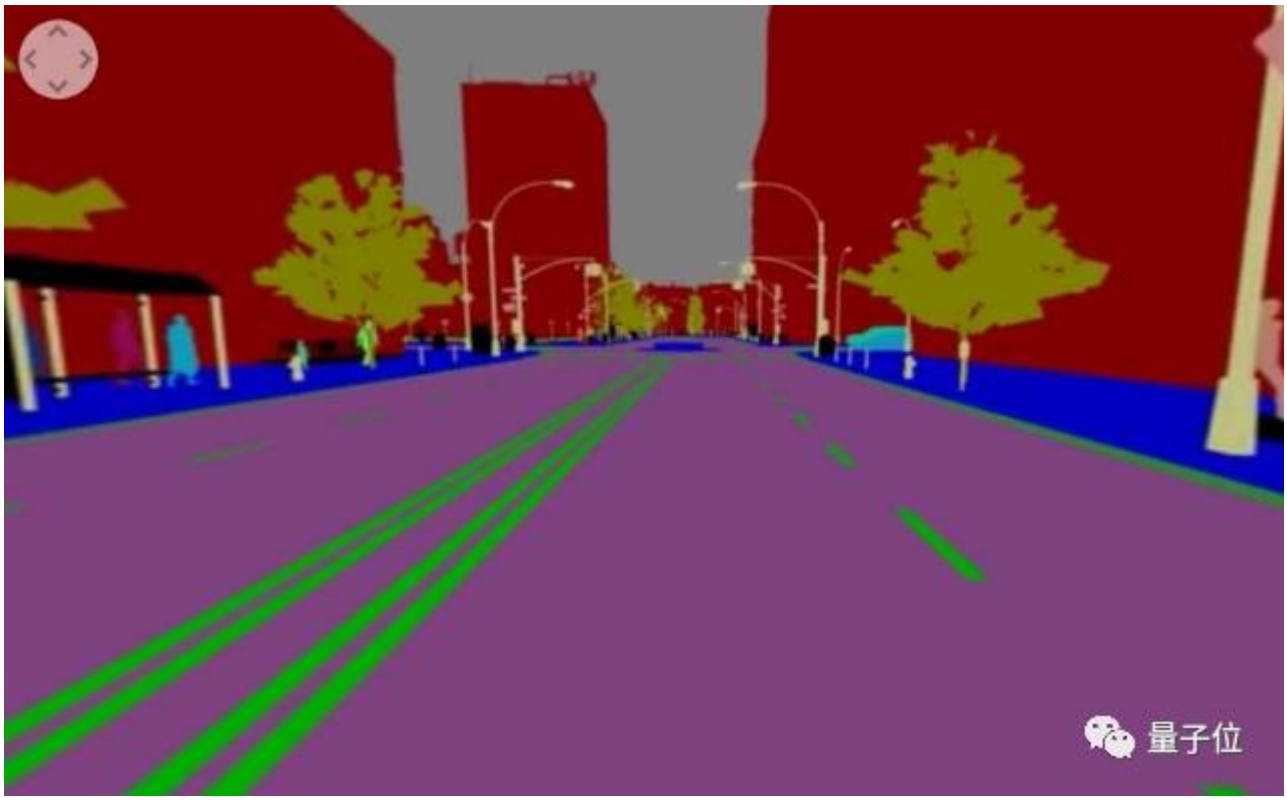
语义分割中的深度学习方法全解：从FCN、SegNet到各版本DeepLab - jacke121的专栏 - CSDN博客

原文：https://www.sohu.com/a/155907339_610300

图像语义分割就是机器自动从图像中分割出对象区域，并识别其中的内容。



量子位今天推荐的这篇文章，回顾了深度学习在图像语义分割中的发展历程。

发布这篇文章的Qure.ai，是一家用深度学习来读取医学影像的公司，他们在官方博客上梳理了语义分割中的深度学习方法。

他们希望通过这份介绍，能让大家了解这个已经在自然图像处理比较成熟、但是在医学图像中仍需发展的新兴技术。

作者Sasank Chilamkurthy三部分介绍了语义分割相关研究：

以下内容编译自Qure.ai官方博客：

语义分割是什么？

语义分割方法在处理图像时，具体到像素级别，也就是说，该方法会将图像中每个像素分配到某个对象类别。下面是一个具体案例。

△左边为输入图像，右边为经过语义分割后的输出图像。

该模型不仅要识别出摩托车和驾驶者，还要标出每个对象的边界。因此，与分类目的不同，相关模型要具有像素级的密集预测能力。

目前用于语义分割研究的两个最重要数据集是VOC2012和MSCOCO。

VOC2012：

http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

MSCOCO：

http://mscoco.org/explore/

有哪些方法?

在深度学习应用到计算机视觉领域之前，研究人员一般使用纹理基元森林(TextonForest)或是随机森林(Random Forest)方法来构建用于语义分割的分类器。

卷积神经网络(CNN)不仅能很好地实现图像分类，而且在分割问题中也取得了很大的进展。

最初，图像块分类是常用的深度学习方法，即利用每个像素周围的图像块分别将各像素分成对应的类别。其中，使用图像块的主要原因是分类网络通常具有全连接层，其输入需为固定大小的图像块。

