the real sense and ignore the spatial correlation. General non-local model (NLM) [11] was applied in video classification and semantic segmentation, which utilized a self-attention mechanism to get the approximate autocorrelation information. In this paper, a novel encoder-decoder attention module (EDAM) based on non-local operation is employed to generate approximate cross-correlation information between encoder feature and decoder feature. In this way, the network can enhance the correlated responses of focused object and weaken the uncorrelated responses in global view through a controllable information flow from encoder. Furthermore, for non-local operation, we explore a low-complexity representation to handle high computational complexity issue. Besides, in order to obtain the high-resolution feature map and accurately segment small lesion, we improve a lightweight residual network [12] with dilated convolution [13] as backbone network to extract feature and employ a hybrid loss consisting of cross-entropy loss and multi-class Dice loss to alleviate the imbalanced data problem.

Consequently, our main contributions include: (1) An efficient encoder-decoder attention network is proposed for joint lesion segmentation in DR OCT images. (2) The proposed encoder-decoder attention module (EDAM) can capture richer global features and model spatial correlation between encoder feature and decoder feature. (3) We achieve impressive results with less parameters compared to state-of-the-art networks on public database: AI-challenger 2018 for automated segmentation of retinal edema lesions.

2 Method

2.1 Proposed Network Architecture

Figure 1 is an overview architecture of our proposed encoder-decoder attention network for joint segmentation of three DR lesions (REA, SRF, PED). In order not to lose the information of small lesions during the downsampling, we improve a residual network with dilated convolution as the encoder to extract high resolution feature map. The dilated convolution with rate of 2 is employed in block3 and rate of 4 is employed in block4 like [14]. Therefore the output size of feature map from encoder is 1/8 of input image. For the convenience of skip-connection, the first downsampling is performed by a 3 × 3 convolution layer with a stride of 2 after a 7 × 7 convolution layer and a bottleneck layer [12]. The next two downsampling layers are in the first bottleneck layer of block1 and block2 respectively. Note that the channel expansion rate is set to 2. In decoder part, bilinear interpolation operation is applied in three upsampling layers to quickly restore the original image size. The boundary refinement (BR) blocks [9] are used to refine the edges of the feature map, which consist of two convolution layers with residual design. It is worth noting that we employ an EDAM between each corresponding stage of encoder and decoder to capture more correlated information about prediction feature map from encoder path.

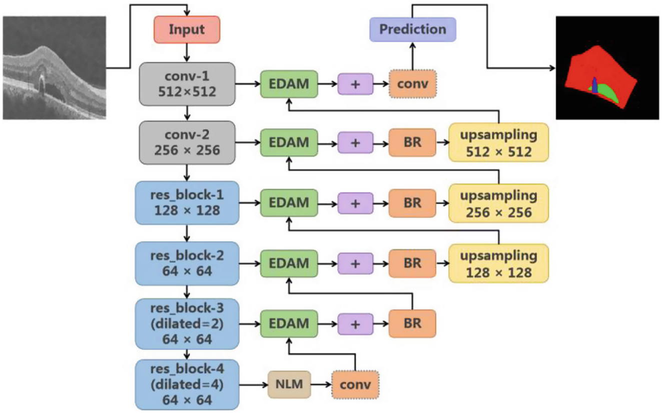

Fig. 1. An overview of our proposed network architecture. EDAM and NLM represent encoder-decoder attention module, general non-local module, respectively. BR and conv represent boundary refinement block and convolutional layer, respectively.

2.2 Encoder-Decoder Attention Module

Attention mechanism is widely used in natural language processing (NLP) [15] and computer vision (CV) [16] field, which can draw global information and obtain rich feature. In this work, we propose an encoder-decoder attention module (EDAM) shown in Fig. 2. To model the spatial correlation over the global view between encoder feature and decoder feature via non-local operation. Here, the feature $\mathbf{E} \in \mathbb{R}^{C \times H \times W}$ from encoder path generates two feature maps $\mathbf{V}$ and $\mathbf{K}$ via two convolution layers with 1×1 filters, respectively, where $\{\mathbf{V}, \mathbf{K}\} \in \mathbb{R}^{C' \times H \times W}$. Meanwhile, the feature map $\mathbf{Q} \in \mathbb{R}^{C' \times H \times W}$ is generated by $\mathbf{D} \in \mathbb{R}^{C' \times H \times W}$ from decoder path through the convolution layer with 1×1 filter. C' is the channel number of feature map $\mathbf{D}$, which is less than C. Then we reshape $\mathbf{V}$, $\mathbf{K}$ and $\mathbf{Q}$ to $\mathbb{R}^{C' \times N}$, where $N = H \times W$. After that, we use a hyperparameter factor $\alpha = \sqrt{C' \times H \times W}$ to normalize the result of matrix multiplication between the transpose of $\mathbf{K}$ and $\mathbf{Q}$ and generate the pixel-wise affinity attention map $\mathbf{A} \in \mathbb{R}^{N \times N}$:

