ParseNet: Looking Wider to See Better

老师们同学们大家好。我今天为大家带来的这篇论文是关于使用深度学习来提升图像识别的准确率，相关的研究也已经有很多，今天这篇是2015年发布的，名字叫做，ParseNet: Looking Wider to See Better。

今天将会提到的技术ParseNet是一种优化过的图像识别技术。这种技术将global context（全局上下文）添加到FCN（完全卷积网络）中以进行semantic segmentation（语义分割）的技术。FCN作为图像语义分割的先河，实现像素级别的分类（即end to end，pixel-wise），为后续使用CNN作为基础的图像语义分割模型提供重要基础。

P2 p3

FCN中的CNN主要由卷积，pooling与全连接构成。卷积主要是获取高维特征，pooling使图片缩小，全连接与传统神经网络相似作为权值训练，最后通过softmax输出概率最高的类别。假设feature map（特征图）大小为227x227,经过卷积与pooling后得到55x55的特征图（一层的特征图可以有多个类别）。注意，不同的卷积操作可能会对图片大小产生影响，FCN论文中pooling使图片缩小1/2。经过多次卷积后特征图大小为13x13,特征图的权值展开为1维与后面的权值实现全连接，最后使用softmax输出类别。这就是CNN的大致网络结构与分类过程。

P4

FCN的upsample意思为上采样，简单来说就是pooling的逆过程，所以pooling也就是下采样，采样后数据数量减少，upsample采样后数据数量增多。FCN作者在论文中讨论了3种upsample方法，最后选用的是反卷积的方法（FCN作者称其为后卷积）使图像实现end to end，可以理解upsample就是使大小比原图像小得多的特征图变大，使其大小为原图像大小。这边不对FCN的具体原理进行赘述。

Global context p5

这篇文章研究fcn的receptive field的问题。receptive field是什么呢。我把它翻译成感知视野。

已知的上下文关系对于改善深度学习的性能是非常有用的，语义分割在本地信息（local information）时通常是不明确的，如果可以从整个图像获得上下文关系就会有很大不同。尽管理论上顶层网络具有非常大的感知视野，但在实际运用中，感知视野的有效尺寸要小得多，不足以捕捉全局的图像信息。为了过扩大有效的感知视野，作者用一个滑动的噪声去干扰输入图像，观察网络的输出，用来探测一个网络的有效感受野具体有多大。考虑这么一个问题：如果你只看到一个物体的一个部分你不会知道他是什么。比如一张床的床脚和椅子的脚长得很像，如果仅仅看着脚，完全不知道是椅子还是床脚。对于这个像素点你就很难判断它到底是什么。如果你不仅仅看脚，你看多一些，那么就可以确定这个东西到底属于什么东西了。

下面我们来看一下这几张图 p5

可以发现FCN的结果输出的东西非常乱，这就是缺乏一个“全局信息”的结果。所以到底什么是receptive field呢。在CNN中，receptive field的意义就是一个像素点能够感受到的关于原图的区域的大小。如果这个区域越大，那么对于这个点能得到的信息就越精准。

RF（receptive field）感知视野 p6

1. 原始图像; （b）来自网络最后一层的channel的 activation map; （c）网络架构中能够接受的最大感知范围 （d）影响最后结果的感知视野。 显然最后的视野不足以捕捉全部的图像。

Early Fusion and late fusion（早期融合和晚期融合）

一旦取得了Global Context的特征，就可以使用两个通用的融合方式，一个是早期融合，然后放入分类器一起分类，另一种就是晚期融合，就是分类后再融合。如果没有额外的处理，则两种方式是一样的。一般来说，早期融合可以很好利用更多特征，这个是晚期融合做不到的。

实验

本次试验，主要使用了VOC 2012、PASCAL-Context和SiftFlow的基准数据集。

IoU(Intersection over Union)

Intersection over Union是一种测量在特定数据集中检测相应物体准确度的一个标准。我们可以在很多物体检测挑战中，例如PASCAL VOC challenge中看多很多使用该标准的做法。

通常我们在 HOG + Linear SVM object detectors 和 Convolutional Neural Network detectors (R-CNN, Faster R-CNN, YOLO, etc.)中使用该方法检测其性能。这个测量方法和你在任务中使用的物体检测算法没有关系。

IoU是一个简单的测量标准，只要是在输出中得出一个预测范围(bounding boxex)的任务都可以用IoU来进行测量。为了可以使IoU用于测量任意大小形状的物体检测，我们需要：

1、 ground-truth bounding boxes（人为在训练集图像中标出要检测物体的大概范围）；

2、我们的算法得出的结果范围。

也就是说，这个标准用于测量真实和预测之间的相关度，相关度越高，该值越高。另外mean IU指的是不同类别识别准确度的平均值，比如一幅图中要识别三个物体，mean IU就是三个物体分别准确度加起来的平均值。

（因为是Conference的关系，篇幅有限，里面并未涉及具体如何实现，这边将会主要介绍测试的环境和结果）

我们先来看PASCAL-Context的结果，我们使用了FCN32s，可以看到他的mean iou，就是准确度大概是三十多这个样子，然后从Pascal context找到的参数带到VOC 2012的结果，作者在PASCAL VOC2012上重现DeepLab和DeepLab-LargeFOV，我们能够看到准确率提升了不少，deeplab下我们重现出来的结果大概好5%，deeplabLargeFOV用了过滤稀释技术（atrous算法）参数少得多，而且更快，这样我们也获得了3.5的进步。

我们将deeplablargeFOV的baseline设定为我们parsenet baseline，然后parsenet是parsenet baseline加上Global context。

P11

我们也使用了siftFlow的dataset，我们使用fcn32s和在PASCAL-context里面找到的参数，去做训练，我们发现添加更多层可以提升性能，我们发现在这种情况下，添加Global Context不会对感知视野有太多的帮助，因为fc7已经足够覆盖到大部分的输入图像了。

P12

最终我们测试了两个模型，parsenet baseline和parsenet 使用voc2012测试集，从图可知，baseline的结果已经高过很多现存的方法，在添加了Global Context的情况下，结果会更好。

结论，parsenet是一个简单的完全卷积神经网络架构，他可以将Global Context加入进语义分割中，在实验中已经很明显发现，fcn网络最大的感知视野不能充分的覆盖所有范围，而parsenet可以明显的提升感知视野，而且显著强于没有Global Context的baseline。