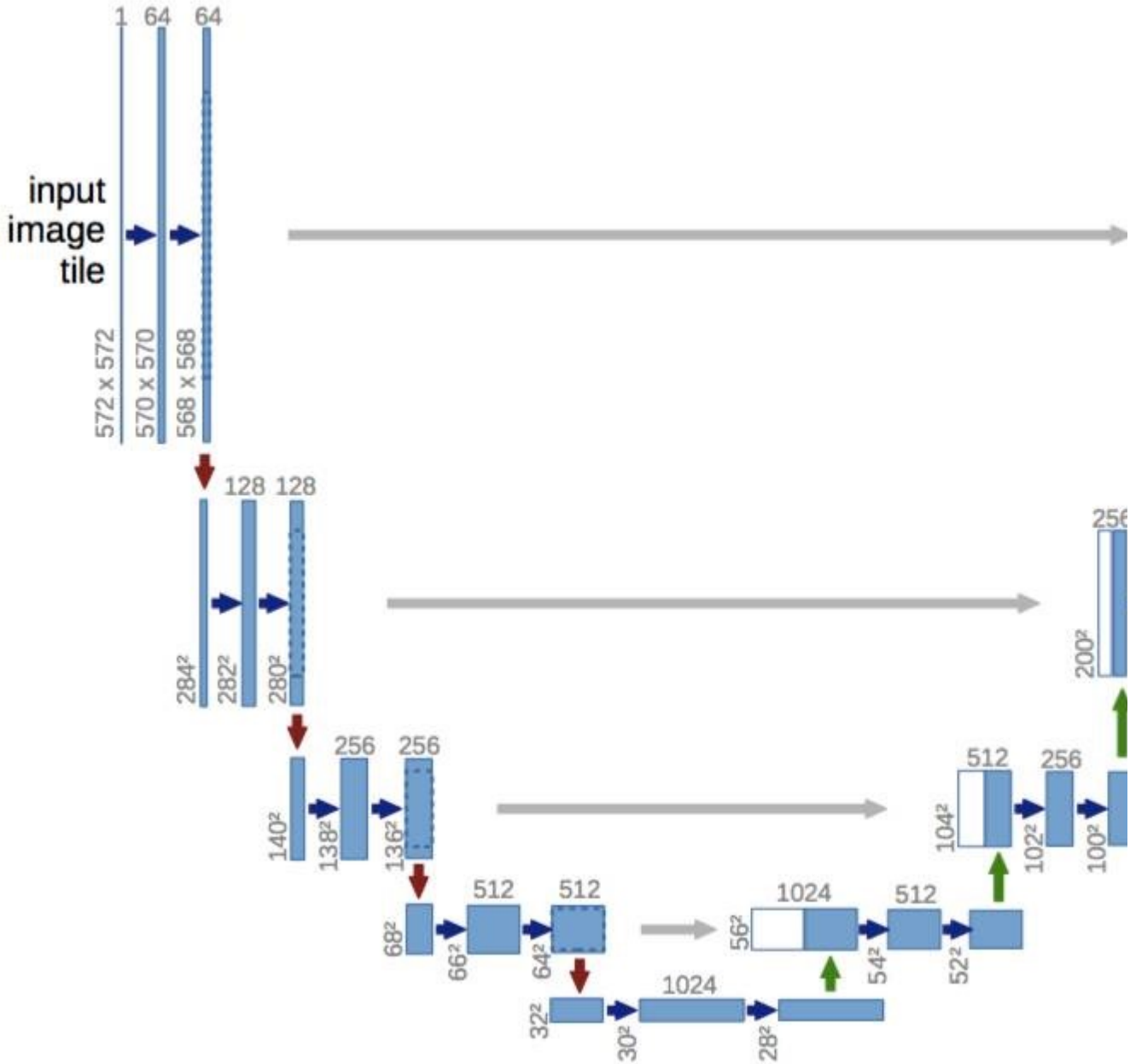
2014年，加州大学伯克利分校的Long等人提出的完全卷积网络(Fully Convolutional Networks)，推广了原有的CNN结构，在不带有全连接层的情况下能进行密集预测。

这种结构的提出使得分割图谱可以生成任意大小的图像，且与图像块分类方法相比，也提高了处理速度。在后来，几乎所有关于语义分割的最新研究都采用了这种结构。

除了全连接层结构，在分割问题中很难使用CNN网络的另一个问题是存在池化层。池化层不仅能增大上层卷积核的感受野，而且能聚合背景同时丢弃部分位置信息。然而，语义分割方法需对类别图谱进行精确调整，因此需保留池化层中所舍弃的位置信息。

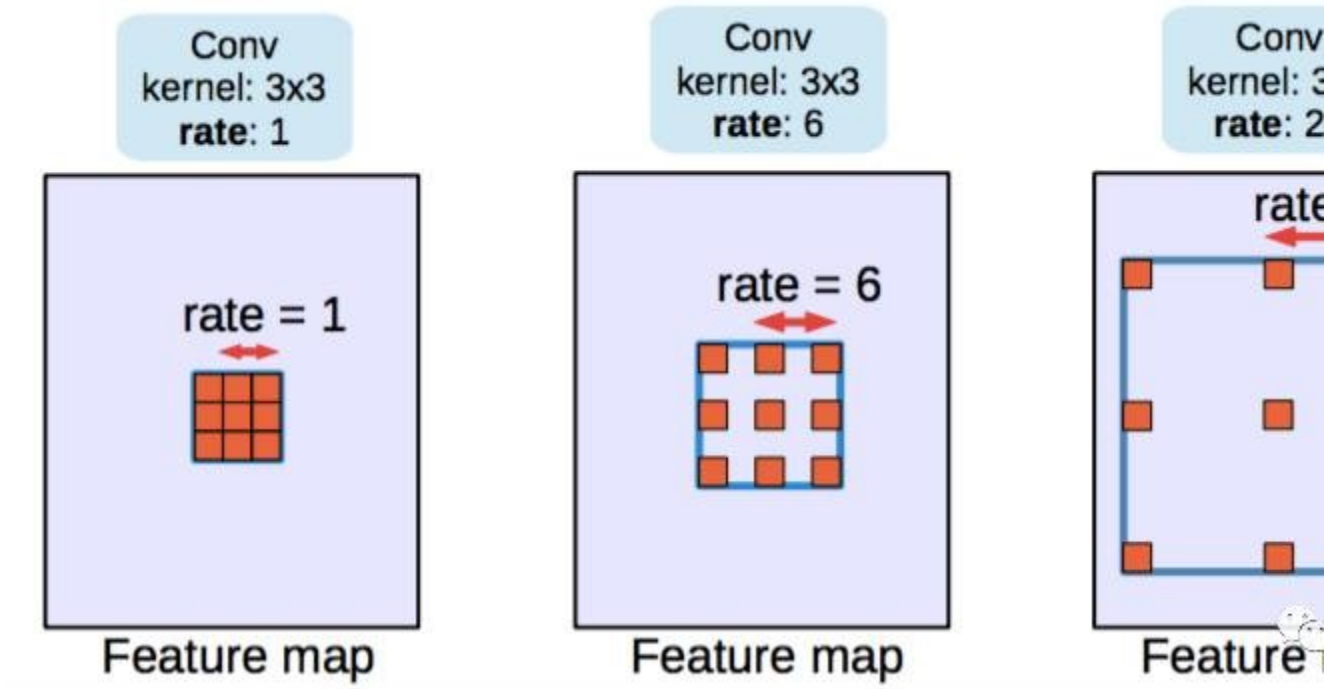
研究者提出了两个不同形式的结构来解决这个问题。

第一种方法是编码器-解码器(encoder-decoder)结构。其中，编码器使用池化层逐渐缩减输入数据的空间维度，而解码器通过反卷积层等网络层逐步恢复目标的细节和相应的空间维度。从编码器到解码器之间，通常存在直接的信息连接，来帮助解码器更好地恢复目标细节。在这种方法中，一种典型结构为U-Net网络。