$$\mathbf{A} = \frac{\mathbf{K}^{\mathrm{T}}\mathbf{Q}}{\alpha} \tag{1}$$

Then we perform matrix multiplication between $\mathbf{V}$ and attention map $\mathbf{A}$ to obtain a final feature map $\mathbf{H}$ and reshape it to $\mathbb{R}^{C' \times H \times W}$, here, $\mathbf{H}$ is feature map that has been weighted by correlated contextual information between encoder feature map and decoder feature map:

$$H_{ic} = \sum_{j=0}^{N} A_{ji} V_{jc} \; \{i, j \in [1, 2, \ldots N], \; c \in [1, 2, \ldots C']\} \tag{2}$$

Where H_{ic} is the weighted sum of all j according to its affinity with i. After this, $\mathbf{H}$ is input to a convolution layer with 1×1 filter and added to decoder path local feature $\mathbf{D}$ to enhance the correlative responses and global representation. In this way, any position in the encoder feature map is aggregated with all other positions in the decoder feature map from the same stage through self-adaption attention maps.

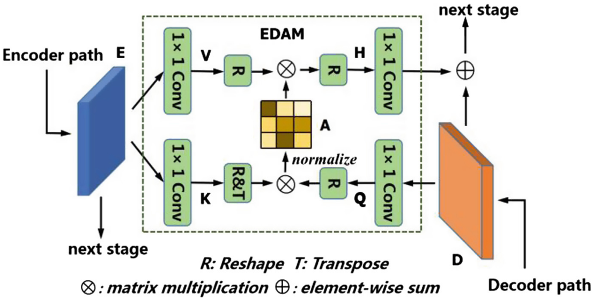

Fig. 2. The detail of encoder-decoder attention module

2.3 Low-Complexity Representation

Representing the relationship between any two pixels requires a complex matrix multiplication operation to obtain a huge attention map, its complexity is $\mathcal{O}(N^2)$ in both time and space, where $N = H \times W$ indicates the spatial dimension of feature map. Because of the high resolution of feature map in semantic segmentation task, we cannot afford for directly implementing EDAM with our limited GPU memory. Fortunately, we explore an efficient and alternative method to achieve the same target with the associative law of matrix multiplication. Note that we obtain the final feature map **H** through two matrix multiplication operations in Sect. 2.2.

$$\mathbf{H} = \frac{\mathbf{V}\left(\mathbf{K}^{\mathrm{T}}\mathbf{Q}\right)}{\alpha} = \frac{\left(\mathbf{V}\mathbf{K}^{\mathrm{T}}\right)\mathbf{Q}}{\alpha} \tag{3}$$

According to the associative law, we could perform $\mathbf{Z} = \mathbf{V}\mathbf{K}^{\mathrm{T}} : \mathbb{R}^{C' \times N} \times \mathbb{R}^{N \times C'} \to \mathbb{R}^{C' \times C'}$ first and then calculate $\mathbf{ZQ}$, where C' is channel of decoder path feature map and much smaller than N. These operations greatly reduce the time and space complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(C'^2)$, and can be easily embedding into other encoder-decoder network. We apply this compatibility representation to each stage between encoder and decoder to replace the original skip-connection except for the bottom of network which employs a NLM.

2.4 Loss Function

To alleviate the problem that Dice loss is sensitive to small structures or absent classes, we employ a hybrid loss consisting of cross-entropy loss and multi-class Dice loss to preform joint lesion segmentation. The total loss can be expressed as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{Dice} + \lambda \mathcal{L}_{CE} \\ &= 1 - \frac{1}{C}\sum_{c=0}^{C-1} \frac{2\sum g * p + \varepsilon}{\sum (g + p) + \varepsilon} - \lambda \frac{1}{N} \sum g \ \log(p) \end{aligned} \tag{4}$$

Where λ is a weighted coefficient between Dice loss $\mathcal{L}_{Dice}$ and cross-entropy loss $\mathcal{L}_{CE}$ and set to 1 in our work, $p \in [0, 1]$ is the predicted probability, $g \in \{0, 1\}$ is the true label, N is the spatial dimension of feature map, C is a sum of the total number of lesion classes and one background class, and ε is a small smoothing factor.

