Rethinking Atrous Convolution for Semantic Image Segmentation

Liang-Chieh Chen George Papandreou Florian Schroff Hartwig Adam Google Inc.

deeplab v3

摘要

本文作者使用atrous convolution，一个可以显式地调节卷积核的感知野，也可以调节通过神经网络计算的特征的分辨率。为了解决分割物体是多尺度的问题，作者设计了使用层叠或者并行的atrous convolution，用于捕捉多尺度上下文信息，通过采用不同的膨胀率。此外，作者加强了之前的atrous Spatial Pyramid Pooling模块，用来挖掘多尺度的卷积特征，使用图片级别的特征编码全局信息，并且进一步推动性能表现。特别的，做特提出的模块包含不同rate的atrous convolution以及批正则化层，使得网络的训练也很简单。

面临的问题

因为连续的池化操作或者卷积步长导致的分辨率下降，这使得DCNNs学习到更加抽象的特征表达，然而这对局部图片变换的不变性，可能会阻碍dense prediction，在dense prediction中，我们需要详细的空间信息。为了克服这个问题，作者提出了使用atrous convolution，也被叫做dilated convolution。可以使用Imagenet愚训练的网络提取更稠密的特征图，通过移除下采样操作并且上采样对应的卷积核，等于在卷积核之间插入洞。通过atrous convolution，我们可以控制特征图计算的对应的分辨率，而不需要学习额外的参数。

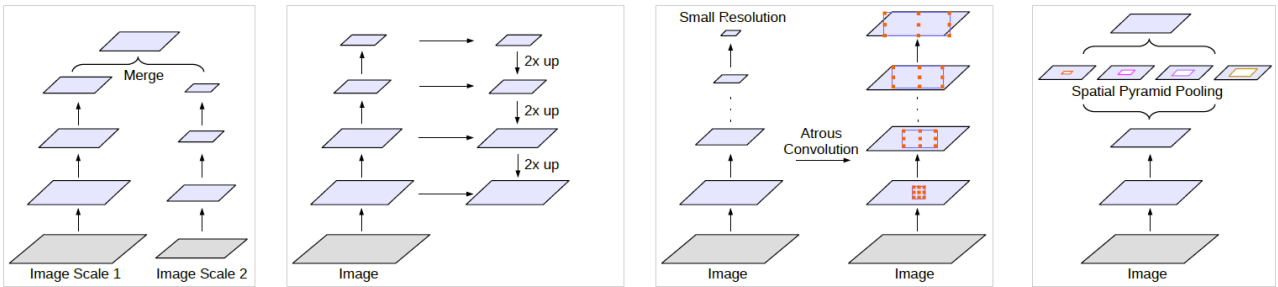
另一个问题来源于多尺度物体的存在，本文主要考虑了四种方法：1.DCNN使用了一个图片金字塔以用来对各种尺度的输入提取特征，在这里不同尺度的物体在不同不同的特征图中变得突出。2.encoder-decoder结构，利用从encoder模块得到的多尺度特征。3.额外的模块在网络的顶部被层叠，以获得捕捉更大尺度的信息，例如DenseCRF被用来编码像素级别的成对相似性。4.空间金字塔被用卷积核或池化操作以多尺度，多有效感受野挖掘输入特征图，一次捕捉不同规模的物体。

该论文解决的问题：

1. 在语义分割中由于连续的池化造成feature Scale减少，丢失位置信息。
2. 分割任务中，有的object很大但是有的object很小，它们需要的感受野大小并不相同。

相关工作

已知全局特征或上下文交流对于语义分割中正确分类像素是非常有益的，我们讨论4种用来提取上下文的全卷积神经网络FCNs，如下图。



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

Figure 2. Alternative architectures to capture multi-scale context.

Image pyramid:

将同样的模型，同样的参数应用到多尺度输入，尺寸小的输入保存范围更大的物体的信息，更大尺寸的输入保存更小的物体的信息。

Encoder-decoder:

在encoder中特征图的空间维度缓慢降低，因此更大范围的信息可以更容易被捕捉到，decoder模块把物体细节和空间信息逐渐恢复出来。

Context module:

这个模型包含额外的模块用层叠的方式来编码大范围的上下文信息，

Spatial pyramid pooling:

使用空间金字塔池化来捕捉上下文信息。

作者提出的网络是把atrous convolution作为context module来使用，并把它当作空间金字塔池化的工具。具体来说，作者复制了resnet的最后一个block，并复制为多份，并把它以层叠的方式连接，并且重新使用了ASPP模块。作者直接将层叠模块应用在特征图而不是置信图上。作者发现用batch normalization训练更好，为了更好地捕捉全局上下文信息，作者提出使用图片级别的特征来增强ASPP。

atrous convolution

atrous convolution可以适应性地调节卷积核的感知野通过改变rate，如下图。

atrous convolution也使得我们可以显式地控制全卷积网络中特征的稠密度。

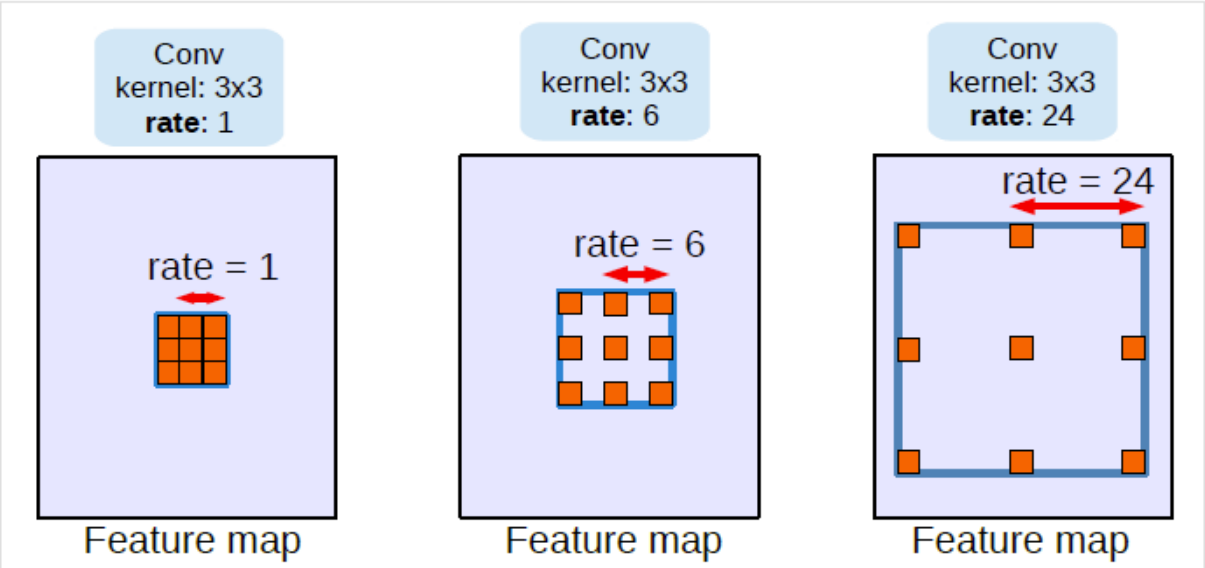


Figure 1. Atrous convolution with kernel size 3×3 and different rates. Standard convolution corresponds to atrous convolution with $rate = 1$. Employing large value of atrous rate enlarges the model's field-of-view, enabling object encoding at multiple scales.

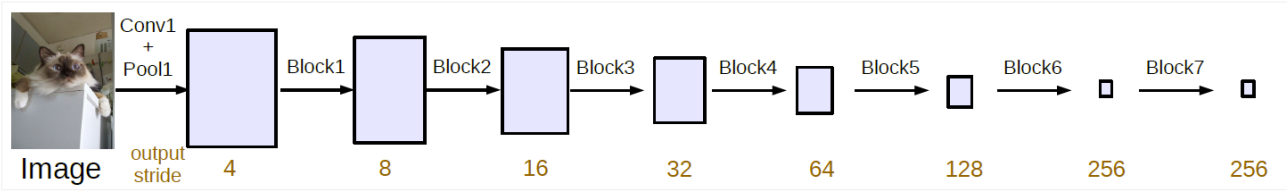
Methods

Going Deeper with Atrous Convolution

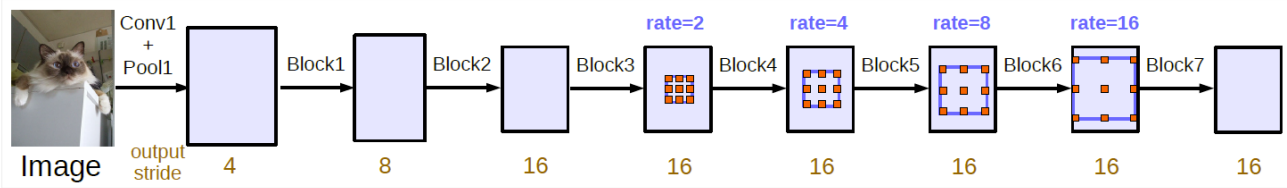
output stride：输入图片对输出图片的比值

作者把resnet的最后一个block复制了几层，如图4中的block4，让后用层叠的方式安排他们，每个block含有3个3*3卷积层，除了最后一层每层的output stride为2.除了最后一层