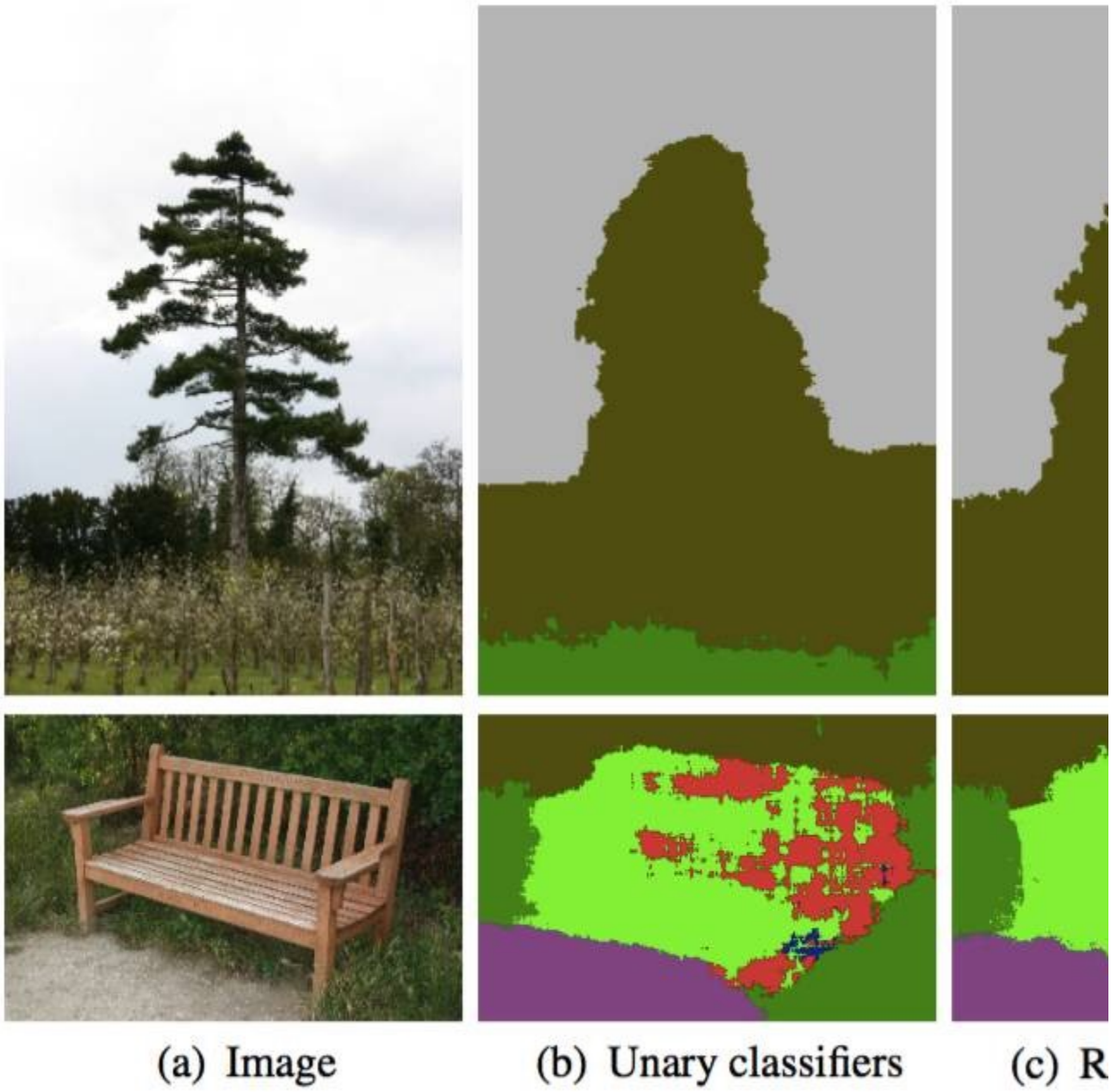
△一种典型的编码器-解码器结构U-Net

第二种方法使用了称作空洞卷积的结构，且去除了池化层结构。



△空洞卷积，当比率为1时，即为经典的卷积结构。

条件随机场(Conditional Random Field，CRF)方法通常在后期处理中用于改进分割效果。CRF方法是一种基于底层图像像素强度进行“平滑”分割的图模型，在运行时会将像素强度相似的点标记为同一类别。加入条件随机场方法可以提高1~2%的最终评分值。



