Image Segmentation with Pyramid Dilated Convolution Based on ResNet and U-Net

Qiao Zhang[1,2], Zhipeng Cui[1], Xiaoguang Niu[1], Shijie Geng[1], and Yu Qiao[1(✉)]

[1] Intelligence Learning Laboratory, Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai, China
qiaoyu@sjtu.edu.cn

[2] The Fu Foundation School of Engineering and Applied Science, Columbia University, New York, USA

Abstract. Various deep convolutional neural networks (CNNs) have been applied in the task of medical image segmentation. A lot of CNNs have been proved to get better performance than the traditional algorithms. Deep residual network (ResNet) has drastically improved the performance by a trainable deep structure. In this paper, we proposed a new end-to-end network based on ResNet and U-Net. Our CNN effectively combine the features from shallow and deep layers through multi-path information confusion. In order to exploit global context features and enlarge receptive field in deep layer without losing resolution, We designed a new structure called pyramid dilated convolution. Different from traditional networks of CNNs, our network replaces the pooling layer with convolutional layer which can reduce information loss to some extent. We also introduce the LeakyReLU instead of ReLU along the downsampling path to increase the expressiveness of our model. Experiment shows that our proposed method can successfully extract features for medical image segmentation.

Keywords: Deep learning · Semantic image segmentation · Convolutional neural network · Medical image · Ultrasound Nerve Segmentation

1 Introduction

It has been widely accepted that CNNs have an impressive performance in computer vision tasks in recent years. CNNs have also been widely applied to the field of medical image segmentation and gain great popularity.

Brebisson et al. [1] apply the CNNs for anatomical brain segmentation and get good result. Zhang et al. [2] has designed deep convolutional neural networks for segmenting isointense stage brain tissues using multi-modality MR images. Li et al. [3] use the CNNs to learn the intrinsic image features of lung image patches. However, a lot of methods were based on the sliding-window technique which was proposed by Ciresan et al. [4]. This method could lead to storage overhead and ineffectiveness if we process a high resolution image. This method would

D. Liu et al. (Eds.): ICONIP 2017, Part II, LNCS 10635, pp. 364–372, 2017.
https://doi.org/10.1007/978-3-319-70096-0_38

also lead to hierarchical global information loss. Long et al. [5] proposed Fully Convolutional Networks (FCN), which is based on VGG-16 [6]. FCN is an end-to-end network which can effectively solve the overstorage problem. It is widely acknowledged that the deeper architecture would achieve better performance. However, the training error rate in a deeper plain network would even be higher because the gradient would disappear more easily in a deeper architecture. He et al. [7] proposed deep residual network which makes the deep network training possible and achieves compelling accuracy. Furthermore, the repeated pooling layers and convolution strides in traditional CNNs would largely reduce receptive filed which is quite important for dense prediction tasks. The deconvolution process would not successfully recover the detail information which are lost in the downsampling process. Fisher et al. [8] proposed dilated convolution, which can effectively enlarge receptive field without losing resolution. It has been proved to improve the performance in VGG-16 network and accelerate convergence.

In this paper, we propose a new network based on ResNet and U-Net [9]. It can effectively combine the features from shallow and deep layers through multi-path confusion. We design a new structure called pyramid dilated convolution, which aims to exploit global context features with multi-scale. Furthermore, we apply the LeakyReLU [10] instead of ReLU [11] at downsampling path to increase the expressiveness of our model. Our network was applied to the Ultrasound Nerve Segmentation task and achieved good result.

2 Methodology

2.1 Pyramid Dilated Res-U-Net

In this paper, we propose a new segmentation architecture named Pyramid Dilated Res-U-Net. It is based on ResNet and U-Net with pyramid dilated convolution unit. This network structure is illustrated in Fig. 1. We use the deformed residual unit as shown in Fig. 2(b) to extract the feature map. We apply U-Net structure to combine multi-path feature maps from intermediate and deep layers. We refine the deep feature map from the 4th block of ResNet with multi-scale

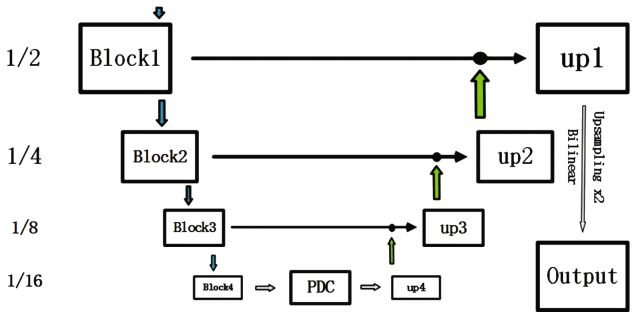

Fig. 1. Pyramid Dilated Res-U-Net

dilated convolution to fuse global context information. As for the first block of ResNet, we apply filter size of 5 instead of 3. Output from fusion is upsampled by bilinear interpolation with a factor of 2 to achieve an end-to-end training.

