MULTIPLE SKIP CONNECTIONS OF DILATED CONVOLUTION NETWORK FOR SEMANTIC SEGMENTATION

Takayoshi Yamashita, Hironori Furukawa, Hironobu Fujiyoshi

Chubu University
1200, Matsumoto-cho, Kasugai, AICHI, Japan

ABSTRACT

Semantic segmentation is a task to estimate class for each pixel. This task also have received benefit from the deep ConvNet and it has achieves high accuracy. In the semantic segmentation from in-vehicle camera image, the object size such as a pedestrian or a vehicle fluctuates according to the distance from the camera. We propose scale aware semantic segmentation method especially small object. The contributions of the method are 1) to feed the features of small region by multiple skip connections, 2) to extract context from multiple receptive field by multiple dilated convolution blocks. The proposed method has achieved high accuracy in the Cityscapes dataset. The comparison with state-of-the-art methods, it has achieved the comparable performance at category IoU and iIoU metrics.

Index Terms— deep learning, convolutional neural network, dilated convolution, semantic segmentation

1. INTRODUCTION

. The semantic segmentation is a task of recognizing classes on a pixel-by-pixel. In earlier work, a bottom-up method has been proposed that clustering is performed by using hand crafted features such as color or gradient histogram and the regions that have similar feature are concatenated [3]. As advanced methods, Fully Convolutional Neural Network (FCN) [9] and encoder - decoder architecture such as Segnet [11] have improved the performance of semantic segmentation task. The various encoder - decoder based methods are proposed [12][19] [21] [22]. Semantic segmentation can be applied to road region extraction and pedestrian or vehicle detection for autonomous driving system. The self-driving robot system is also one of the applications of semantic segmentation. One of the critical problems is variation of object size. Even object of the same class, it has different sizes and appearances depending on the position from camera. To address this problem, we propose multiple dilated convolution blocks to deal with various object sizes. Dilated convolution convolves elements separated by a certain stride in convolution process. We also propose multiple skip connections in order to obtain robustness against appearance changes. This multiple skip connections is inspired from Residual Network [17] which achieves superior performance than human in object recognition task. U-Net[16] which connects the correspond layers from encoder to decoder is also related of our skip connections. Our multiple skip connections consists of 1) skip connection in encoder like ResNet, 2)skip and concatenate output of multiple dilated convolution blocks, and 3)skip connection like U-Net. The proposed method which employs multiple dilated convolution blocks and multiple skip connections is enable to extract fine segmentation result to small objects.

2. RELATED WORKS

Semantic segmentation learns network that recognizes class labels for each pixel by end-to-end approach. The network of this task is based on these object recognition network structures. Fully Convolutional Network (FCN) employ pre-trained VGG16 which is learned using Imagenet and fine-tune the network[9]. In order to accommodate different input data sizes, the fully connected layers are replaced with convolution layers which have 1×1 filter. It is adopted skip connections to capture global information of the entire image and local information of each class. Because the layer which closes the input layer of the network captures fine information of the object. On the other hand, by repeating the pooling, the feature maps become small and lack fine information. The skip connection concatenates output feature maps and intermediate feature maps to obtain fine class label. Segnet consists of encoder - decoder network structure[11]. The encoder has convolution and pooling layers that is based on VGG16 structure and extracts features from the input image. The decoder has a structure paired with the encoder that consists of deconvolution and upsampling layers. In addition, the pooling indices are recorded, and the feature value is substituted to recorded position at upsampling layer, and 0 is substituted at the other positions. Thus, fine class labels can be recognized by decoding features.

3. PROPOSED METHOD

We propose a network that can segment the details of small objects. The proposed network structure is shown in Figure

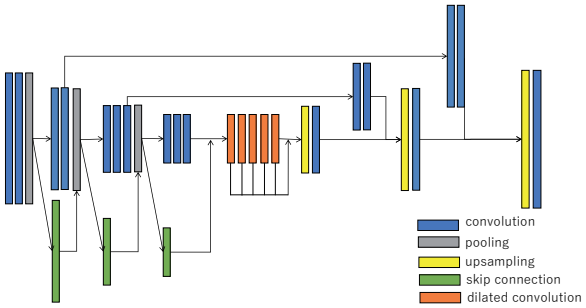

Fig. 1. Network architecture of our semantic segmentation. Blue box is convolution layer, gray box is pooling, yellow box is upsampling, orange box is Dilated Convolution and green box is skip connection. The network is based on encoder - decoder structure. The multiple Dilated convolution layers are arranged between encoder and decoder. Skip connection are introduced at them.

