## **Research paper-3: Rethinking Atrous Convolution for Semantic Image Segmentation (DeeplabV3)**

Source: <https://arxiv.org/pdf/1706.05587.pdf>

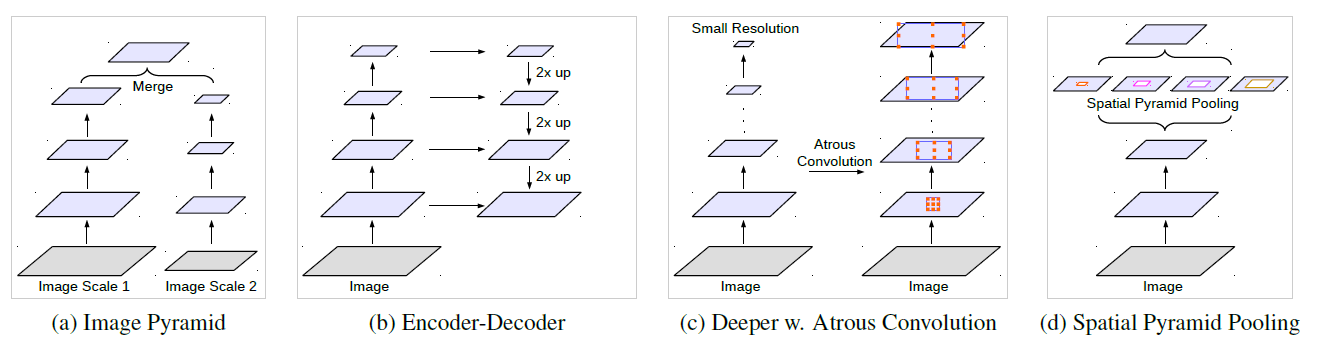
## **Summary**

Introduction:

* This paper is proposed by some of the top researches in google where they try to rethink and reinvent some of their previous work on deeplabV1 and deeplabV2 network which was mainly created for image segmentation.
* The architecture in the paper solves the problem that involves signal decimation and learning multi-scale contextual features in the neural network.
* The author in this paper discusses atrous convolution, which is a powerful tool to explicitly adjust filter’s field-of-view as well as control the resolution of feature responses computed by Deep Convolutional Neural Networks.
* This paper proposes a module that employs atrous convolution in cascade or in parallel to capture the multi-scale context by adopting multiple atrous rates in the network.
* Atrous convolution chooses to preserve the feature resolution and some of the spatial information in the Deep neural network.

Related Work:

* In recent years there has been much research works on Image segmentation like deeplabv1 and deeplabv2 which involves Preprocessing using Conditional random field (CRF), Pspnet with context information, etc where some work to some extent.



**Various Architectures from Related works**

* Most of the recent works include convolution networks with different architectures as below used for image segmentation.

1. Image Pyramid
2. Encoder-Decoder
3. Context modules with Atrous Convolution
4. Spatial Pyramid Pooling

**a) Image Pyramid:** This architecture consists of Deep Convolution networks that are applied to an image pyramid to extract features for different scale input where objects at different scales give different feature maps. Small scale inputs encode the long-range context, while the large scale inputs preserve the small object details to obtain optimum results.

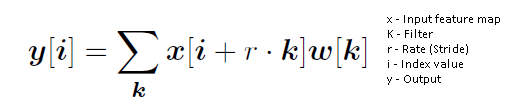
**b)** **Encoder-Decoder:** This architecture consists of two parts Encoder and Decoder. In Encoder the spatial information of feature maps is gradually reduced and longer range information is captured and the Decoder gradually recovers spatial information and object details by upsampling or inverse convolution and to finally obtain predictions, example of the encoder-decoder network are SegNet and U-Net.

**c) Context module:** This architecture consists of modules laid out in cascade to encode long-range context. Some of the modules include efficient high-dimensional filtering algorithms applied to a Deep Convolution neural network and some work include jointly train both the CRF and DCNN components to capture context information.

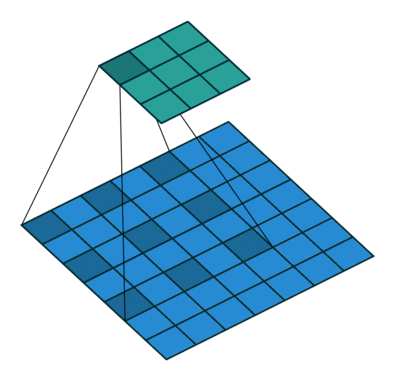
**d) Spatial pyramid pooling:** This type of architecture consists of some form of pooling module to capture Global context and Multiscale information with various parallel operations performed on features example Pspnet has a spatial pyramid pooling Module and Deeplabv2 has an atrous spatial pyramid pooling.