2.2 BN-LeakyReLU Residual Unit

The basic residual unit in ResNet is shown in Fig. 2(a). The following form denotes the basic unit:

$$y_k = F(x_k; W_l) + h(x_k) \tag{1}$$

$$x_{k+1} = f(y_k) \tag{2}$$

where x_k and x_{k+1} represent the input and output of the k-th unit, and h is an identity mapping function, F is a residual function and f represents activation function. He et al. [12] proposed that pre-activation of the weight layers (Fig. 2(b)) would be much easier to train and generalize better than post-activation structure (Fig. 2(a)). According to [12], we can use the chain rule of backpropagation [13] to get the following form:

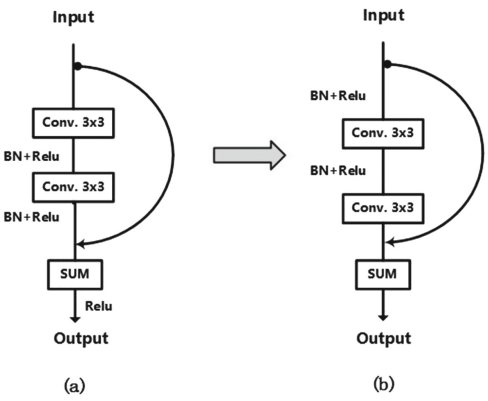

Fig. 2. Basic residual unit (a) and deformed residual unit (b).

$$\frac{\partial \varepsilon}{\partial x_k} = \frac{\partial \varepsilon}{\partial x_K}\frac{\partial x_K}{\partial x_k} = \frac{\partial \varepsilon}{\partial x_K}(1 + \frac{\partial}{\partial x_k}\sum_{i=k}^{K-1} F(x_i, W_i))) \tag{3}$$

where ε denotes the loss function, x_k denotes the feature of k-th layer and x_K denotes the feature of K-th layer. This structure could propagate information directly and through weight layers. Therefore, we implement this technique into our Network (Fig. 2(b)). As for the first block, we use a filter size of 5 instead of 3 in order to get a better basic feature map. The activation function is LeakyReLU instead of ReLU. LeakyReLU is denoted as the following form.

$$f(x) = \begin{cases} \alpha x & \text{if}(x < 0) \\ x & \text{if}(x > 0) \end{cases} \tag{4}$$

It allows a small, non-zero gradient when the unit is not active. So it would enlarge the expressiveness of our network to some extent.

2.3 Pyramid Dilated Convolution Unit

Fisher et al. [8] proposed dilated convolution which can exponentially enlarge receptive field without losing resolution. It is widely known that the receptive field affects the extent to which we exploit the context information. The context information is of great importance for accurate segmentation. However, Zhou et al. [14] presents that the actual receptive field of CNNs in deep layer is much smaller than the theoretical calculation.

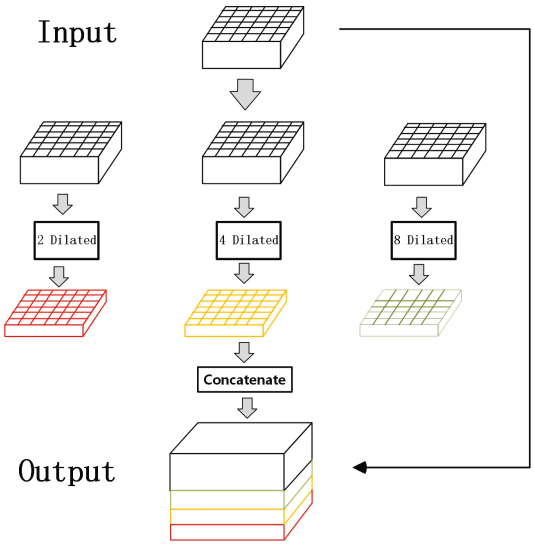

Fig. 3. Given an input feature map, we separately use dilated convolution with different factors to extract information. The corresponding three extracted feature maps are then concatenated with the input feature map to get the output.