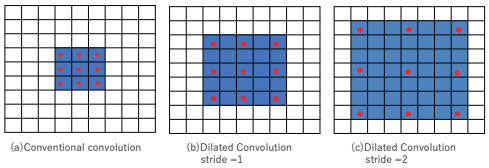

Fig. 2. Conventional Convolution and Dilated Convolution. While former convolution applied filter densely at elements, later convolution applied sparsely with stride s.

1. The basis network is an encoder - decoder structure. Multiple skip connections are introduced at encoder, decoder and between them to keep local information. In order to perceive global context, the multiple dilated convolution blocks is arranged between the encoder and the decoder . The feature maps are bypassed and entered with the feature maps of following layer. This multiple dilated convolution blocks has also skip connection that merges feature maps of them. In order to propagate the fine information of the small object to the decoder, the feature maps on the encoder are connected to the pair on the decoder layers. Each skip connection performs convolution of 1×1. The multiple dilated convolution blocks consists of multiple dilated convolution layers. In addition, each convolution layer performs batch normalization [14]. Batch normalization reduces variations in data between batches in mini batch learning, It accelerates the convergence of learning and becomes robust to variations such as brightness.

3.1. Basis network

As shown in Fig. 1, the encoder consists of 10 layers of convolution layers and the decoder consists of 3 deconvolution layers. Filter size and number of filters for each layer are shown in the Table 1. The filter size in each layer is 3×3, and the number of filters is doubled after 2×2 max pooling. The

Table 1. Structures of each layer.

Layer	Filter size	# of filters	pooling
1	3×3	32	–
2	3×3	32	max pooling
3	3×3	64	–
4	3×3	64	max pooling
5	3×3	128	–
6	3×3	128	–
7	3×3	128	max pooling
8	3×3	256	–
9	3×3	256	–
10	3×3	256	–
11	3×3, s=2	512	–
12	3×3, s=4	512	–
13	3×3, s=8	512	–
14	3×3, s=16	512	–
15	3×3, s=32	512	–
16	3×3	256	upsampling
17	3×3	128	upsampling
18	3×3	# of classes	upsampling

decoder performs deconvolution for the number of times of pooling. For each deconvolution, feature maps are upsampled by a factor of 2 and convolution is performed. The filter size of decoder is 3×3. Segnet has the same number of convolution layers as the encoding side at each deconvolution. Unlike structure of Segnet, our network has one convolution layer at deconvolution. ReLU is applied as the activation function of each layer on both the encoder and decoder.

3.2. Multiple dilated convolution blocks

Multiple dilated convolution blocks that capture the global context are arranged between the encoder and the decoder. Dilated convolution is a convolution process that separates elements to be convoluted with stride, as shown in Figure 2. When 3×3 filter convolves to input map, conventional convolution is performed for densely 3×3 region as shown in Fig. 2 (a). The input value and the filter value are multiplied at each element to obtain the correspond value. On the other hand, dilated convolution has stride which is distance between convolved elements. When it is set to 1 as shown in Fig. 2(b), 3×3 filter is applied to 5×5 region. When stride is set to 2, the convolution is performed on 7×7 region as shown in Fig. 2(c). The dilated convolution is a sparse connection to a wider region than conventional convolution. We stacks 5 dilated convolution layers with different strides. Although the dilated Convolution has the same filter size as the conventional convolution, it enable to perceive wider range to capture global context by stacking them.

3.3. Multiple skip connections

In the ResNet, the error can be propagated close to the input layer even in ultla deep network structure by introducing skip

Table 2. Comparison result of network structure.

Multiple Dilation Convolution Blocks	Multiple Skip Connections	class[%] IoU	class[%] iIOU	catgory[%] IoU	catgory[%] iIoU
none	none	54.9	37.8	83.6	73.2
use	use	56.1	40.2	84.3	76.1
use no skip	none	67.3	45.8	87.8	74.1
use no skip	use	72.5	52.5	89.2	78.2
use with skip	use	73.0	55.6	89.2	81.9

connection. In FCN and U-Net, high-resolution segmentation results are obtained using the feature maps of the intermediate layers. This idea can also be regarded as a kind of skip connection. In our network, skip connection like the ResNet is introduced to the encoder. The skip connection like the FCN that is connected between pair of encoder and decoder is also introduced in our network. In the skip connection on the encoder, the value of each element of the feature map is added. When the number of channels of the feature map is inequivalent, the convolution with 1×1 is performed as much as the number of channels of the upper layer to adjusts the number of channels. In the case of adding the 32 channel feature map in the 2nd layer to the 64 channel feature map in the 4th layer 64 of 1×1 filters are applied to channel feature maps in the 2nd layer to obtain a feature map of 64 channels. In skip connection of FCN, the feature map of the encoder is added to the feature map of the decoder for each element. While skip connection like ResNet concatenates feature maps, it is known that concatenation and addition are equivalent [20]. By performing skip connection by adding, it is possible to suppress the amount of memory usage without increasing the number of feature maps. Through these skip connections, fine information on the small object can be propagated through two paths. We also introduce skip connections that merge the feature maps of each dilated convolution layers. It is possible to input feature maps that captures information in various respective regions.