△发展中的CRF方法效果。b图中将一维分类器作为CRF方法的分割输入；c、d、e图为CRF方法的三种变体；e图为广泛使用的一种CRF结构。

接下来，我们会梳理一些代表性论文，来介绍从FCN网络开始的分割结构演变历程。

这些结构都使用了VOC2012数据集来测试实际效果。

一些有趣的研究

接下来将按照论文的发表顺序来介绍以下论文：

1．FCN网络；

2．SegNet网络；

3．空洞卷积(Dilated Convolutions)；

4．DeepLab (v1和v2)；

5．RefineNet；

6．PSPNet；

7．大内核(Large Kernel Matters)；

8．DeepLab v3；

对于上面的每篇论文，下面将会分别指出主要贡献并进行解释，也贴出了这些结构在VOC2012数据集中的测试分值IOU。

FCN

论文：

Fully Convolutional Networks for Semantic Segmentation

于2014年11月14日提交到arvix

https://arxiv.org/abs/1411.4038

主要贡献：

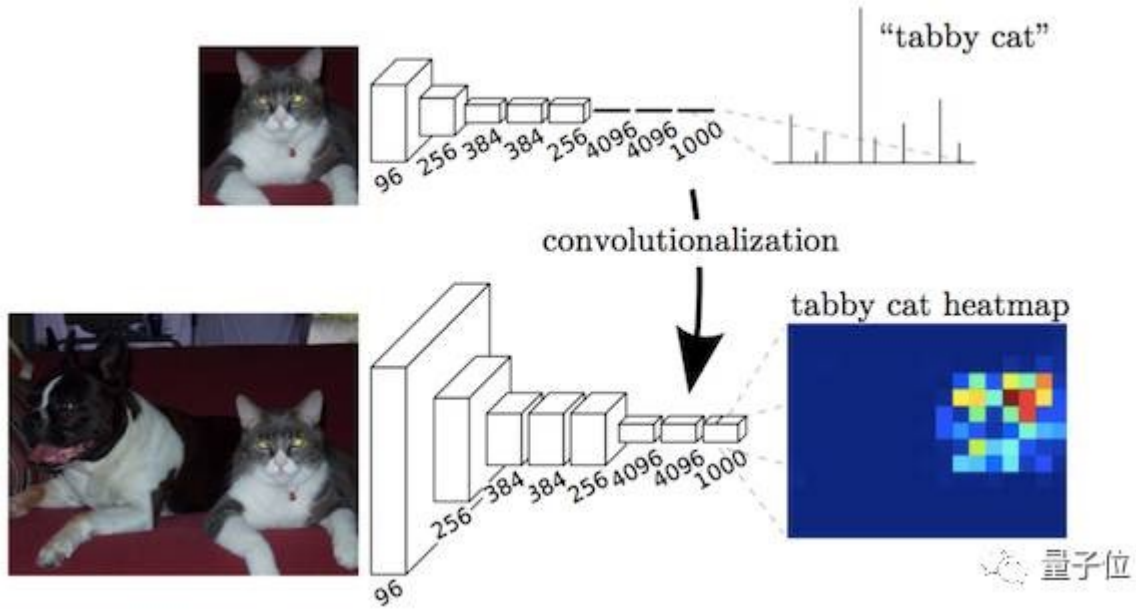
- 将端到端的卷积网络推广到语义分割中；
- 重新将预训练好的Imagenet网络用于分割问题中；
- 使用反卷积层进行上采样；
- 提出了跳跃连接来改善上采样的粗糙程度。