3 Experiments

Databases: We evaluated our proposed network in a public database: AI-challenger 2018 for automated segmentation of retinal edema lesions, which contains of 100 cubes with the size of 1024 × 512 × 128 and is annotated manually by experts. All of these cubes contain REA, but PED and SRF only involve some cubes and are surrounded by REA. Since the annotation problem, 68 cubes was chosen for training and 30 cubes was chosen to averagely divided into Part A and Part B for testing. In the databases, background accounts for 92.94% of all the voxels, REA, SRF and PED takes up 6.19%, 0.84% and 0.03%, respectively, which indicates the segmentation of SRF and PED will suffer from the extremely imbalanced data distribution problem. Besides, the 3D OCT cube is a huge burden for limited GPU memory. To address these problems, we extract 2D B-scan image as the input image of segmentation task.

Patch Extraction and Data Augmentation: The size of 2D B-scan slice is 1024 × 512. According to our statistics, the lesion area always lie in a square of size $l \times l$, where $l = 512$. Therefore, for the slice containing the lesion, we randomly cropped a $l \times l$ patch which include whole lesion, and for the slice with only background, a $l \times l$ patch was randomly cropped on single slice. After that a single slice containing PED was reapplied to random cropping 9 times to balance data. Now we got a more balanced training data than before with about 12300 patches. In the inference phase, we applied a sliding window strategy with the window size $l \times l$ to take tiles, and the stride is $l/4$. To alleviate the border effects in segmentation task, we put more weight on the middle of prediction feature map during the splicing process.

Implementation: We employed Keras and tensorflow to implement our proposed method. The optimizer was stochastic gradient descent (SGD) with the "poly" learning rate policy. The basic learning rate and momentum were set to 0.01 and 0.9, respectively. The batch size and epoch were set to 12 and 40, respectively. For data augmentation, we only applied online random left-right flipping. All of these were preformed on three NVIDIA Tesla K40 GPUs with 12 GB memory.

4 Evaluation and Results

We compared feature maps from common skip-connection and EDAM. Figure 3 displays the visualization of feature maps, in which shows that common skip-connection introduces unnecessary noise and has no any discrimination for feature. However, our EDAM can enhance the correlated responses of focused object via modeling long-range spatial dependencies in global view and suppress irrelevant noise.

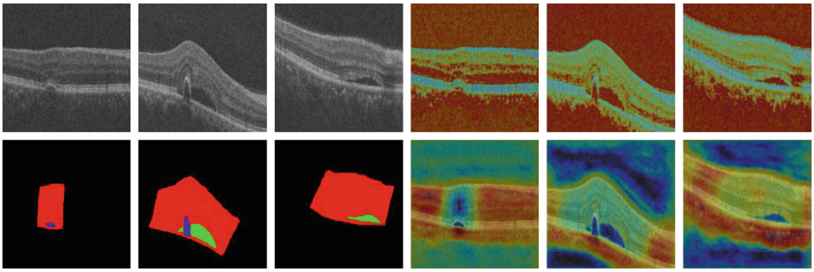

Fig. 3. The visualization of feature maps. Left top: original patch. Left bottom: ground truth. Right top: feature map from common skip-connection. Right bottom: feature map from EDAM.

We carried out comprehensive experiments on Part A and Part B OCT cubes with the measure of average Dice scores in Table 1. In order to evaluate the effectiveness of our proposed EDAM, an ablation experiment was conducted. Baseline represents our network employing the common skip-connections, EDAM represents encoder-decoder attention module inserted in each stages, we observe that it yields the more gain when EDAM is inserted into network comparing with Baseline network. Moreover, there is only a little increase of parameters because EDAM is inserted in relatively shallow layers with a few channels. We also report the performance of our proposed network compared with other state-of-the-art segmentation methods. To be fair, we employed the same hybrid loss to these comparable methods. It can be seen that all the Dice

Table 1. The performance of different segmentation models and our proposed method on Part A and Part B OCT cubes, measured with average global Dice scores (±standard deviation).

Method	Part A				Part B				#Para
	Dice (%)				Dice (%)				
	REA	SRF	PED	Ave	REA	SRF	PED	Ave	
U-Net	73.23 (±14.6)	60.96 (±37.3)	41.35 (±38.7)	58.51	75.25 (±13.6)	60.13 (±38.1)	28.59 (±26.2)	54.65	31.04 M
Attention U-Net [18]	74.56 (±14.8)	62.07 (±36.7)	40.94 (±35.3)	59.19	76.53 (±16.5)	61.34 (±35.9)	29.67 (±23.4)	55.84	29.82 M
PSPNet [14]	75.99 (±14.3)	59.42 (±38.4)	42.76 (±36.7)	59.39	77.58 (±14.1)	59.42 (±37.3)	31.69 (±25.4)	56.23	46.77 M
Res-FCN [10]	74.15 (±16.0)	62.88 (±35.9)	43.50 (±37.22)	60.17	76.83 (±13.4)	61.24 (±36.3)	29.51 (±24.4)	55.86	37.05 M
V-Net [17]	76.23 (±15.2)	57.33 (±36.2)	39.48 (±36.24)	57.68	77.04 (±12.7)	58.45 (±38.0)	26.16 (±23.1)	53.88	45.60 M
Baseline	75.58 (±14.3)	64.67 (±35.7)	39.59 (±35.2)	59.95	76.25 (±14.6)	62.14 (±36.7)	31.82 (±24.2)	56.74	19.79 M
Baseline + EDAM	**76.51 (±14.2)**	**67.75 (±35.2)**	**45.14 (±34.9)**	**63.13**	**78.04 (±13.1)**	**64.83 (±37.4)**	**32.01 (±23.6)**	**58.29**	**19.87 M**