3.4. Learning of our network

The proposed network is enable to learn end-to-end approach. Unlike conventional ConvNet based semantic segmentation methods, it does not use pre-trained network. This makes it possible to flexibly change the network structure. The mini-batch size is 16. Adam [7] is used for learning optimization method. Cropped region with a fixed size from image randomly is input during learning process. It is possible to consider various scenes and augment variation of the learning data. The input size is 720×720 which is cropped from 0.75 times to 1.25 times of the size and is resized. Our method is implemented by chainer framework and learning with NVIDIA DGX-1. Because the memory size of the Tesla P100 in the DGX-1 is 16GB, the mini batch size that can be processed with one GPU is 2. Therefore, mini batch learning is performed in data parallel by using 8 GPUs.

Table 3. Comparison result on test dataset of Cityscapes

method	class[%] IoU	class[%] iIoU	category[%] IoU	category[%] iIoU
SegModel	78.5	56.1	89.8	75.9
ResNet-38[21]	78.4	59.1	90.9	81.1
Dilation10 [15]	67.1	42.0	86.5	71.1
FCN-8s[9]	65.3	41.7	85.7	70.1
Segnet basic[11]	57.0	32.0	79.1	61.9
proposed method	71.6	49.4	89.3	78.3

4. EXPERIMENTS

We evaluate the proposed network in Cityscapes dataset [18]. This dataset is taken in 50 cities in Europe during a day in fine weather and the number of classes is 30 classes. Some classes frequently occur frequently, therefore, 19 classes are used for evaluation. The Annotation includes fine annotations and coarse annotations. While fine annotations are annotated in detail for 5000 images, the Coarse annotations are rough annotations that surround the area for 20000 images. In this experiment, we use fine annotations data. It includes 2975 images for learning, 500 images for validation, and 1525 images for testing. The annotation data for the image for testing is not published. Testing results can be obtained by upload the result. Therefore, in this experiment, comparison experiments are performed using validation dataset.

4.1. Evaluation metrics

Evaluation of semantic segmentation coincides with class annotated pixel by pixel. It judges whether or not to calculate the average IoU (Intersection over Union) [4]. In Cityscapes, 19 classes to be evaluated are classified into 7 categories, and the average category IoU is also evaluated. On the other hand, the IoU has a problem that accuracy with respect to a small class of areas tends to deteriorate. There are many classes whose size varies depending on the distance from the camera, such as pedestrians, cars, signs, etc., in the image of the target in-vehicle camera. Therefore, Cityscapes has criteria to calculate the IoU at the instance level and evaluate the average. IoU at the instance level is obtained as iIoU. As with IoU, iIoU evaluates by calculating the average of each class and category.

4.2. Comparison of network structures

In the proposed method, multiple dilated convolution blocks and multiple skip connections are introduced in the encoder - decoder structure. We compare the accuracy of these structures. The evaluation result is shown in the Table 2. By intro-

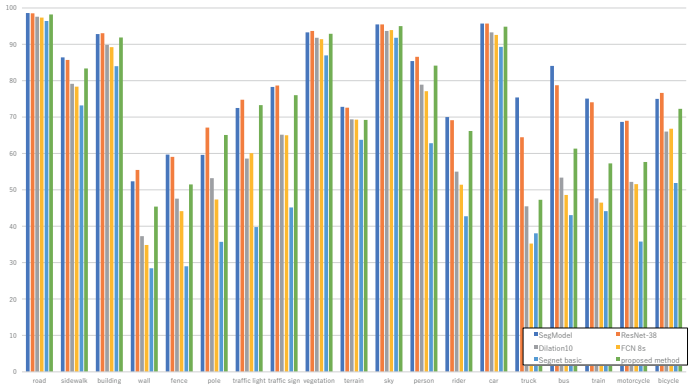

Fig. 3. IoU of each class on test dataset of Cityscapes. We compared with 6 methods that are common semantic segmentation methods and the state of the art methods.

ducing skip connections in encoder and between encoder and decoder, accuracy is improved from plain encoder-decoder structure from 2% to 3%. By introducing multiple dilated convolution blocks without skip connection, the average class IoU increases from 54.9% to 67.3% and the average class iIoU has improved significantly from 37.8% to 45.8%. Also, the category IoU and the category iIoU are improved to about 5% and the accuracy is improved. When both multiple skip connections and multiple dilated convolution blocks are introduced, the accuracy is further improved, the average class IoU is 72.5%, the average class i IoU is 52.5%, the average category IoU is 89.2% and the average category i IoU is 78.2%. These two processes greatly contribute to improvement in accuracy. When skip connection is also introduced for multiple dilated convolution blocks, the iIoU precision of classes and categories has improved by about 3%. The multiple dilated convolution blocks capture the context of wide region.

