Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi- Supervised Semantic Segmentation

(2018)

Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, Thomas S. Huang

Notes

1 Introduction

The big gap between the supervised semantic segmentation and weakly supervised segmentation, this is mainly due to the quality of the generated masks based on the image level labels, that are then used to train the segmentation network, the authors propose a method to use different dilation rates to enlarge the receptive feild and obtain more precise localization maps, that can then be merge for a more precise pseudo labels:

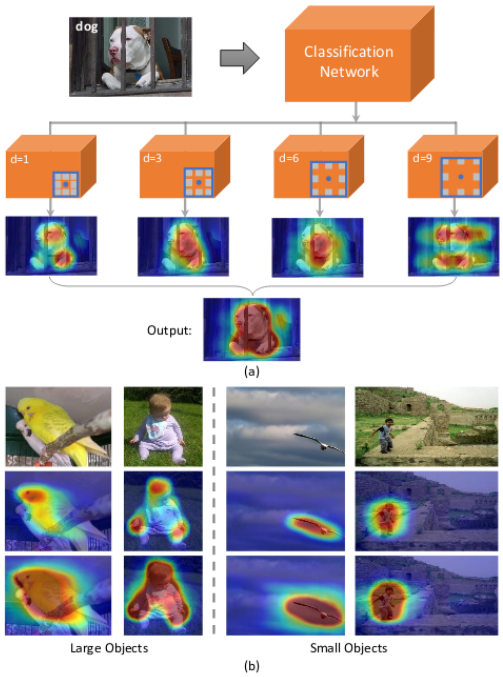

Figure 1. (a) Our proposed approach: equipping a standard classification network with multiple dilated convolutional blocks of different dilation rates for dense object localization. (b) Comparison between the state-of-the-art CAM [46] (the 2nd row) and ours (the last row) on quality of the produced object localization maps. Our approach localizes target objects more accurately even in presence of great scale variation.

The authors utilize CAM to generate an object localization map for each convolutional block with different dilation rates, the convolution block can only localize two small discriminative regions without enlarging dilation rate i.e. d = 1. By gradually increasing the dilated rates (from 3 to 9), more object-related regions are discovered.

2 Method

Instead of only detecting discriminative regions in the image, with different dilation rates we can enlarge the receptive field and enabling the information to transfer from discriminative regions to adjacent non-discriminative like we see in the figure bellow, By enlarging the dilated rate from 1 to 3 of a 3x3 kernel, the location near the *head* will be perceived and get their discriminativeness enhanced. By further increasing the dilated rates (to d= 6,9), some further locations will perceive the head and similarly facilitate the classification model to discover these regions.

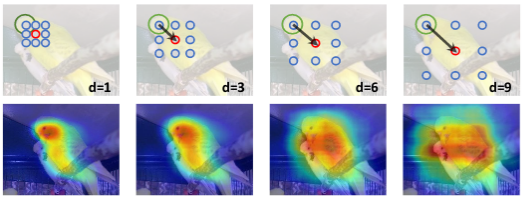

Figure 2. Motivation of our approach: information can be transferred from the initially discriminative region to other regions by varying dilated rates of convolutional kernels. The corresponding localization maps are shown in the 2nd row. Best viewed in color.

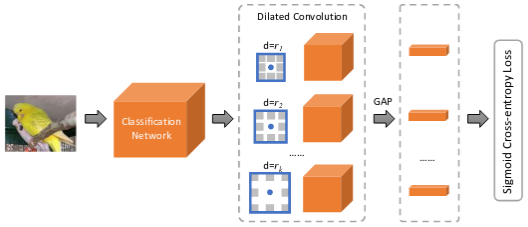

Figure 3. Illustration on training the network with multiple dilated convolutional blocks.

To apply different dilations, they use VGG16 and they remove the fully connected layers and one pooling layer to enlarge the resolution of feature maps, then convolutionnal blocks with different dilation rates are appended to *conv5* to localize object related region with different receptive fields, after then we apply global average pooling for each output of different dilation rates, and pass them though the fully connected layer for the classification scores to train the network to minimize the sigmoid cross entropy loss, we then create the CAM activation maps for class specific localization maps, thus for each dilation rates we'll have the localization maps of the target classes, with different localization map, and to reduce the false region, we take the average over the localization maps by different dilation rates that are greate then 1 (d = 3,6,9) and then this localization map is added to the localization map of the standard convolutionnal block (d=1) to not miss the accurate regions:

$$H = H_0 + \frac{1}{n_d}\sum_{i=1}^{n_d} H_i$$

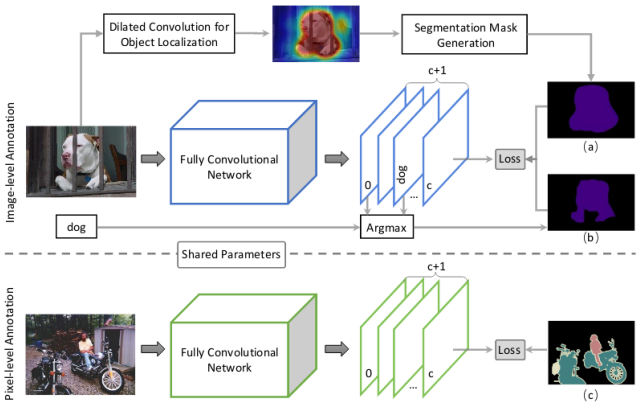

Figure 4. Details of training semantic segmentation in weakly- or semi- supervised manner with our proposed approach. In particular, (a) is the segmentation mask inferred from the dense localization map; (b) is the online predicted segmentation mask; (c) is the human annotated segmentation mask.

Weakly-supervised In a Weakly-supervised setting, to get the localization maps, we extract the confidence scores of the classification that corresponds to the ground truth and then use CAM to

extract the different localization maps, and then we calculate the loss using both the predicted segmentation $\hat{M}_w$ mask and the pseudo segmentation mask produced by the network M_w.

$$J_w\left(f\left(I_w;\theta\right)\right) = -\frac{1}{\sum_{c\in\mathcal{C}}\left|M_w^c\right|}\sum_{c\in\mathcal{C}}\sum_{u\in M_w^c}\log f_{u,c}\left(I_w;\theta\right) - \frac{1}{\sum_{c\in\mathcal{C}}\left|\hat{M}_w^c\right|}\sum_{c\in\mathcal{C}}\sum_{u\in\hat{M}_w^c}\log f_{u,c}\left(I_w;\theta\right)$$

Semi-supervised Learning With additionnal strong labels, we add a second loss function where we some the log of the predicted probabilities over the correct classes, which is simplt a cross entropy:

$$J_s\left(f\left(I_w;\theta\right)\right) = -\frac{1}{\sum_{c\in\mathcal{C}}\left|M_s^c\right|}\sum_{c\in\mathcal{C}}\sum_{u\in M_s^c}\log f_{u,c}\left(I_s;\theta\right)$$

3 Experiments

The authors use DeepLab-CRF-LargeFOV that used for both classification and segmentation, with a mini batch of 30 images that are randomly croped to a size of $321x321$ pixels, the model is trained for 15 epochs, with an initial learning rate of 0.001 and is decreased by a factor of 10 after 6 epochs.

Table 1. Comparison of weakly-supervised semantic segmentation methods on PASCAL VOC 2012 validation and test sets.

Methods	Training Set	validation	test
Supervision: Scribbles			
Scribblesup CVPR2016 [17]	10K	63.1	-
Supervision: Box			
WSSL ICCV2015 [21]	10K	60.6	62.2
BoxSup ICCV2015 [3]	10K	62.0	64.2
Supervision: Spot			
1 Point ECCV2016 [28]	10K	46.1	-
Scribblesup CVPR2016 [17]	10K	51.6	-
Supervision: Image-level Labels			
MIL-FCN ICLR2015 [23]	10K	25.7	24.9
CCNN ICCV2015 [22]	10K	35.3	35.6
EM-Adapt ICCV2015 [21]	10K	38.2	39.6
MIL-seg* CVPR2015 [24]	700K	42.0	40.6
SN_B* PR2016 [35]	10K	41.9	43.2
TransferNet* CVPR2016 [7]	70K	52.1	51.2
DCSM ECCV2016 [31]	10K	44.1	45.1
BFBP ECCV2016 [29]	10K	46.6	48.0
SEC ECCV2016 [14]	10K	50.7	51.7
AF-MCG* ECCV2016 [26]	10K	54.3	55.5
STC TPAMI2017 [34]	50K	49.8	51.2
Saleh *et al.*TPAMI2017 [30]	10K	50.9	52.6
Ray *et al.*CVPR2017 [27]	10K	52.8	53.7
AE-PSL CVPR2017 [33]	10K	55.0	55.7
Hong *et al.*CVPR2017 [8]	970K	58.1	58.7
Kim *et al.*ICCV2017 [13]	10K	53.1	53.8
MDC (Ours)	10K	**60.4**	**60.8**

(* indicates methods implicitly use pixel-level supervision)

Table 2. Comparison of semi-supervised semantic segmentation methods on PASCAL VOC 2012 validation and test sets.

Methods	validation	test
Weakly Supervision: Boxes		
BoxSup ICCV2015 [21]	63.5	66.2
WSSL ICCV2015 [21]	65.1	66.6
Khoreva *et al.*CVPR2017 [12]	65.8	66.9
Weakly Supervision: Image-level Labels		
WSSL ICCV2015 [21]	64.6	66.2
MDC (Ours)	**65.7**	**67.6**