SEMANTIC SEGMENTATION WITH MULTI-PATH REFINEMENT AND PYRAMID POOLING DILATED-RESNET

Zhipeng Cui[1], *Qiao Zhang*[1], *Shijie Geng*[1], *Xiaoguang Niu*[1], *Jie Yang*[1], *Yu Qiao*[*1]

[1]Institute of Image Processing and Pattern Recognition
Department of Automation, Shanghai Jiao Tong University, China

ABSTRACT

Recently, fully convolutional network (FCN) and dilated convolution have shown significantly improvment in semantic segmentation task. Deep residual network (ResNet) has shown strong ability in object recognition. However, FCN-based methods utilize intermediate layers and spatial context information ineffectively. Repeated 2-step striding in ResNet is harmful for segmentation tasks. In this paper, we propose a new semantic segmentation method based on FCN and ResNet. Here, we combine the dilated convolution designed for semantic segmentation with residual unit to enlarge receptive field of ResNet. Meanwhile, "pre-activation" method is used in dilated residual unit. We propose a new segmentation architecture which intergrate multiple intermediate layers and global context information. We extract low-level features of intermediate layers with multi-path refinement which consists of relu-conv unit and chained residual pooling. Global context is gathered by a pyramid pooling method which is connected to the final output of ResNet. Outputs of these 2 modules are then fused to reach high-resolution prediction. In this way, global and local information from pyramid pooling can be enhanced by multi-path refinement. Fully-connected conditional random field is added as a "post-processing" after fusion to receive accurate boundary performance. Our proposed approach achieves 74.7% IoU on Cityscapes benchmark.

Index Terms— Segmentation, Residual Networks, Dilated convolution, Multi-path refinement, Pymraid pooling

1. INTRODUCTION

These years, image semantic segmentation is becoming increaslingly important among various of computer vision tasks. Goal of semantic segmentation is to allocate each pixel a corresponding category label. It is associated with varieties of potential development in the near future such as robot intelligence, artificial management system, auto-driving and etc.

There have been so many trials on dense pixel prediction with various network. Badrinarayanan et al. [1] proposed SegNet which combined layers as encoder and decoder. Long et al. [2] trained a end-to-end fully convolutional network (FCN) for dense pixel prediction. Those networks were implemented based on VGG-16 which had remarkable performance but were limited by shallow architecture compared to the deep residual network (ResNet) [3]. ResNet can achieve compelling accuracy and efficiently combine information with extremely deep architectures. Key challenge of adapting classification CNNs to semnantic segmentation task is that invariance required for classification will cause reduced semantics in higher layers. Meanwhile, pooling operations and 2-step striding convolution in classification networks will shrink receptive field which is intended to be larger in semantic segmentation tasks.

As for dense prediction task, Resnet which has best performance on object recognition could also lose fine structure due to repeated 2-step striding convolution . Low-level features are very necessary for accurate high resolution semantic segmentation on the boundaries. Hyeonwoo et al. [4] proposed a deep deconvolution network but coarse output could not easily be recovered by deconvolution. Fisher et al. [5] proposed a new model for semantic segmentation based on dilated convolution which can enlarge the receptive field and extract more context information. Dilated convolution was also implemented in DeepLab [6]. This technique has been proved that it can improve the performance in VGG-16 and speed up convergence.

Another challenge for presented FCN-based method is lack of making use of intermediate layers and global context information for semantic segmentation. FCN and DeConvNet methods both combine feature maps from middle layer. Guosheng et al. [7] proposed a Refinement network which fused the multiple layers' output in ResNet-101 and retained rich visual information named RefineNet. Hengshuang et al. [8] introduced a method for preseving local and global spatial information for high-resolution prediction named pyramid scene parsing network (PSPNet). These strategies are successful and obtain good results. But context information is confined to repeated striding in Refinenet. PSPNet lacks in extracting intermediate layer information from lower blocks.