We address this issue by designing a new structure, called Pyramid Dilated Convolution Unit shown in Fig. 3. We apply dilated convolution with 2, 4, 8 factors at the 4th block of ResNet to refine the feature map. It can effectively extract global context information through multi-scale dilated convolution. This unit could also enlarge receptive field without losing resolution.

The refined feature maps of different factors generated by dilated convolution are finally concatenated together with input image. Through concatenation operation, we can combine the raw feature information and the information in hierarchical structure. Then the fused feature map is fed to upsampling process. Experiment results show that the Pyramid Dilated Convolution Unit can successfully refine feature map with global context information.

2.4 Multi-path Fusion

As we know the feature map in the deep layer is usually of small size and it would lead to drastically information loss if we upsample directly. The low-level features embedded in intermediate layers are very necessary for accurate high resolution segmentation. In our network, we implement the U-Net-like structure to deal with multi-path fusion. Therefore, the shallow layer information and deep layer information together make the final segmentation more reliable. Specifically, feature map from the 5th block of ResNet is fed to the ReLU-Conv Unit (Fig. 4). This unit could be used to fine tune the weights effectively. The output of it is upsampled by bilinear interpolation and then concatenates the feature map from 4th block. In this way, we get fused output of half the input image size.

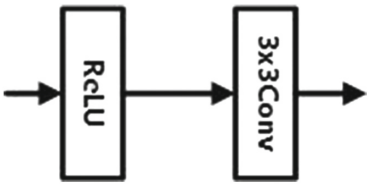

Fig. 4. ReLU_Conv Unit

3 Experiments

Our proposed method is applied on the segmentation problem of medical image. The method is evaluated in the Ultrasound Nerve Segmentaiton datasets and it achieves good result.

3.1 Implementation Details

Our network is based on top of keras with the backend of tensorflow. We implement data augmentation method to generate more training data. Specifically, we adopt small rotation, translation, random resize and random mirror. Inspired by [9], we use the "Adam" gradient descent optimizer with 0.00002 learning rate. For the training process, we assume that "batchsize" is of great importance because it affects the stability of the gradient and batch normalization [15]. However, we set the "batchsize" to 12 during training because of limitation of physical memory on GPU.

3.2 Ultrasound Nerve Segmentation

Ultrasound Nerve Segmentation task is required to identify nerve structures called the Brachial plexus in ultrasound images. This help inserting a patient's

Table 1. As for the network, baseline is ResNet54 (with ReLU and Pyramid Dilated Convolution). In our test, $\alpha = 0.2$ yields the best corresponding this network structure.

Parameter α	Dice coefficient(%)
ResNet54 (without LeakyRelU)	64.21
ResNet54 (with $\alpha = 0.1$)	67.12
ResNet54 (with $\alpha = 0.2$)	69.15
ResNet54 (with $\alpha = 0.3$)	65.26
ResNet54 (with $\alpha = 0.4$)	63.17

pain management catheter. The dataset are consisted of grayscale images with the corresponding binary masks. However, the dataset contains quite a lot contradictory images, therefore we pre-process the images and keep 4102 training images out of the 5500 in the end. The original images have a size of 580×420, we resize the images into 160×128 since the images are quite noisy and limitation of our memory resources. For the evaluation part, we use dice coefficient as a loss and also try binary cross-entropy. The two methods get roughly the same result.

To evaluate our network, we conduct experiments with several different settings. As for downsampling, we do experiment with pooling downsampling and convolution downsampling. We try different alpha of LeakyReLU in the downsampling process (Table 1).

Table 2. Deeper structure could yield better performance. However, deep network would be harder to train and occupy more resources. So in our experiment, we choose to use the ResNet54. PDC means Pyramid Dilated Convolution Unit.

Depth of ResNet	Dice coefficient(%)
ResNet34+PDC	68.52
ResNet54+PDC	69.15
ResNet72+PDC	69.31
ResNet101+PDC	69.39

It is widely known that deeper neutral networks could yield better segmentation accuracy, however the deep architecture could result in astounding cost of training time and GPU resources. We conduct experiments for various depths of deformed ResNet of 34,54,72,101 as shown in Table 2. We try different filter size to extract features of first block. We find that filter size of 5 could yield a better result than 3 and 7 in our problem.