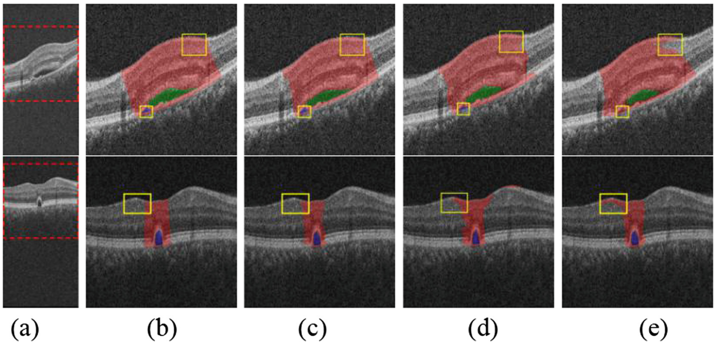

Fig. 4. Examples of segmentation results. (a) Original OCT B-scans, (b) Ground truth, (c) The segmentation results of Baseline + EDAM, (d) The segmentation result of Baseline, (e) The segmentation result of Res-FCN [12]. Note that red area, green area and blue area represent REA, SFR and PED, respectively (Color figure online).

scores of REA, SRF, PED and the average Dice score of our method were superior to all the other methods both on Part A and Part B. Especially, the number of parameters in our network (20 M) is only 65% of U-Net (31 M), while it achieved a dramatic improvement of 6% in average Dice scores. And the similar phenomena has been also shown in other results. Our method also achieved a significant improvement (student's t-test, p-value = 0.04) on SRF segmentation with a 3%–5% Dice increase to the second-best one. Due to the false positive prediction of cube without SRF and PED, the variance is relatively large. All these performances indicated that our network is more efficient than all compared methods, which benefits from EDAM capturing global correlation information between encoder and decoder.

Two examples of segmentation results are shown in Fig. 4. In the first row, the size of PED areas is too small to Res-FCN, while our method could segment it precisely. In addition, for relatively large REA lesion, our method achieved a very similar result to the ground truth, which benefited from EDAM can aggregate the correlative information between encoder and decoder in global view. It can be also seen that the proposed method can eliminate more false positives than Res-FCN method from second row. Overall, the encoder-decoder attention network is capable of the accurate and effective segmentation for DR joint lesion due to its strong global correlative information aggregation ability.

5 Conclusion

We proposed a novel network with encoder-decoder attention module for the multi-class joint lesion segmentation of diabetic retinopathy. The proposed method models the long-range spatial dependencies and captures more correlative contextual information between encoder feature and decoder feature. Experiment results showed that it has a great potential for imbalanced data medical image segmentation with its efficient and compelling performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) (61622114, 81401472) and Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University (No. IIC1702).

References

1. Shi, F., et al.: Automated 3-D retinal layer segmentation of macular optical coherence tomography images with serous pigment epithelial detachments. IEEE Trans. Med. Imaging **34**(2), 441–452 (2015)
2. Sun, Z., et al.: An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images. Sci. Rep. **6**, 21739 (2016)
3. Chiu, S.J., et al.: Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed. Opt. Express **6**(4), 1172–1194 (2015)

4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. Roy, A.G., et al.: ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. BOE **8**(8), 3627–3642 (2017)
6. Venhuizen, F.G., et al.: Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomed. Opt. Express **9**(4), 1545–1569 (2018)
7. Badrinarayanan, V., et al.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. PAMI **39**(12), 2481–2495 (2017)
8. Jégou, S., et al.: The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: CVPR Workshop, pp. 11–19 (2017)
9. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: CVPR, pp. 4353–4361 (2017)
10. Liu, Z., et al.: Towards clinical diagnosis: automated stroke lesion segmentation on multi-spectral MR image using convolutional neural network. IEEE Access **6**, 57006–57016 (2018)
11. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR, pp. 7794–7803 (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2015)
13. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint. arXiv:1511.07122 (2015)
14. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR, pp. 2881–2890 (2017)
15. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
16. Hu, J., et al.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
17. Milletari, F., et al.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, pp. 565–571 (2016)
18. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint. arXiv:1804.03999 (2018)