4.3. Comparison on test dataset

The comparison result on the Cityscapes test dataset is shown in Table 3. As a result, compared with common segmentation methods such as Segnet and FCN, the accuracy can be improved significantly in each evaluation metrics. In addition, compared with the method using dilated convolution, the proposed method can obtain better results. When compared with the state of the art methods (SegModel, ResNet - 38) recorded in the benchmark result of Cityscapes, these methods are better on class IoU and class iIoU. Meanwhile, our method achieves equivalent accuracy on the category IoU and the category iIoU. Since the proposed method can classify at the category level, the multiple dilated convolution blocks and multiple skip connections have the effect of improving the accuracy of semantic segmentation. The IoU for each class of these comparison methods is shown in Figure. 3. Our method achieves high accuracy for classes with high occurrence frequency such as road, building, sky etc., and it is similar to comparison methods. It also obtain equivalent accuracy for

Fig. 4. Result images on test dataset of Cityscapes. First and third columns are input images and second and forth columns are our results.

classes with large size fluctuation such as person and car. It is highly accurate for classes such as road, person, and car, which are important for autonomous driving system.

Figure. 4 shows semantic segmentation results for test dataset. It obtains fine segmentation results for specific classes such as road, building, person, car. In addition, it can handle small regions and various scenes. On the other hand, although bus and truck are roughly segmented, the region in the object is segmented as another class. Therefore, these Class IoU precision is deteriorating. On the other hand, although it obtain roughly segmentation for bus and truck classes, the area in the object is segmented as another class.

5. CONCLUSION

We propose semantic segmentation method with multiple skip connections and multiple dilated convolution blocks. The skip connections are consists, skip connection in encoder like ResNet and skip connection in paired between encoder and decoder like FCN, and merged feature maps of dilated convolution layers. We achieved segmentation with higher performance than FCN and Segnet that are common segmentation methods on the Cityscapes dataset. Moreover, compared to the state of the art methods, our network achieves equivalent accuracy on average category IoU and average category.

6. REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition" , Proceedings of the IEEE, 1998.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems (NIPS2012), 2012.

[3] C. Farabet, C. Couprie, L. Najman, Y. LeCun, "Learning hierarchical features for scene labeling", IEEE transactions on pattern analysis and machine intelligence (PAMI), 2013.

[4] M. Everingham, A. S. M. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes challenge: A retrospective", International Journal of Computer Vision (IJCV), 2014.

[5] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", IEEE conference on computer vision and pattern recognition (CVPR2014), 2014.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2014), 2014.

[7] D. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

[8] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", International Conference on Learning Representation (ICLR2015), 2015.

[9] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), 2015.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du,P. H. Torr, "Conditional random fields as recurrent neural networks", IEEE International Conference on Computer Vision (CVPR2015), 2015.

[11] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", arXiv preprint arXiv:1511.00561, 2015.

[12] A. Kendall, V. Badrinarayanan, R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding", arXiv preprint arXiv:1511.02680, 2015.

[13] H. Noh, S. Hong, B.Han, "Learning deconvolution network for semantic segmentation", IEEE International Conference on Computer Vision (ICCV2015), 2015.

[14] S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", arXiv preprint arXiv:1502.03167, 2015.

[15] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions", International Conference on Learning Representation (ICLR2016), 2016.

[16] P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[17] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele, "The cityscapes dataset for semantic urban scene understanding", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[19] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", arXiv preprint arXiv:1606.00915, 2016.

[20] P. O. Pinheiro, T. Y. Lin, R. Collobert, P. Dollar, "Learning to refine object segments", European Conference on Computer Vision (ECCV), 2016.

[21] Z. Wu, C. Shen, A. van den Hengel, "Wider or Deeper: Revisiting the ResNet Model for Visual Recognition", arXiv preprint arXiv:1611.10080, 2016.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid Scene Parsing Network", arXiv preprint arXiv:1612.01105, 2016.