Review: Multi-Scale Context Aggregation by Dilated Convolutions

Jiho Bak
School of Electrical and Electronic Engineering, Yonsei University

Abstract

The previous state-of-the-art semantic segmentation models were based on the state-of-the-art classifiers. Yu [1] claims that the models specifically designed for dense prediction, will perform better for dense prediction. Therefore, Yu [1] proposes 'dilated convolution' for multi-scale context aggregation with larger receptive field, but without loss of resolution. The 'context module' and 'front end module' are proposed in the paper, which improves the performance of previous state-of-the-art segmentation models with end-to-end training.

1 Dilated Convolution

The concept of 'convolution with a dilated filter' was used in wavelet decomposition signal processing [2]. The basic idea is 'dilating' the position of the multiplying computation in convolution. The original convolution can be represented as in 1, where F is discrete function, k is discrete filter of size $(2r+1)^2$, and $*$ is discrete convolution operator. Then dilated convolution can be represented as in 2 with dilated factor l and dilated convolution operator $*_l$. This makes the original convolution as 1-dilated convolution.

$$(F * K)(p) = \sum_{s+t=p} F(s)k(t) \tag{1}$$

$$(F *_l K)(p) = \sum_{s+lt=p} F(s)k(t) \tag{2}$$

The dilated convolution remains the number of the parameter while expanding receptive field. Fig 1 is a visualization of receptive field where (a), (b), and (c) are 1, 2, 3 dilated examples each. As a result, receptive field of each element in F_{i+1}(i-dilated) is $(2^{i+2}-1)\times(2^{i+2}-1)$.

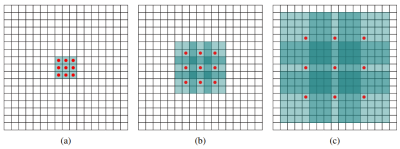

Figure 1: The receptive fields of dilated convolution

2 Context Module

2.1 Architecture

The context module is the actual architecture that aggregates multi-scale contextual information. This module consists of dilated convolution layers and the truncation function $\max(\cdot,0)$, which is ReLU. The context module uses 1, 1, 2, 4, 8, 16, 1 dilated 3×3 convolution layers and one last 1×1 convolution layer. The Basic context module uses same number of channel as input for every layers. The Large context module uses deeper channels. The architecture of the module are organized in Fig 2.

Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C

Figure 2: The Basic and Large context module structure.

2.2 Initialization

The basic initialization for dilated convolution network is identity initialization. Through experiments, not only the intuition, Yu claims that identity initialization helps the learning for the segmentation tasks. For large context module, since it uses deeper channels, the initialization weights are normalized as in Fig 3.

$$k^b(\mathbf{t},a) = \begin{cases} \frac{C}{c_{i+1}} & \mathbf{t}=0 \text{ and } \left\lfloor \frac{aC}{c_i} \right\rfloor = \left\lfloor \frac{bC}{c_{i+1}} \right\rfloor \\ \varepsilon & \text{otherwise} \end{cases}$$

Figure 3: The initialization equation for dilated convolution filter.

3 Front End Module

The front-end prediction module were designed based on VGG-16. For dense prediction, the last two pooling and striding layers were removed, and the subsequent convolution layers were dilated by factor of 2 and 4 each. This enables the initialization with the parameters of the original classification network, but produces higher-resolution output[1]. Finally, the front-end module takes padded images and produces 64×64 resolution feature map.

4 Experiments

4.1 Evaluation of front end module

The front end module were evaluated by Pascal VOC dataset and compared with FCN-8s, DeepLab, DeepLab-Msc. Front end module outperformed these models.

4.2 Evaluation of Context Aggregation

The context modules plugged after the front end and before other networks, were train and tested in VOC and Microsoft COCO datasets. As in Fig 4, the models with larger context module and with additional model, performed best. This result shows the effect of adding the context module to three different architectures for semantic segmentation.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean IoU
Front end	86.3	38.2	76.8	**66.8**	63.2	87.3	78.7	82	33.7	76.7	53.5	73.7	76	76.6	83	**51.9**	77.8	44	79.9	**66.3**	69.8
Front + Basic	86.4	37.6	78.5	66.3	64.1	89.9	79.9	84.9	**36.1**	79.4	**55.8**	77.6	81.6	79	83.1	51.2	81.3	43.7	82.3	65.7	71.3
Front + Large	**87.3**	**39.2**	**80.3**	65.6	**66.4**	**90.2**	**82.6**	**85.8**	34.8	**81.9**	51.7	**79**	**84.1**	**80.9**	**83.2**	51.2	**83.2**	**44.7**	**83.4**	65.6	**72.1**
Front end + CRF	89.2	38.8	80	**69.8**	63.2	88.8	80	85.2	33.8	80.6	55.5	77.1	80.8	77.3	84.3	**53.1**	80.4	45	80.7	**67.9**	71.6
Front + Basic + CRF	89.1	38.7	81.4	67.4	65	91	81	86.7	**37.5**	81	**57**	79.6	83.6	79.9	**84.6**	52.7	83.3	44.3	82.6	67.2	72.7
Front + Large + CRF	**89.6**	**39.9**	**82.7**	66.7	**67.5**	**91.1**	**83.3**	**87.4**	36	**83.3**	52.5	**80.7**	**85.7**	**81.8**	84.4	52.6	**84.4**	**45.3**	**83.7**	66.7	**73.3**
Front end + RNN	88.8	38.1	80.8	**69.1**	65.6	89.9	79.6	85.7	36.3	83.6	57.3	77.9	83.2	77	**84.6**	**54.7**	82.1	**46.9**	80.9	66.7	72.5
Front + Basic + RNN	89	38.4	82.3	67.9	65.2	91.5	80.4	87.2	**38.4**	82.1	**57.7**	79.9	85	79.6	84.5	53.5	84	45	82.8	66.2	73.1
Front + Large + RNN	**89.3**	**39.2**	**83.6**	67.2	**69**	**92.1**	**83.1**	**88**	**38.4**	**84.8**	55.3	**81.2**	**86.7**	**81.3**	84.3	53.6	**84.4**	45.8	**83.8**	**67**	**73.9**

Figure 4: Controlled evaluation of the effect of context module.

4.3 Final Evaluation on the Test Set

The final evaluation is done in Pascal VOC 2012 evaluation server. The context module boosted the accuracy over the front end. The context module alone, without subsequent models, outperformed CRF. And finally, the model with context module and CRF-RNN showed the state-of-the-art performance. As a result, we can assure that the context module boosts the segmentation accuracy, which is efficient design for dense prediction.

[1] Vladlen Koltun Fisher Yu. Multi-scale context aggregation by dilated convolutions. *ICLR*.

[2] Mark J Shensa. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE*.