Rethinking Atrous Convolution for Semantic Image Segmentation
(2017)

Liang-Chieh Chen et al.
Resume

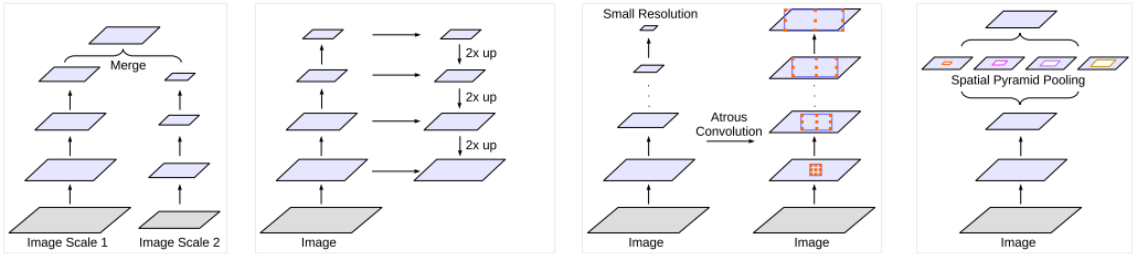
March 1, 2019

---

1 Introduction

In this version of deeplab, the authors propose the usage of cascaded trous convolutions (starting from the fourth block of resnet), and similary to PSPNet and adding image level features to encode global context, removing the need to use fully connected CRFs to improve the detections.

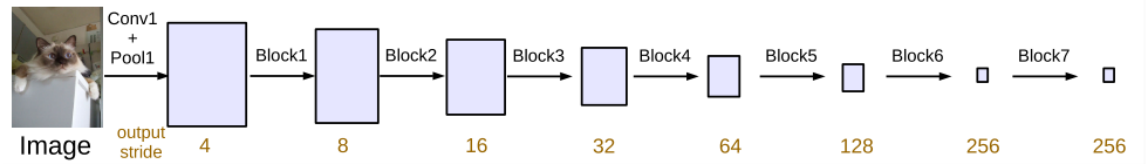
Architecures for capturing multi-scale context



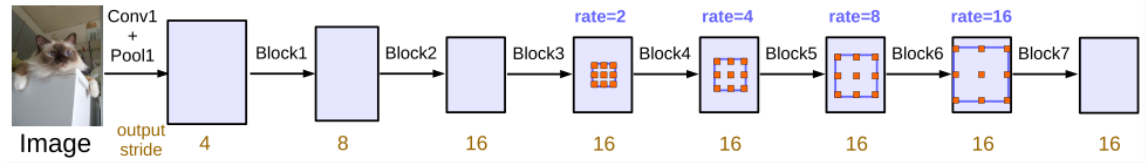
(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

2 Cascaded atrous convolutions

The authors explore the usage of cascaded convolutions after the fourth block of the resnet architecture, thus adding three block with different rates of dilation. In the fourth block of the resnet, the output stride (the factor by wchich the spatial dimensions of the input image are reduced) is 32, if we added three block with the same architecture the reduction will be 256.



(a) Going deeper without atrous convolution.



To avoid such a severe disimation of the signal, they replace the strided convolutions in each added block starting from block 4 with dilated convolution of different rates (2 - 4 - 8 -16)

Multi-Grid	block4	block5	block6	block7
(1, 1, 1)	68.39	73.21	75.34	75.76
(1, 2, 1)	70.23	75.67	76.09	**76.66**
(1, 2, 3)	73.14	75.78	75.96	76.11
(1, 2, 4)	73.45	75.74	75.85	76.02
(2, 2, 2)	71.45	74.30	74.70	74.62

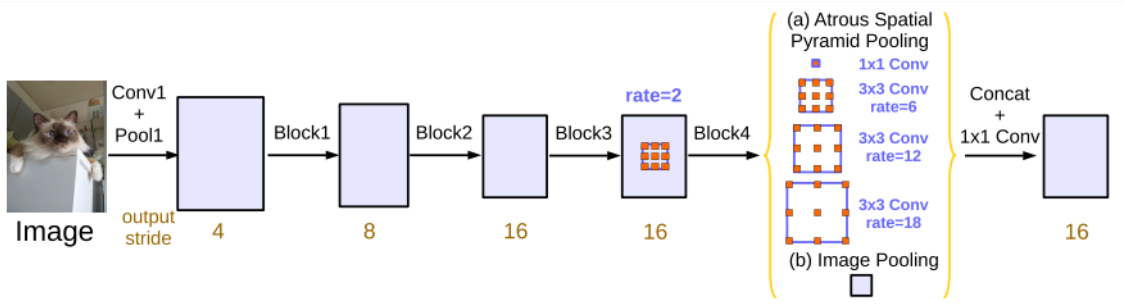
Table 3. Employing multi-grid method for ResNet-101 with different number of cascaded blocks at *output_stride* = 16. The best model performance is shown in bold.

Multi-grid: In each one of the blocks, we have 3 conv3x3, and so, for a rate of 2 (block 4), and $MultiGrid = r_1, r_2, r_3$, the dilation rate for each one of the three 3x3convs in block 4 will be : $2 \times r_1$, $2 \times r_2$ and $2 \times r_3$ respectively.

3 Atrous Spatial Pyramid Pooling

The problem with the ASSP proposed in deeplab v2, is the usage of vety big dilation rates, for 3x3 convs, when the dilation rate is small, all the 9 filter weights are applied to most of the valid region on feature map, when the atrous rate gets larger, the 3 3 filter degenerates to a 11 filter since only the center weight is effective.

So to incorporate the global context without the need of using large dilation rates, the autors use average pooling $- >$ conv1x1 of the incoming volume, the same approch as in PSPNet, and also adding batchnorm to every branch.



So the improved ASPP consists of (a) one 11 convolution and three 3 3 convolutions with rates = (6, 12, 18) when output stride = 16 (doubled when the output stride == 8) (all with 256 filters and batch normalization), and (b) the image-level features (average pooling), The resulting features from all the branches are then concatenated and pass through another 1 1 convolution (also with 256 filters and batch normalization) before the final 1 1 convolution which generates the final logits.

Multi-Grid			ASPP		Image	
(1, 1, 1)	(1, 2, 1)	(1, 2, 4)	(6, 12, 18)	(6, 12, 18, 24)	Pooling	mIOU
✓			✓			75.36
	✓		✓			75.93
		✓	✓			76.58
		✓		✓		76.46
		✓	✓		✓	77.21

Table 5. Atrous Spatial Pyramid Pooling with multi-grid method and image-level features at *output_stride* = 16.

4 Training

- Learning rate policy: $lr = lr \times \left(1 - \frac{iter}{max\_iter}\right)^{0.9}$

- Crop size: Large croping is important to avoid the problem of conv1x1 with higher dilation rates.

Crop Size	UL	BN	mIOU
513	✓	✓	77.21
513	✓		75.95
513		✓	76.01
321		✓	67.22

- Batch normalization: all the added modules on top of the resnet includes batch normalization
- Training: They begin by employing $output_stride = 16$ (starting from block 4), with a batch of 16, and $LR = 0.007$ for 30K iterations ($\tilde{4}8$ Epochs), then they freeze batch norm parameters, use $output_stride = 8$ (from block 3) and fine ture with a $LR = 0.001$.