*Corresponding author: Yu Qiao, qiaoyu@sjtu.edu.cn. This research is partly supported by NSFC, China (No: 61375048, 61572315, 6151101179) and 973 Plan, China (No: 2015CB856004).

Our network is designed based on the above considerations of the issue in semantic segmentation tasks. We propose a new semantic segmentation network based on ResNet and FCN with dilated residual unit and fusion of multi-path refinement and pyramid pooling.

2. PROPOSED METHOD

We propose a new segmentation framework based on ResNet-101 with new dilated residual unit illustrated in Fig. 1. Multiple resolution of feature maps from intermediate layers are combined to refine the output precision. Then a cascaded architecture is employed to the output of different resolutions. Two main modules are included in the cascaded architecture, residual unit without batch-normalization and chained residual pooling. Besides, interpolation and fusion layers are plugged in the architecture. The feature map from the last dilated residual unit is used as the input of pyramid pooling. Outputs from multi-path are then upsampled and fused with output of pyramid pooling. Finally, fused output is upsampled by a factor of 4 to achieve the original size dense output. Final output is then connected to a fully-connected CRF to refine boundary information.

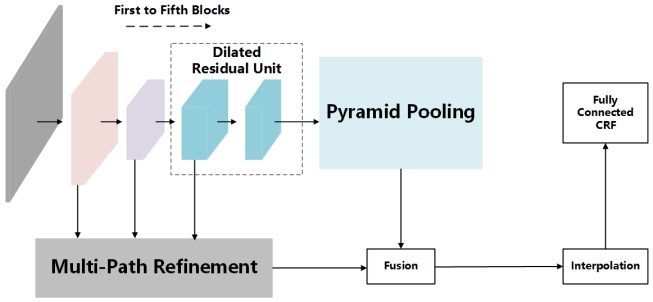

Fig. 1. Network Architecture

2.1. BN-ReLU Dilated Residual Unit

ResNet [3] could efficiently converge information to extract deep architectures with compelling accuracy. Following expression denotes the ordinary residual:

$$y_l = h(x_l) + F(x_l; W_l), x_{l+1} = f(y_l) \tag{1}$$

x_l denotes the input of lth unit, y_l denotes the output. W_l is the weights related to the unit. The h function in identity mapping [9]: $h(x_l) = x_l$ achieves the lowest training error and fastest optimization. Following the identity map rule, the output of $l+1$ layer is denoted as $x_{l+1} \equiv y_l$. Recursively, we will get L layer's output:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{2}$$

It has been verified in classification network that if we put batch normalization and relu before the convolution unit, we can get better result through pre-activation, which could make optimization more efficiently and raise regularization in case of overfitting. So we replace original residual unit to BN-ReLU residual unit as in Fig. 2.

However, when using regular convolution operations to extract features, it would diminish view of field, which could not entirely rebuilt by upsampling or interpolation. Therefore, we employ dilated convolution in 3×3 stage of residual unit. A dilated convolution can be defined as:

$$(F *_f k)(\boldsymbol{p}) = \sum_{s+f\boldsymbol{t}=\boldsymbol{p}} F(s)k(\boldsymbol{t}) \tag{3}$$

f denotes the dilated factor. Dilated rate are set as 2 and 4 respectively in 4th and 5th block.

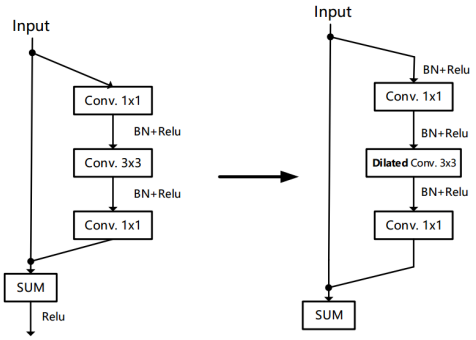

Fig. 2. BN-ReLU Dilated ReLU Residual Unit

2.2. Multi-path Refinement

Output feature map in 5th block of Dilated-ResNet is reduced by a factor of 8 due to the repeated 2-step striding in previous layers. Simply upsampling the feature map by a factor of 8 will result in heavy loss in region boundary information. However, stacked of deconvolution layers will result in increasing parameters and training time. As a result, multi-path refinement with intermediate layers is deployed in our network.

