**FCN: Fully Convolutional Networks for Semantic Segmentation**

**Abstract**

卷积网络是能够产生特征层次结构的强大视觉模型。论文表明，通过卷积网络训练端到端，像素到像素，都超过了最先进语义分割模型。论文关键是构建“完全卷积”网络，这些网络接受任意大小的输入，并通过有效的推理和学习产生相应大小的输出。我们定义并详细描述完全卷积网络，解释了它们在空间密集预测任务中的应用，并与之前的模型建立了联系。我们将当代分类网络（AlexNet，VGGnet，GoogleNet）应用到完全卷积网络中，并通过微调将其学习到的表征转移到分割任务中。然后，我们去细化了一种新的结构，它将来自深、粗层的语义信息与来自浅、细层的外观信息相结合，以产生准确和详细的分割。

我们的模型转化最近分类上的成功继续密集预测，通过重新解释分类网络以全卷积网络和微调它们学习表现的方法

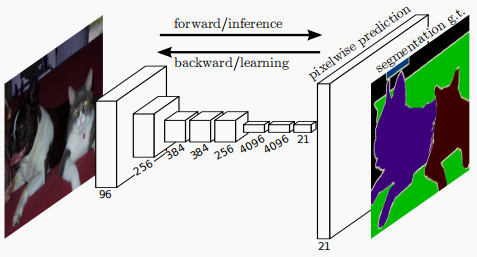


图1 全卷积网络能有效学习进行语义分割（对每个像素点进行分类）

2.Related work

解释FCN设计以及密度预测折中方案，介绍了我们网络内上采样与多层组合的结构体系，并且描述我们的实验框架。

我们的方法基于最近深度网络中图像分类以及迁移学习的成功。迁移第一次被使用在多种视觉识别任务，然后检测，以及在混合建议分类器的实例与语义分割。我们现在重构以及微调分类网络以构建直接且密集预测的语义分割模型。我们绘制FCNs的空间，并在这个框架中定位了先前的历史模型和最近的模型。

Fully convolutional networks

据我们所知，将卷积网络扩展任意大小输入的想法首次被Matan等人所提出，它扩展了经典的LeNet网络结构来识别数字字符串（主要是手写字体）。由于他们的网络仅限于一维的输入字符串，Matan等人使用维特比解码来获得它们的输出。Wolf和Platt [40]将邮箱地址输出扩展为邮政地址块四个角的检测分数的二维地图。这些历史操作都是为了检测而进行推理和学习的全卷积。Ning等人用完全卷积推断定义线虫组织的粗糙细胞的分类分割。

全卷积计算在当今许多的多层次网络也被利用。 比如Sermanet等人的滑动窗口检测,Pinheiro和Collobert的语义分割以及Eigen等人的图像恢复都使用了全卷积推理. 全卷积训练是很少见的，但Tompson等人有效地使用了学习一种端到端的局部检测和姿态估计的空间模型方法。尽管他们不去解释或分析这种方法。

除此之外，He等人使用丢弃分类网络的非卷积部分来制作特征提取器。他们将proposals和空间金字塔池合并在一起，以产生用于分类的本地化的固定长度特征。尽管快速且有效，但是这种混合模型不能进行端到端的学习。

基于卷积网的密集预测近期的一些工作已经将卷积网应用于密集预测问题，其中包括Ning等人的语义分割。,Farabet等人以及Pinheiro和Collobert ；Ciresan等人的电子显微镜边界预测以及Ganin和Lempitsky的通过混合卷积网和最邻近模型的处理自然场景图像; 还有Eigen等人的图像修复和深度估计。这些方法的相同点包括如下：

限制容量和接受范围的小模型；

patchwise学习;

超像素投影的预处理，随机正则化、滤波和局部分类；

输入移位和密度输出的隔行交错输出；

多尺度金字塔处理；

饱和双曲线正切非线性；

集成。

然而我们的方法确实没有这种机制。但是我们研究patchwise（3.4）训练以及从FCNs角度出发的“shift-and-stitch”dense输出（3.2）。我们也讨论网络内的上采样（3.3）。其中Eigen等人的全连通预测，是一个特例。

不像这些存在的方法，我们采用以及扩展深度分类网络，使用图像分类作为有监督的预训练，并且从输入图像和整个图像ground thruths通过全卷积网络去微调进行简单且有效的学习。

Hariharan等人和Gupta等人也改编深度分类网到语义分割，但是也在混合proposal-classifier模型中这么做了。这些方法通过采样边界框和region proposal进行微调了R-CNN系统,用于检测、语义分割和实例分割。这两种办法都不能进行端到端的学习。

我们通过跨层和融合特征来定义一种非线性的局部到整体的表述用来协调端到端。在现今的工作中Hariharan等人也在语义分割的混合模型中使用了多层。

3.Fully convolutional networks

在网络中每层的数据是尺寸为h\*w\*d的三位数组，其中h和w是空间维度，d是特征或者通道维度。第一层是像素尺寸为h\*w、颜色通道为d的图像。高层的位置和图像中它们连通的位置相对应，被称为接收域。