Atrous Convolution

Deep Convolutional Neural Networks have shown to be effective for semantic segmentation. But, the repeated combination of max-pooling and striding at consecutive layers of these networks significantly reduces the spatial resolution of the resulting feature maps. Deconvolutional layers or transposed convolution can be used to recover the spatial resolution but it’s computationally expensive, The better solution is Atrous convolution which is developed mainly for efficient computation and to preserve spatial Information.



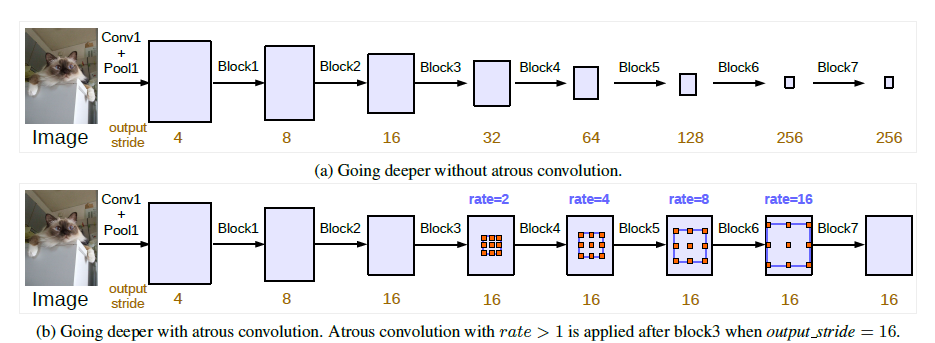
Consider two-dimensional vectors input feature map **x**, output **y**, weight **w** where for each location **i** on the output **y** and a filter **w**, atrous convolution is applied with atrous rate **r**, which is equivalent to convolving the input x with upsampled filters produced by inserting r - 1 zero between two consecutive filter values along each spatial dimension.



* Above figure represents Atrous convolution with Atrous rate=2
* If Atrous rate=1 then it’s same as normal convolution
* If Atrous rate>1 then it’s Atrous convolution or Dilated convolution with rate=r

Going Deeper with Atrous Convolution:

Atrous convolution also allows us to explicitly control how densely to compute feature maps in fully convolutional networks. Here, its called output stride the ratio of input image spatial resolution to final output resolution example if output feature map has output stride=16 that means the output resolution of feature map is 16 times smaller than the original input image.



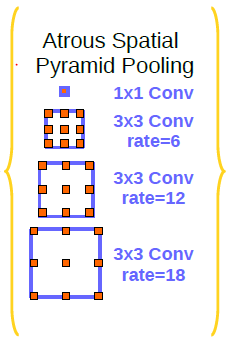
The above figure represents Restnet50 with few extra block4 repeated as block5 to block7

* **Going deeper without atrous convolution:** The first figure has an output Stride=256 where the output has a very small resolution resulting in loss of spatial information and worst performance which can be observed in experiments.
* **Going deeper with atrous convolution:** The second figure has atrous convolutional applied after block 3 and keeping output stride=16 as constant for future blocks with the increasing field of view and finally which results in preserving long-range spacial information with good performance as observed in experiments.

**MultiGrid method:**

* Multigrid represent the different atrous rate that is applied to consecutive convolutions within each block from block4 as in above figure
* Multigrid can be represented as r . (r1, r2, r3) where r1,r2,r3 are atrous rates applied to three Conv layers within a block
* Example: consider block-4 with multigrid=(1,2,4) with rate=2 then the consecutive convolutions layers in block-4 will have an atrous rate of 2.(1,2,4) = (2,4,8) respectively.

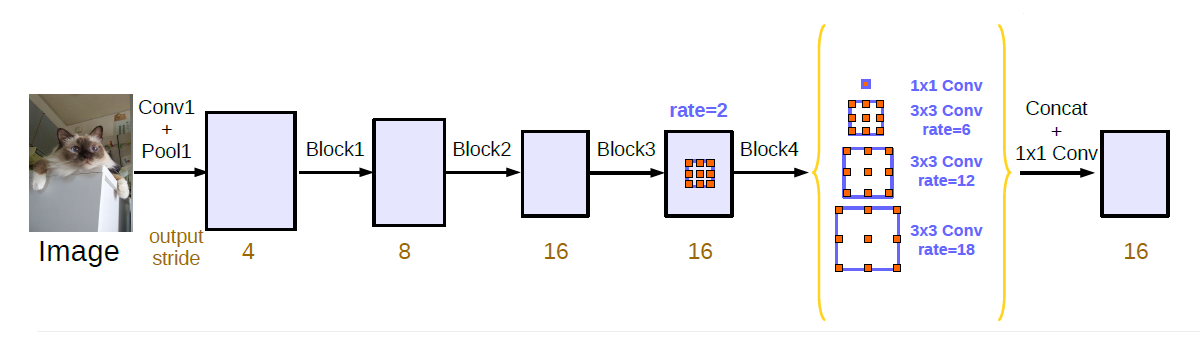
Atrous Spatial Pyramid Pooling:



* The Atrous Spatial Pyramid Pooling(ASPP) module contains parallel convolutions layer with different atrous rates to captures multi-scale information.
* The ASPP module has one 1x1 Conv and three 3x3 Conv with the atrous rate of (6,12,18) and all convolutional layers are followed by batch normalization
* if the Atrous rate value is close to the feature map size, instead of capturing the whole image context, it simply degenerates to a simple 1 x 1 filter, To overcome this, global average pooling is used before 1x1 Conv in the first layer in Atrous Spatial Pyramid Pooling

**Implementation of DeepLabV3**

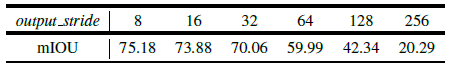
Deeplabv3 can be implemented in various configurations with varying network architecture and with increasing complexity however below is an architecture implemented as in the paper.



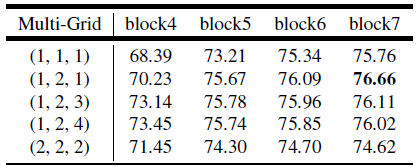
* The Deeplabv3 is built by using Restnet50 which has 4 blocks excluding Conv1+ Pool1.
* Initially Conv1 and Pool1 is performed on the image before it is fed into next block
* The block 1,2 has output stride of 4,8 and Block 3,4 has output stride=16 respectively
* Atrous convolution is only used from Block4 that has three 3x3 convolutions as in Restnet50 architecture.
* Block-4 has multigrid=(1,2,4) with rate=2 {2.(1,2,4)=(2,4,8)} which means the three 3x3 convolutions in the Block-4 has Atrous rate of 2,4,8 respectively
* The ASPP module performs parallel convolutions with 256 filters and batch normalization each on the output of Block4, First global average pooling followed by 1x1 Conv, second 3x3 Conv with rate=6, third 3x3 Conv with rate=12, fourth 3x3 Conv with rate=18 and the rates are doubled if block4 has output stride=8.
* After the above mentions parallel operations, all obtain feature maps are concatenated and pass through another 1x1 convolution with 256 filters and batch normalization.
* Finally, the feature maps are upsampled to ground truth resolution and Final prediction is obtained with a softmax layer.

**Experiments on DeeplabV3**

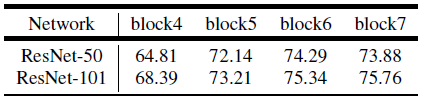
The experiments are done with the test set of Imagenet with pre-trained restnet50 for below results



* The above results are observed with deeplabV3 with the deeper network up to block7 with different output stride where it is clear that output stride has an influence on performance
* The network with output stride=8 yields better segmentation results when compared to others.

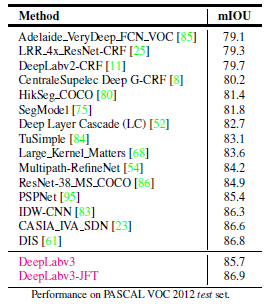


* The Multigrid (1,1,1) which is the same as normal convolution has the lowest Score and the multigrid (1,2,1) has the highest score for the deeplabV3 network.
* Multigrid settings with different Atrous rates in the network can affect the segmentation results.



* When deeper Restnet are used for deeplabV3 there is a slight increase in performance with an increase in numbers of blocks
* Deeper networks tend to give better results on the test set of Imagenet.

DeeplabV3 with PASCAL VOC Dataset



* The Above results ResNet-101 model has been pre-trained on both ImageNet and the JFT dataset resulting in a performance of 86.9% on the PASCAL VOC 2012 test set.
* More experimental observation and results in the research paper.

**Conclusion:**

* DeeplabV3 employs atrous convolution in cascade or in parallel to capture multi-scale context by adopting multiple atrous rates.
* Atrous convolution, a powerful tool to explicitly adjust filter’s field-of-view as well as control the resolution of feature responses in the application of semantic image segmentation.
* DeeplabV3 chooses to preserve long-range context information and extract dense features in the network for better results.
* From the experimental observation, it’s clear that DeeplabV3 significantly performs better than the previous version, DeeplabV1, and DeeplabV2 for semantic segmentation.