具体解释：

本文的关键在于：分类网络中的全连接层可以看作是使用卷积核遍历整个输入区域的卷积操作。

这相当于在重叠的输入图像块上评估原始的分类网络，但是与先前相比计算效率更高，因为在图像块重叠区域，共享计算结果。

尽管这种方法并不是这篇文章中所特有的，还有一篇关于overfeat的文章也使用了这种思想，但是确实显著提高了在VOC2012数据集上的实际效果。



△用卷积运算实现的全连接层结构

在将VGG等预训练网络模型的全连接层卷积化之后，由于CNN网络中的池化操作，得到的特征图谱仍需进行上采样。

反卷积层在进行上采样时，不是使用简单的双线性插值，而是通过学习实现插值操作。此网络层也被称为上卷积、完全卷积、转置卷积或是分形卷积。

然而，由于在池化操作中丢失部分信息，使得即使加上反卷积层的上采样操作也会产生粗糙的分割图。因此，本文还从高分辨率特性图谱中引入了跳跃连接方式。

分值	评论	来源
62.2	无	排行榜
67.2	增大动量momentum(原文未描述)	排行榜

△FCN网络在VOC2012上测试的基准分值

个人评论：

本文的研究贡献非常重要，但是最新的研究已经很大程度地改进了这个结果。

SegNet

论文：

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

于2015年11月2日提交到arvix

https://arxiv.org/abs/1511.00561

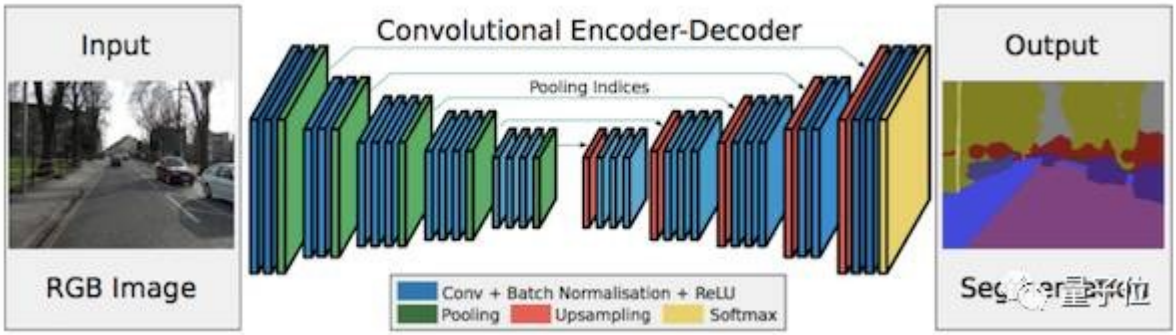
主要贡献：

将最大池化指数转移至解码器中，改善了分割分辨率。

具体解释：

在FCN网络中，通过上卷积层和一些跳跃连接产生了粗糙的分割图，为了提升效果而引入了更多的跳跃连接。

然而，FCN网络仅仅复制了编码器特征，而Segnet网络复制了最大池化指数。这使得在内存使用上，SegNet比FCN更为高效。



△SegNet网络结构

△SegNet在VOC2012上测试的基准分值

个人评论：

FCN网络和SegNet网络都是最先出现的编码器-解码器结构，但是SegNet网络的基准分值还不能满足可实际使用的需求。

空洞卷积

论文：

Multi-Scale Context Aggregation by Dilated Convolutions

于2015年11月23日提交到arvix

https://arxiv.org/abs/1511.07122

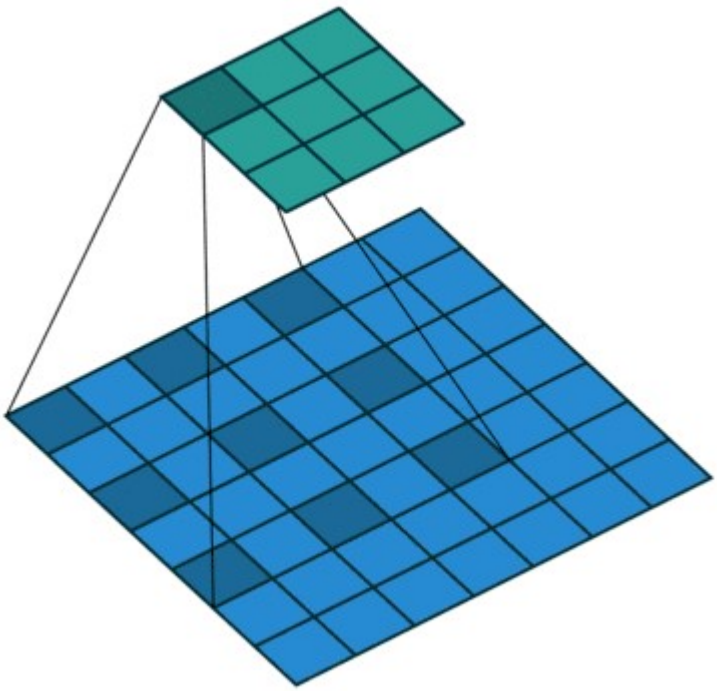
主要贡献：

- 使用了空洞卷积，这是一种可用于密集预测的卷积层；
- 提出在多尺度聚集条件下使用空洞卷积的“背景模块”。

具体解释：

池化操作增大了感受野，有助于实现分类网络。但是池化操作在分割过程中也降低了分辨率。

因此，该论文所提出的空洞卷积层是如此工作的：



△空洞卷积示意图

空洞卷积层在不降低空间维度的前提下增大了相应的感受野指数。

在接下来将提到的DeepLab中，空洞卷积被称为多孔卷积(atrous convolution)。

从预训练好的分类网络中(这里指的是VGG网络)移除最后两个池化层，而用空洞卷积取代了随后的卷积层。

特别的是，池化层3和池化层4之间的卷积操作为空洞卷积层2，池化层4之后的卷积操作为空洞卷积层4。

这篇文章所提出的背景模型(frontend module)可在不增加参数数量的情况下获得密集预测结果。

这篇文章所提到的背景模块单独训练了前端模块的输出，作为该模型的输入。该模块是由不同扩张程度的空洞卷积层级联而得到的，从而聚集多尺度背景模块并改善前端预测效果。

分值	评论	来源
71.3	前端	空洞卷积论文
73.5	前端+背景	同上
74.7	前端+背景+ CRF	同上
75.3	前端+背景+ CRF - RNN	同上

△空洞卷积在VOC2012上测试的基准分值

个人评论：

需要注意的是，该模型预测分割图的大小是原图像大小的1/8。这是几乎所有方法中都存在的问题，将通过内插方法得到最终分割图。

DeepLab(v1和v2)

论文1：

Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs

于2014年12月22日提交到Arvix

https://arxiv.org/abs/1412.7062

论文2：

DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

于2016年6月2日提交到Arxiv

https://arxiv.org/abs/1606.00915

主要贡献：

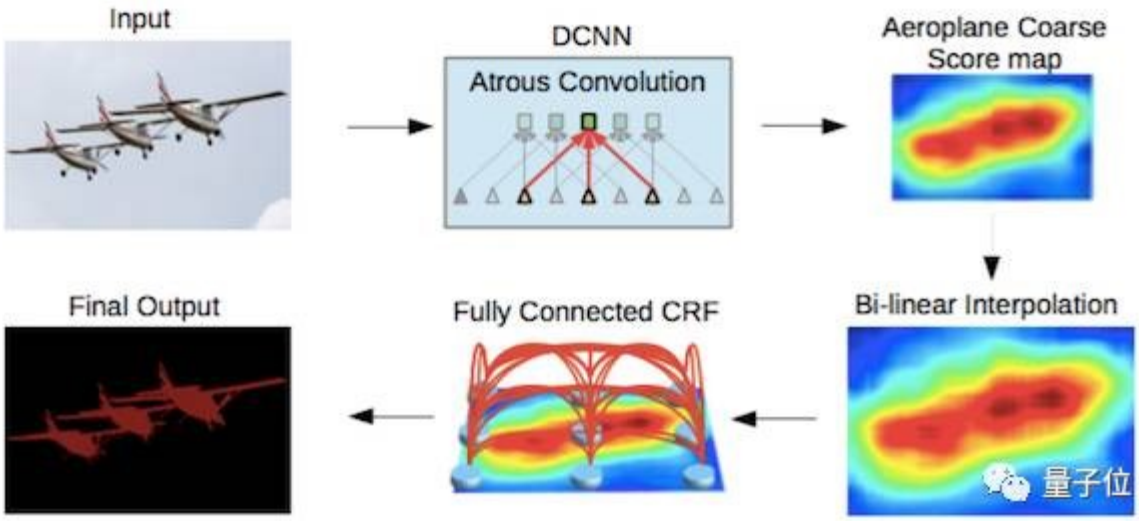
- 使用了空洞卷积；
- 提出了在空间维度上实现金字塔型的空洞池化atrous spatial pyramid pooling(ASPP)；
- 使用了全连接条件随机场。

具体解释：

空洞卷积在不增加参数数量的情况下增大了感受野，按照上文提到的空洞卷积论文的做法，可以改善分割网络。

我们可以通过将原始图像的多个重新缩放版本传递到CNN网络的并行分支(即图像金字塔)中，或是可使用不同采样率(ASPP)的多个并行空洞卷积层，这两种方法均可实现多尺度处理。