Inspired by Refinenet [7], we adopt the relu-conv residual unit without batch normalization and chained conv-pooling layers in multi-path module. Relu-conv residual unit is illustrated in Fig. 3. This kind of residual unit is used to fine tune the weights for segmentation tasks. Two sequential of relu-conv residual units are connected to the output of intermediate layer. Number of feature maps is a quarter of input feature maps. Output of this block is then fused with upsampled low resolution output from higher layer.

The fused output is then passed through the chained conv-pooling residual unit and three continuous relu-conv residual unit. Chained conv-pooling residual unit consists of 3 convolution and pooling modules as demonstrated in Fig. 4. The

difference of chained residual unit is that the second and the third block takes the output from previous layer. The module can efficiently extract feature with convolution and pooling in multiple region.

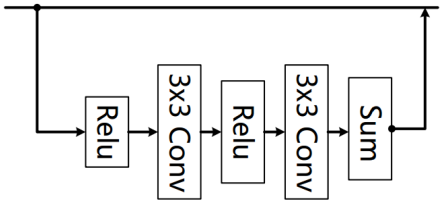

Fig. 3. ReLU-Conv Residual Unit

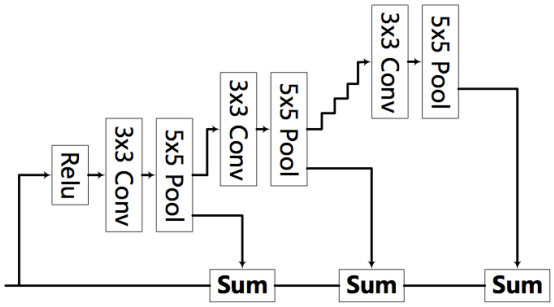

Fig. 4. Chained Conv-Pooling Residual Unit

2.3. Pymraid Pooling Module

Global average pooling was first used in classification CNNs between feature maps and softmax layer. It can extract relativance between feature and categories effectively. Wei et al. [10] introduced ParseNet which added global average pooling in FCN. Hengshuang et al. [8] employed the pyramid pooling module and achieved the best results. Global average pooling can be implemented to extract pooled context features from any layer. The final block in our network consists of higher semantic and global context information. So a pyramid of global average pooling connected to the final output can provide context information of multiple sub-regions. A set of average pooling is key to obtain robust high level semantic information.

Multi-path refinement takes the output from 2nd, 3rd and 4th blocks. Final output from 5th block is then passed through the pyramid pooling module. Different levels of average pooling is cascaded to the output which is 1/8 of original resolution. Using 4 different sizes of pooling kernels allows us to cover different part in output feature map. Pooled feature is then delivered to a convolution layer with a reducing number of feature maps, followed by bi-linear interpolation and concatenation layer.