这种网络模型设计的动机，引入的 stride 能更容易的捕获较深的blockes中的大范围信息. 例如，整体图像feature可以融合到最后一个小分辨率的 feature map 中，如Figure3(a). 不过，这种连续的步长式设计，对于语义分割是不利的，会破坏图像的细节信息. 因此，这里采用由期望 *output_stride* 值来确定 rates 的atrous convolution 进行模型设计，如Figure3(b). 采用串行的ResNet， 级联block为block5、block6、block7，均为block4的复制，如果没有 atrous convolution， 其*output_stride*=256.



(a) Going deeper without atrous convolution.

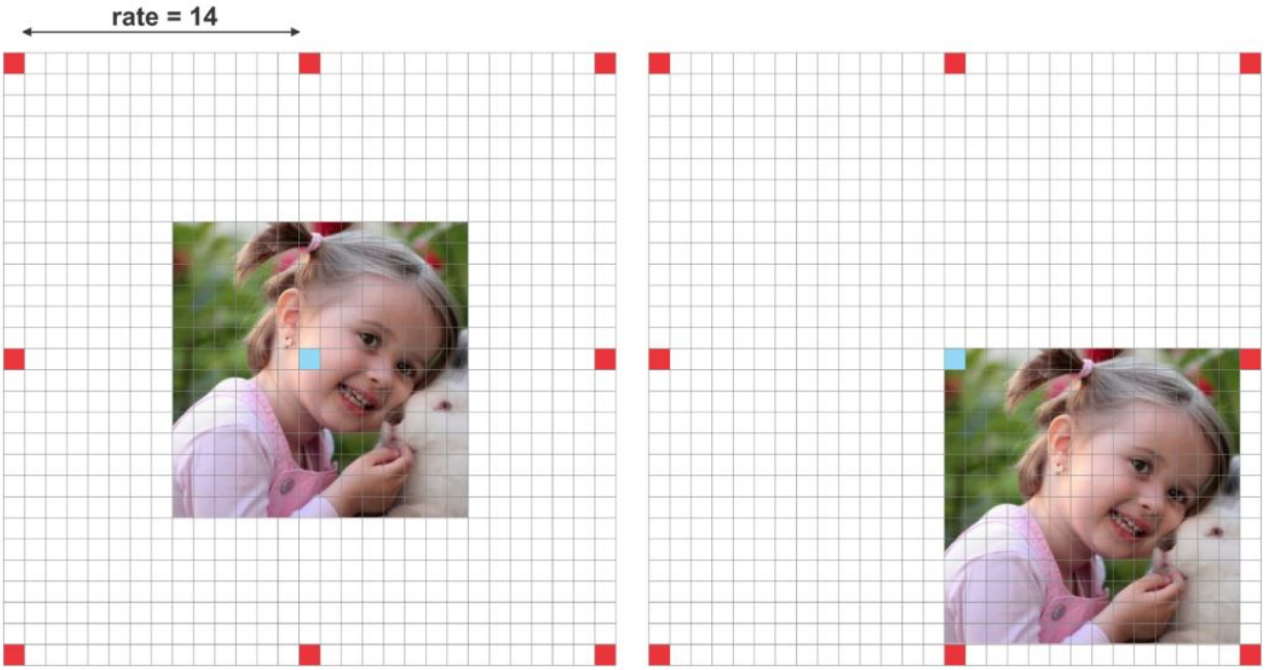


(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

Figure 3. Cascaded modules without and with atrous convolution.

Atrous Spatial Pyramid Pooling

aspp是被spp启发的，spp表现出它可以把以不同的规模重新采样图片以高效准确地把任意尺度的区域进行分类，作者还加入了批正则化。不同间隔的ASPP可以高效地捕捉多尺度信息，然而，作者发现随着采样率增大，有效卷积核权重就越少，在极端情况下，atrous 卷积退回成为1*1卷积，因为只有中间的卷积核是有效的。如下图



为了克服这个问题，并把全局上下文信息合并到这个模型中，作者采用图片级别的特征，作者在最后一张特征图上使用全局池化，把产生的图片级别特征送给一个有256个卷积核的1*1卷积中，然后双线性上采样特征到想要的空间维度。作者提出的ASPP包含一个1×1卷积，3个3×3卷积，膨胀率分别是6，12，18，（当输出步长为16时），所有分支生成的特征被串连，然后通过另一个1×1卷积，在最终1×1卷积以生成最终logits之前。