卷积网是以平移不变性为基础。其基本组成部分（卷积，池化以及激活函数）作用于局部输入域，并且只依赖相对空间坐标。Xij记为在特定层中位置（i，j）的数据向量。Yij对于下一层，其计算公式如下：



k指的是卷积核大小，s是步长或下采样因素，fks决定了层的类型：一种用于卷积或者平均池化的矩阵乘法，空间池化需要空间最大化，或者一个有激活作用的elementwise nonlinearity，亦或是层的其他种类。

其卷积核大小和步长遵从转换规则，则其表示形式如下：



当一个普通深度网络计算一个普通的非线性函数，一个只有这种形式的层的网络计算一个非线性滤波，其我们称之为深度滤波或全卷积网络。正常情况下FCN可以计算任意尺寸的输入并且产生相应（或许重采样）空间维度的输出。

一个实值损失函数由FCN定义的任务决定。如果损失函数是最后一层的空间维度综合，，它的梯度将是它每层空间组成的梯度总和。所以在全部图像上的基于l的随机梯度下降计算将和基于l'的梯度下降结果一样，将最后一层的所有接收域作为minibatch（分批处理）。在这些接收域显著重叠的情况下，前反馈计算和反向传播计算整图的叠层都比独立的patch-by-patch有效的多。

我们接下来解释怎样将分类网络转化为产生粗糙输出图的全卷积网络。对于像素级预测，我们需要将粗输出连接回像素。章节3.2描述了一个技巧，其目的是为了OverFeat。我们通过重新诠释等效网络的修改来直观了解这个技巧。作为一个有效的替换，我们引入的deconvolution layers为了上采样在3.3节。在3.4节，考虑训练通过像素级采样，并且在4.3中证明，我们的整个图像训练更快且同样有效。

3.1.Adapting classifiers for dense prediction

经典的识别网络，包括LeNet，AlexNet，以及一些后继者，表面上采用固定大小的输入，并产生非空间输出。这些网络的全连接层有着固定的尺寸并且丢弃了空间坐标。然而，这些全连接层可被视为覆盖整张输入区域的卷积层。将它们转换为完全卷积网络就可以接受任何大小的输入和输出分类图。转化如下图所示：

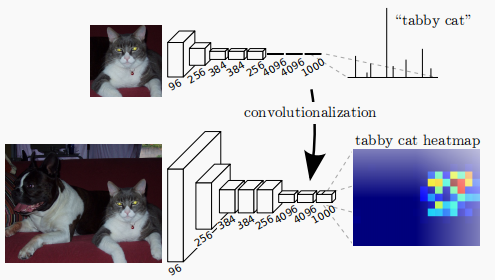


图2：将全连接层转换为卷积层，使分类网络输出热图，添加层与空间损失实现高效端到端密集学习机制。

此外，当结果图在特殊的输入patches等于原始网络的估计，那些patches在重叠区域的计算是均匀分摊。

3.2.Shift-and-stitch is filter rarefaction

3.3．Upsampling is backwards strided convolution

我们最好的分割架构使用这些层来学习上采样用以微调预测。见4.2节。

3.4.Patchwise training is loss sampling

全图式训练是有效且高效的。

4.分割架构

我们将ILSVRC分类器应用到FCNs并结合内网上采样和像素级损失来增大dense prediction的能力。我们通过微调来练分割。接下来，我们构建了一个新的跨层结构，它结合粗、语义、和局部的表征信息来重新预测。

为此，我们训练并验证了PASCAL VOC2011分割挑战赛。我们使用每像素多项式逻辑损失进行训练，并使用联合上平均像素相交的标准度量进行验证，平均值覆盖所有类别，包括背景。该训练忽略了在ground truth（模糊或困难）的像素。

4.1.From classifier to dense FCN

我们选择了VGG16网络，发现其在此实验中相当于19层网络。对于GooLeNet，我们仅使用最后的损失层，并且通过丢弃最后的平均池化层来提高性能。我们通过丢弃最后的分类层、切去每层网络头，用卷积层来替换全连接层。我们附加了一个1\*1的、通道维数为21的卷积来预测每个PASCAL分类（包括背景）的得分在每个粗糙的输出位置，后面紧跟一个去卷积层用来双线性上采样粗糙输出到像素密集输出。我们发现最好的结果在以一个固定的学习速率得到（最少175个epochs)。

从分类到分割的微调对每层网络有一个合理的预测。甚至最坏的模型也能达到大约75%的良好表现。内设分割的VGG网络（FCN-VGG16）已经在val上平均IU 达到了56.0取得了最好的成绩，相比于52.6。在额外数据上的训练将FCN-VGG16提高到59.4，将FCN-AlexNet提高到48.0。尽管相同的分类准确率，我们的用GoogLeNet并不能和VGG16的分割结果相比较。