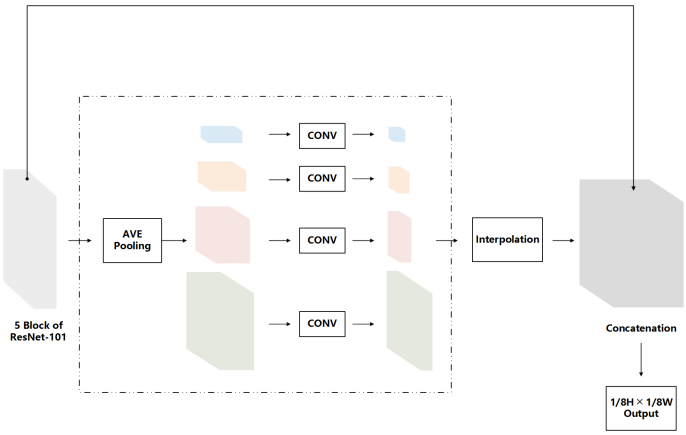

Fig. 5. Pyramid Pooling

2.4. Fusion of Multi-Path Refinement and Pyrmaid Pooling Dilated ResNet-101

We then fuse the result feature maps from multi-path refinement and pymraid pooling, yielding the final dense output. On the one hand, multi-path process combine low level information in 2nd block with high level semantics in 3rd block and 4th block. On the other hand, pyramid pooling can provide more context information for high resolution results. Relu-conv unit and chained conv-pooling residual unit follow the identity map to ensure efficient backpropagation training in network. Backward propagation properties of ResNet in identity map is of better performance, following the chain rule, we will get:

$$\frac{\partial l}{\partial x_l} = \frac{\partial l}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial l}{\partial x_L}(1 + \frac{\partial}{\partial x_l}\sum_{i=1}^{L-1} F(x_i, W_i)) \quad (4)$$

$loss$ denotes the loss function. Accordingly, we can find that:

$$\frac{\partial loss}{\partial W_l} = \frac{\partial loss}{\partial x_l}\frac{\partial x_l}{\partial W_l} = \frac{\partial x_l}{\partial W_l}(\frac{\partial loss}{\partial x_L} + \frac{\partial loss}{\partial x_L}\sum_{i=l}^{L-1}\frac{\partial F(x_i; W_{i+1})}{\partial x_l}) \quad (5)$$

Weights in lth layer only depends on $\frac{\partial l}{\partial x_L}$ and $\frac{\partial l}{\partial x_L}\sum_{i=l}^{L-1}\frac{\partial F(x_i;W_{i+1})}{\partial x_l}$. Identity mapping ensures the gradients flowing from Lth layer to the shallow layers. This ensures that our architecture can be trained end-to-end.

2.5. Boundary Refinement with Fully-connected CRF

We then refine the output from fusion of multi-path refinement and pyramid pooling with fully-connected CRF [11]. CRF is combined with our deep architecture as a post-processing step. In this way, we harness the boundary of dense output with multiple layers information and fully connected CRF.

Fully-connected CRF was first introduced in Deeplab which can eliminate shortage of previous CRFs. Energy function of fully-connected CRF can be denoted as:

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \quad (6)$$

$\boldsymbol{x}$ denotes the labels of each pixel. Unary potential is denoted as $\theta_i(x_i) = -logP(x_i)$. $P(x)$ in above expression is the dense output at pixel i computed by fused deep network.

3. EXPERIMENTS

Our proposed method achieve better results in semantic segmentation more efficently. We evaluate our proposed method on recently released dataset Cityscapes [12]. The cityscapes benchmark consists of 5000 high resolution annotated urban scene image recored from 50 cities, including 2975 training images, 500 validation images, 1525 test images. Our method is implemented by the Caffe framework [13] and Deeplab-v2 [6] on NVIDIA TITAN X GPU.

3.1. Implementation Details

Motivated by [6], we choose "poly" learning rate as our training policy. Base learning rate is set as 0.001 which is used in original ResNet training procedure and power is set as 0.9. Adding up iteration can lead to more accurate results but will introduce increased training time and iteration is set as 20K. We set momentum to 0.9 and 0.0001 for weight decay.

Before training, data augmentation is utilized to improve segmentation performance. To avoid introducing extra GPU memory allocated, we randomly pick a method from mirror, scaling, rotation and aspect ratio setting. The transformed width and height of aspect ratio setting is: $\omega_f = \omega/f, h_f = hf$, f denotes the factor. We randomly select float between 0.5 and 2 as mirror and scaling factor, rotation with -10 and 10 degrees, and aspect ratios from 0.7 to 1.3.

CRF parameters are finetuned following the rules in [6]. ω_2 and σ_γ in function (7) are both set as 3. Cross validation is used for validation set to find the best fit parameters. The range of initialization is 3 to 6, 30 to 100, 3 to 6 for ω_1, σ_α and σ_β respectively. In CRF layer, 10 mean field iterations is implemented. CRF layer is added in TEST phase.