我们也可通过全连接条件随机场实现结构化预测，需将条件随机场的训练和微调单独作为一个后期处理步骤。



△DeepLab2网络的处理流程

分值	评论	来源
79.7	ResNet-101 + 空洞卷积 + ASPP + CRF	排行榜

△DeepLab2网络在VOC2012上测试的基准分值 RefineNet

论文：

RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation

于2016年11月20日提交到Arxiv

https://arxiv.org/abs/1611.06612

主要贡献：

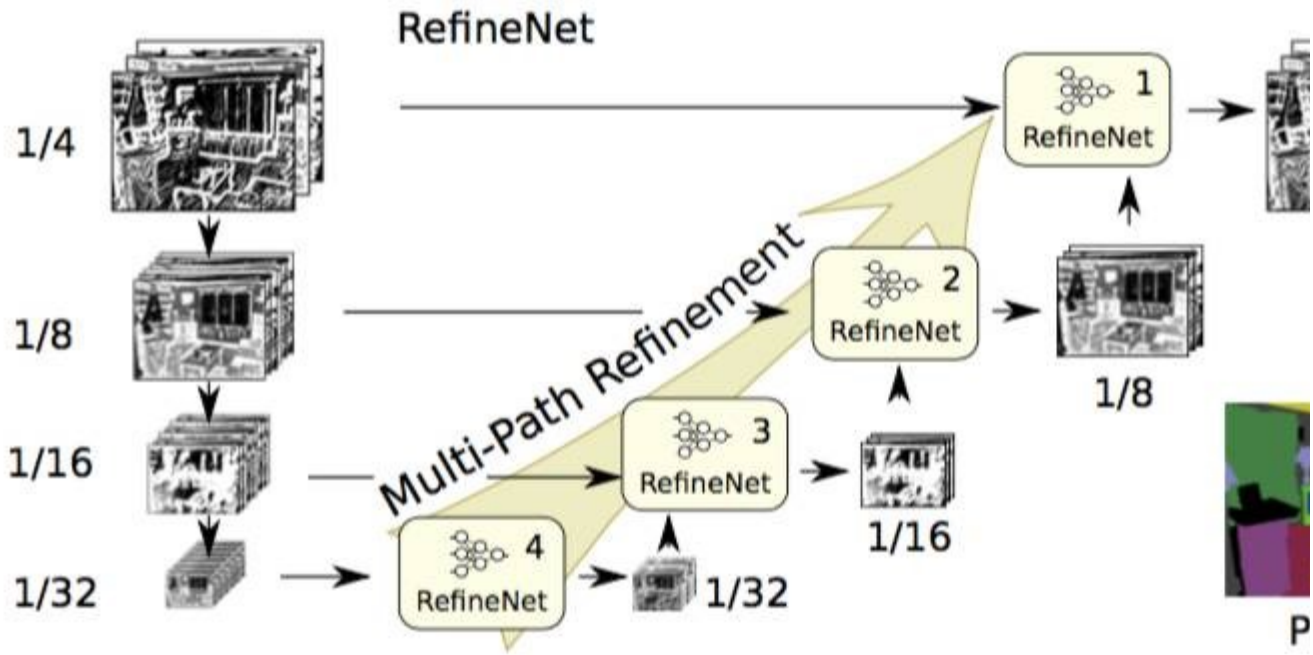
- 带有精心设计解码器模块的编码器-解码器结构；
- 所有组件遵循残差连接的设计方式。

具体解释：

使用空洞卷积的方法也存在一定的缺点，它的计算成本比较高，同时由于需处理大量高分辨率特征图谱，会占用大量内存，这个问题阻碍了高分辨率预测的计算研究。

DeepLab得到的预测结果只有原始输入的1/8大小。

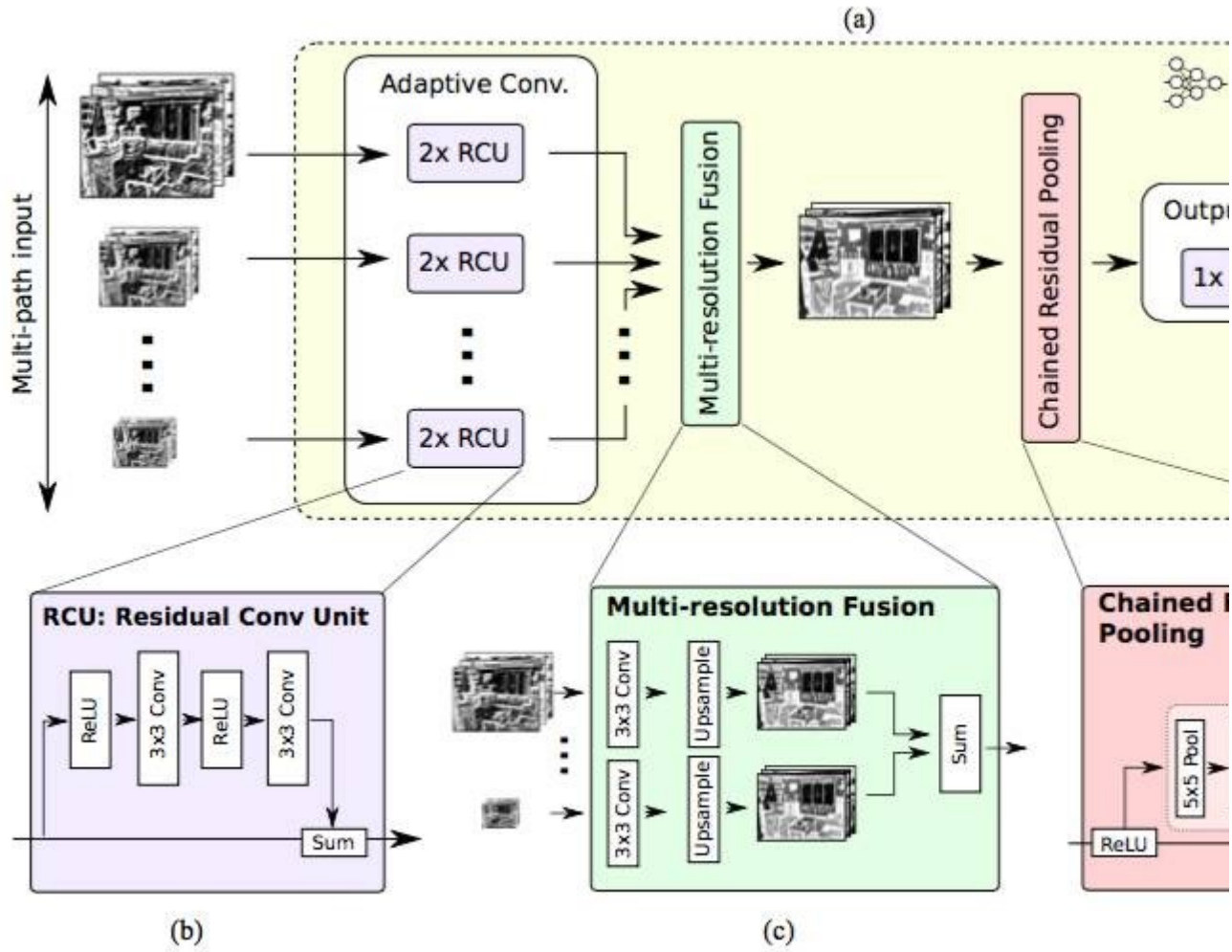
所以，这篇论文提出了相应的编码器-解码器结构，其中编码器是ResNet-101模块，解码器为能融合编码器高分辨率特征和先前RefineNet模块低分辨率特征的RefineNet模块。