We also compare Dilated Res-U-Net with other architectures (Table 3). Figure 5 presents the segmentation results of ultrasound nerve images with

Table 3. In this table, fs means filter size of the first block of ResNet. All experiments are on the preprocessed dataset and the U-Net experiment is based on the original dataset.

Method	Dice coefficient(%)
ResNet54+PDC+fs3	69.01
ResNet54+PDC+fs5(Ours)	**69.15**
ResNet54+PDC+fs7	69.11
ResNet54+PDC+pooling	68.73
ResNet54(fs3)	64.52
U-Net(without prepocess)	56.00

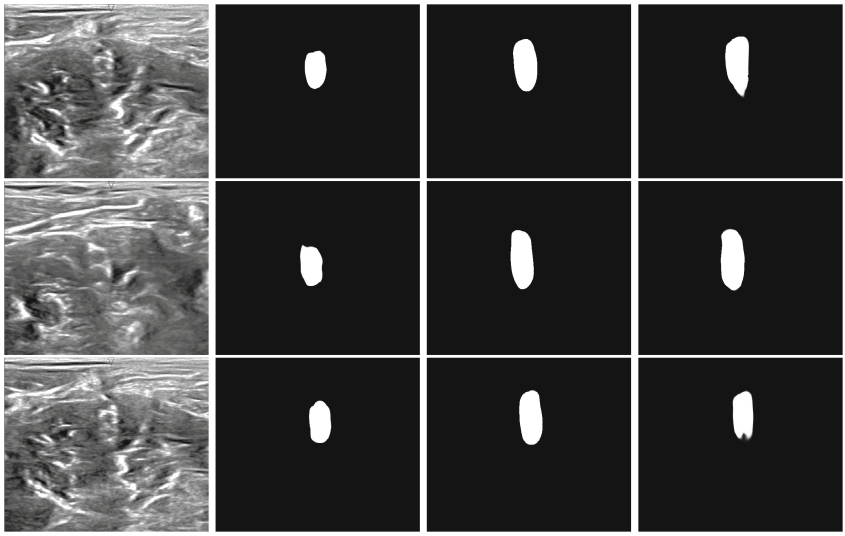

Fig. 5. Samples of ultrasound nerve segmentation with different CNNs. From left to right: (a) Input image, (b) U-Net, (c) Dilated-Res-U-Net, (d) Dilated-Res-U-Net(without PDC). PDC means Pyramid Dilated Convolution Unit

different CNNs. U-Net is restored following the link https://github.com/jocicmarko/ultrasound -nerve-segmentation. Figure 5 shows that Dilated-Res-U-Net could get a more complete structure than the network without Pyramid Dilated Convolution Unit. Table 3 demonstrates that this structure could effectively improve the accuracy by 4.6%. Therefore, the Pyramid Dilated Convolution Unit can successfully refine feature map with global context information.

4 Conclusions

In this paper, we have proposed an effective semantic segmentation network based on ResNet and U-net. We have developed a new structure Pyramid

Dilated Convolution Unit for exploitation of global context information. This unit also enlarges the receptive field without losing resolution. We also introduce LeakyReLU in the downsampling process instead of ReLU. We designed a structure without pooling operation and conduct experiment of different filter size in the extraction of basic features. Experiment results on Ultrasound Nerve Segmentation dataset show that our proposed method could effectively extract features in medical image for segmentation.

Acknowledgments. This research is partly supported by NSFC (No: 61375048).

References

1. Brebisson, A.D., Mountana, G.: Deep neural networks for anatomical brain segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2015)
2. Zhang, W., Li, R., Deng, H., Wang, L.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage **108**, 214–224 (2015)
3. Li, Q., Cai, T., Wang, X., Zhou, Y., Feng, D.: Medical image classification with convolutional neural network. In: the 13th International Conference on Control Automation Robotics & Vision (ICARCV). IEEE (2014)
4. Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J.: Flexible, high performance convolutional neural networks for image classification. In: Twenty-Second International Joint Conference on Artificial Intelligence (2011)
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ArXiv preprint arXiv:1409.1556 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition (2016)
8. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2015)
9. 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). doi:10.1007/978-3-319-24574-4_28
10. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
11. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010) (2010)
12. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). doi:10.1007/978-3-319-46493-0_38
13. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)

14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856 (2014)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)