Due to the limitation of amounts of GPUs, the convolution layers in our network is initialized and fixed by the PSP-Net [8] caffemodels. Learning rate in those layers is set as 0. At the same time, mean and variance parameters in batch normalization layers are not updated by the batch instead a pre-set parameter as illustrated in [7].

3.2. Results on Cityscapes

We evaluate our network on Cityscapes and compare the result to the advanced semantic segmentation architectures, CRFasRNN [14], FCN [2], Dilation [5], LRR [15] and Deeplab [6]. Segmentation results are measured with usual evaluation metrics: pixel accuracy, mean accuracy for each class and intersection over union (IoU). Table. 1 shows IoU of our method compared to other methods and per-class IoU results comparison.

	[14]	[2]	[5]	[15]	[6]	Ours
road	96.3	97.4	97.6	97.7	**97.9**	**97.9**
swalk	73.9	78.4	79.2	79.9	81.3	**81.4**
build.	88.2	89.2	89.9	**90.7**	90.3	90.1
wall	47.6	34.9	37.3	44.4	**48.8**	48.7
fence	41.3	44.2	47.6	48.6	47.4	**59.0**
pole	35.2	47.4	53.2	58.6	49.6	**59.6**
tlight	49.5	60.1	58.6	**68.2**	57.9	**68.2**
sign	59.7	65.0	65.2	72.0	67.3	**75.2**
veg.	90.6	91.4	91.8	**92.5**	91.9	92.3
terrain	66.1	69.3	69.4	69.3	**69.4**	63.8
sky	93.5	93.9	93.7	**94.7**	94.2	86.5
person	70.4	77.1	78.9	81.6	79.8	**82.7**
rider	34.7	51.4	55.0	60.0	59.8	**63.5**
car	90.1	92.6	93.3	94.0	93.7	**95.2**
truck	39.2	35.3	45.5	43.6	56.5	**66.1**
bus	57.5	48.6	53.4	56.8	67.5	**83.7**
train	55.4	46.5	47.7	47.2	57.5	**67.8**
mbike	43.9	51.6	52.2	54.8	57.7	**65.0**
bike	54.6	66.8	66.0	69.7	68.8	**71.7**
Mean IoU	62.5	65.3	67.1	69.7	70.4	**74.7**

Table 1. Per-class and Mean IoU results on Cityscapes

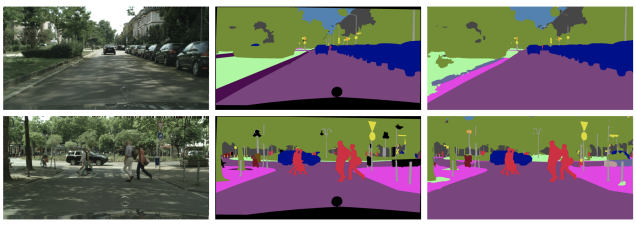

Fig. 6. Segmentation Result

4. CONCLUSION

In this paper, we propose a novel semantic segmentation architecture based on FCN and ResNet-101 which incorporate techniques raised for semantic segmentation tasks: dilated convolution, multi-path refinement and pyramid pooling. We further propose a novel network which fuses outputs from multi-path and pyramid pooling modules. The fusion output consists of low-level features and spatial context information. We also connect the output to a fully-connected CRF. Finally, we achieve good result of 74.7% mean IoU on Cityscapes.

5. REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[2] 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.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

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

[6] 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," *arXiv preprint arXiv:1606.00915*, 2016.

[7] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation," *arXiv preprint arXiv:1611.06612*, 2016.

[8] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[10] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[11] Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Adv. Neural Inf. Process. Syst*, vol. 2, no. 3, pp. 4, 2011.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[14] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[15] Golnaz Ghiasi and Charless C Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 519–534.