4.2.Combining what and where

我们为分段定义了一个新的完全卷积网络（FCN），它结合了特征层次的各个层，并细化了输出的空间精度。参见图3。

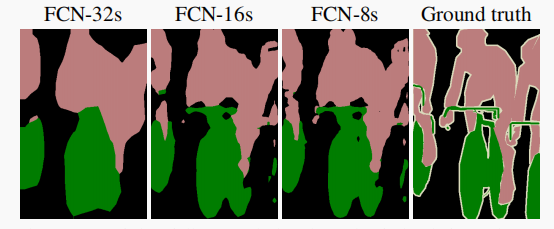


图3

虽然完全卷积分类器可以按照4.1所示进行分割，甚至在标准度量上获得高分，但它们的输出却非常粗糙。最终预测层的32像素步长限制了上采样输出中的细节比例。

我们通过添加链接来解决这个问题，这个链接将最后的预测层与步幅更小的较低层相结合。

我们首先通过预测16pixel的步长层将输出步长分成两半。我们在pool4之上添加了一个1\*1的卷积以产生额外的预测。我们通过添加一个2倍的上采样层并将两个预测相加。将该输出与在第32步conv7（卷积FC7）上计算的预测进行融合。我们为了双线性插值进行2倍的上采样初始化，但允许参数按照第3.3节所述学习。最后，将步长为16的预测上采样回图像。我们将这个网络称为FCN-16s。FCN-16s端到端学习，最后用一个较粗的网络进行参数初始化，这称之为FCN-32s。作用于pool4的新参数初始化为零，以便网络以未修改的预测开始。学习率降低了100倍。

我们继续以这种方式将来自pool3的预测与来自pool4和conv7加上2倍上采样的预测进行融合，构建FCN-8s网络。发现平滑度和细节略有改善。所以不再融合更低的层。

Refinement by other means

减少池化层的步长是获得好性能最直接的方法。然后，这样做对于我们的VGG-16基础网络是有问题的。将pool5设置为步幅为1需要我们的卷积fc6具有14\*14的内核大小，以保持其感受野的尺寸。

另一种获得精细预测的方法就是利用3.2节中描述的shift-and-stitch技巧。在有限的实验中，我们发现从这种方法的提升速率比融合层的方法花费的代价更高。

4.3.Experimental framework

Optimization

我们用SGD momentum进行训练。们使用20个图像的小批量和固定学习率10-3,10-4和5-5用于FCN-AlexNet，FCN-VGG16和FCN-GoogLeNet分别由线搜索选择。 我们使用动量0.9，重量衰减为5-4或2-4，并加倍学习偏见，虽然我们发现了训练只对学习率敏感。 我们零初始化class scoring layer，因为随机初始化都没有产生更好的性能，更快的收敛。Dropout包括在内在原始分类器网中使用的位置。

Fine-tuning

我们通过反向传播微调所有图层。）对于粗糙的FCN-32s，一个普通的GPU需要3天的训练时间，升级到FCN-16s和FCN-8s版本训练大约需要1天。

Patch Sampling

我们的全图有效地训练每张图片batches到常规的、大的、重叠的patches网格。相反的，先前工作随机样本patches在一整个数据集，可能导致更高的方差batches，可能加速收敛。我们通过空间采样之前方式描述的损失研究这种折中，以1-p的概率做出独立选择来忽略每个最后层单元。为了避免改变有效的批次尺寸，我们同时以因子1/p增加每批次图像的数量。我们发现采样相对于全图式训练在收敛速率上没有很显著的效果，但是由于每个批次都需要大量的图像，很明显的需要花费更多的时间。

Class balancing

完全卷积训练可以通过加权或采样损失来平衡类。虽然我们的标签有点不平衡（大约有3/4是背景），但我们发现没有必要进行类平衡。

Dense Prediction

通过网络内的反褶积层将分数向上采样到输入维度。最后一层反褶积滤波器固定为双线性插值，而中间上采样层初始化为双线性上采样，然后学习。不使用移位和缝合（第3.2节）或过滤器稀疏等效物。

Augmentation

我们尝试通过随机镜像和“抖动”图像来增加训练数据，将图像在每个方向上转换到32像素（最粗略的预测尺度）。这没有产生明显的改善。

More Training Data

PASCAL VOC 2011分割训练设置1112张图片的标签。Hariharan等人为一个更大的8498的PASCAL训练图片集合收集标签，被用于训练先前的先进系统,SDS。训练数据将FCV-VGG16得分提高了3.4个百分点到59.4。

Implementation

所有型号均使用Caffe在单一NVIDIA Tesla K40c上进行培训和测试。

结论：

全卷积网络是一类丰富的模型，是现代分类网络其中的特例。认识到这一点，将这些分类网络扩展到分割，并通过多分辨率层组合改进体系结构，极大地提高了最先进的技术水平，同时简化和加快了学习和推理。