△RefineNet网络结构

每个RefineNet模块包含一个能通过对较低分辨率特征进行上采样来融合多分辨率特征的组件，以及一个能基于步幅为1及5×5大小的重复池化层来获取背景信息的组件。

这些组件遵循恒等映射的思想，采用了残差连接的设计方式。



△RefineNet模块

分值	评论	来源
84.2	CRF + 多维度输入 + COCO预训练	排行榜

△RefineNet网络在VOC2012上测试的基准分值 PSPNet

论文：

Pyramid Scene Parsing Network

于2016年12月4日提交到Arxiv

https://arxiv.org/abs/1612.01105

主要贡献：

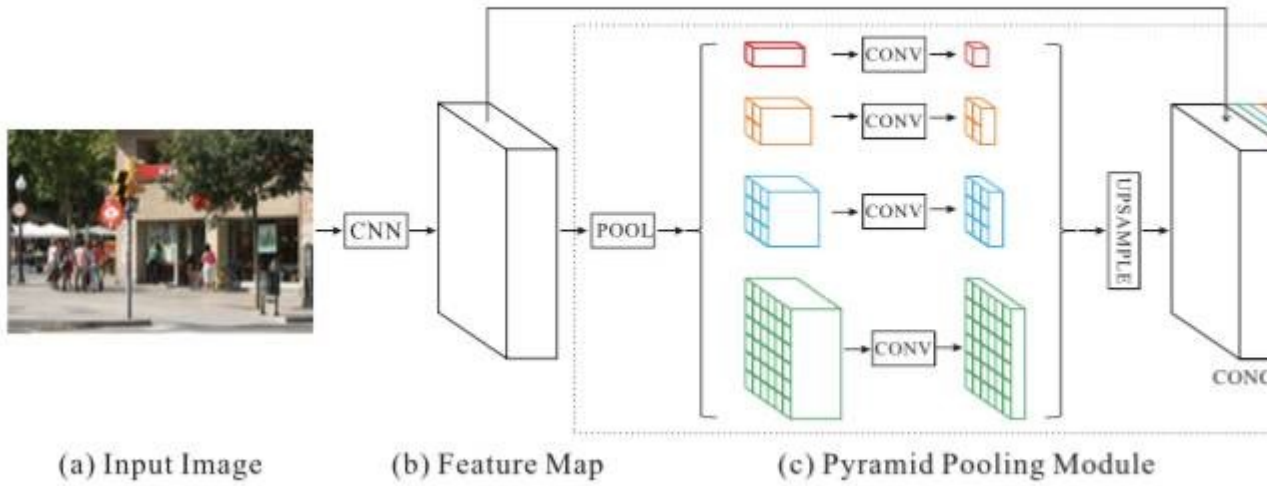
- 提出了金字塔池化模块来聚合背景信息；
- 使用了附加损失(auxiliary loss)。

具体解释：

全局场景分类很重要，由于它提供了分割类别分布的线索。金字塔池化模块使用大内核池化层来捕获这些信息。

和上文提到的空洞卷积论文一样，PSPNet也用空洞卷积来改善Resnet结构，并添加了一个金字塔池化模块。该模块将ResNet的特征图谱连接到并行池化层的上采样输出，其中内核分别覆盖了图像的整个区域、半各区域和小块区域。

在ResNet网络的第四阶段(即输入到金字塔池化模块后)，除了主分支的损失之外又新增了附加损失，这种思想在其他研究中也被称为中级监督(intermediate supervision)。



△PSPNet网络结构

分值	评论	来源
85.4	COCO预训练，多维度输入，无CRF方法	排行榜
82.6	无COCO预训练方法，多维度输入，无CRF方法	PSPNet论文

△PSPNet网络在VOC2012上测试的基准分值 大内核

论文：

Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network

于2017年3月8日提交到Arxiv

https://arxiv.org/abs/1703.02719

主要贡献：

提出了一种带有大维度卷积核的编码器-解码器结构。

具体解释：

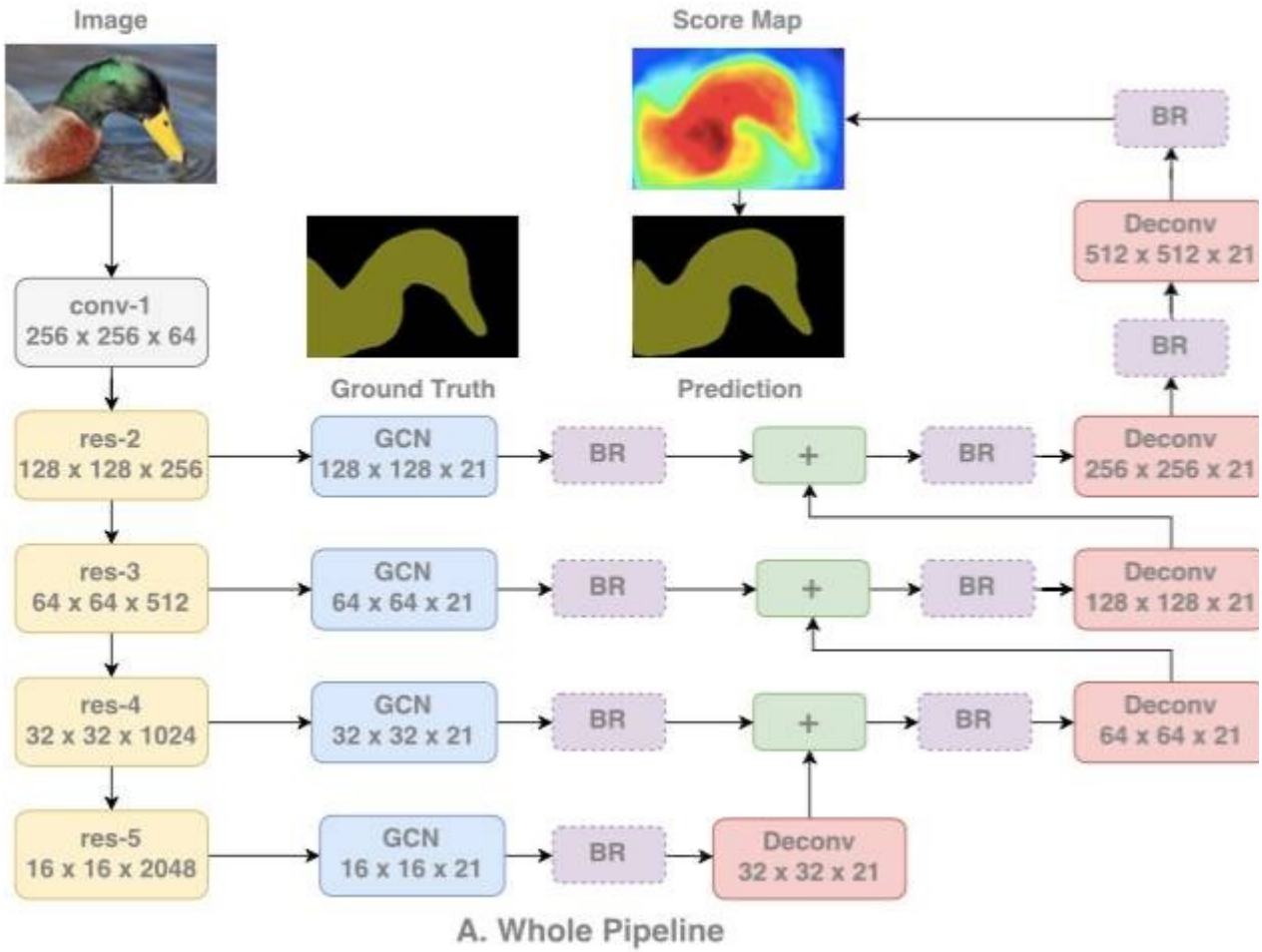
这项研究通过全局卷积网络来提高语义分割的效果。

语义分割不仅需要图像分割，而且需要对分割目标进行分类。在分割结构中不能使用全连接层，这项研究发现可以使用大维度内核来替代。

采用大内核结构的另一个原因是，尽管ResNet等多种深层网络具有很大的感受野，有相关研究发现网络倾向于在一个小得多的区域来获取信息，并提出了有效感受野的概念。

大内核结构计算成本高，且具有很多结构参数。因此，k×k卷积可近似成1×k + k×1和k×1 + 1×k的两种分布组合。这个模块称为全局卷积网络(Global Convolutional Network, GCN)。

接下来谈结构，ResNet(不带空洞卷积)组成了整个结构的编码器部分，同时GCN网络和反卷积层组成了解码器部分。该结构还使用了一种称作边界细化(Boundary Refinement，BR)的简单残差模块。



△GCN网络结构

分值	评论	来源
82.2	-	详情见本论文
83.6	改进训练过程，未在本文中详细描述	排行榜

△GCN网络在VOC2012上测试的基准分值 DeepLab v3

论文：

Rethinking Atrous Convolution for Semantic Image Segmentation

于2017年6月17日提交到Arxiv

https://arxiv.org/abs/1706.05587

主要贡献：

- 改进了空间维度上的金字塔空洞池化方法(ASPP)；
- 该模块级联了多个空洞卷积结构。

具体解释：

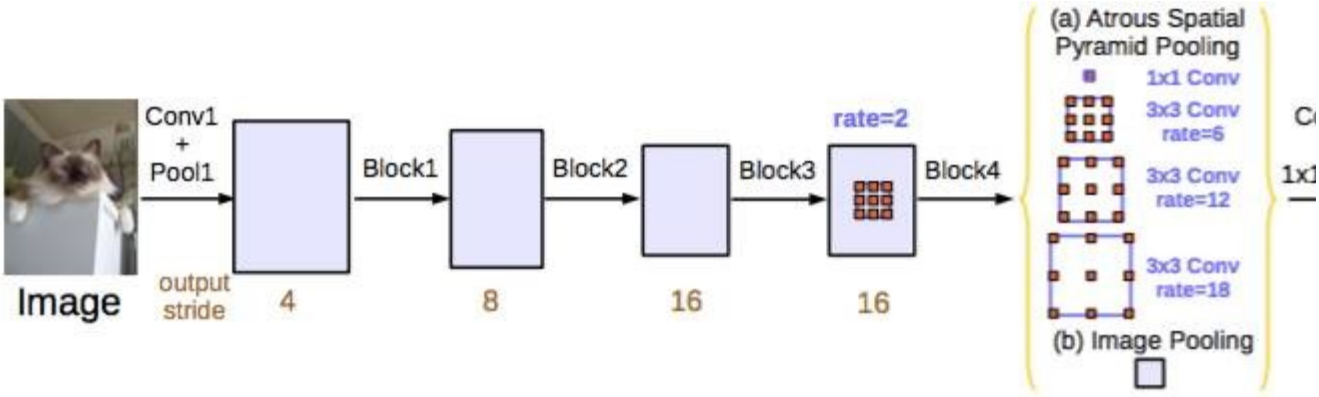
与在DeepLab v2网络、空洞卷积中一样，这项研究也用空洞卷积/多空卷积来改善ResNet模型。

这篇论文还提出了三种改善ASPP的方法，涉及了像素级特征的连接、加入1×1的卷积层和三个不同比率下3×3的空洞卷积，还在每个并行卷积层之后加入了批量归一化操作。

级联模块实际上是一个残差网络模块，但其中的空洞卷积层是以不同比率构建的。这个模块与空洞卷积论文中提到的背景模块相似，但直接应用到中间特征图谱中，而不是置信图谱。置信图谱是指其通道数与类别数相同的CNN网络顶层特征图谱。

该论文独立评估了这两个所提出的模型，尝试结合将两者结合起来并没有提高实际性能。两者在验证集上的实际性能相近，带有ASPP结构的模型表现略好一些，且没有加入CRF结构。

这两种模型的性能优于DeepLabv2模型的最优值，文章中还提到性能的提高是由于加入了批量归一化层和使用了更优的方法来编码多尺度背景。



△DeepLabv3 ASPP结构

分值	评论	来源
85.7	使用了ASPP结构，且不带有级联模块	排行榜

△DeepLabv3 ASPP结构在VOC2012上测试的基准分值

原文地址：

http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review