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

Abstract

在这项工作中，我们通过深度学习解决了语义图像分割的任务，并做出三个主要贡献，通过实验证明具有实质性的实用价值。首先，我们使用上采样滤波器或“萎缩卷积”来强调卷积，作为密集预测任务中的强大工具。 Atrous卷积允许我们明确地控制在深度卷积神经网络中计算特征响应的分辨率。它还允许我们有效地扩大滤波器的视野，以在不增加参数数量或计算量的情况下结合更大的上下文。其次，我们提出了一个不稳定的空间金字塔池（ASPP）来在多个尺度上稳健地分割对象。 ASPP使用多个采样率和有效视场的过滤器探测进入的卷积特征层，从而捕获多个尺度的对象和图像上下文。第三，我们通过结合DCNN和概率图形模型的方法来改进对象边界的定位。 DCNN中通常部署的最大池和下采样的组合实现了不变性，但是对定位精度有影响。我们通过将最终DCNN层的响应与完全连接的条件随机场（CRF）相结合来克服这个问题，CRF在定性和定量方面都有所提高，以提高定位性能。我们提出的“DeepLab”系统在PASCALVOC-2012语义图像分割任务中设置了新的先进技术，在测试集中达到了79.7％mIOU，并将结果推进到其他三个数据集：PASCAL-Context，PASCAL-Person-部分和城市景观。我们所有的代码都在网上公开发布。

Introduction

深度卷积神经网络（DCNNs）[1]已经将计算机视觉系统的性能推向了各种高级问题的飙升，包括图像分类[2]，[3]，[4]，[5]， [6]和目标检测[7]，[8]，[9]，[10]，[11]，[12]，其中以端到端方式训练的DCNN比依赖于系统的系统提供了明显更好的结果 手工制作的功能。 这种成功的关键在于DCNN对局部图像变换的内置不变性，这使得他们可以学习越来越抽象的数据表示[13]。 这种不变性对于分类任务显然是可取的，但是可能妨碍密集预测任务，例如语义分割，其中空间信息的抽象是不期望的。

特别地，我们考虑将DCNN应用于语义图像分割中的三个挑战：（1）降低的特征分辨率，（2）多尺度的对象的存在，以及（3）由于DCNN不变性而降低的定位精度。接下来，我们将讨论这些挑战以及我们在我们提出的DeepLab系统中克服它们的方法。

第一个挑战是由最初为图像分类设计的连续DCNN层执行的最大池和下采样（'跨步'）的重复组合引起的[2]，[4]，[5]。当DCNN以完全卷积的方式使用时，这导致特征图的空间分辨率显着降低[14]。为了克服这个障碍并有效地生成更密集的特征映射，我们从DCNN的最后几个最大池化层中移除下采样运算符，而是对后续卷积层中的滤波器进行上采样，从而得到以更高采样率计算的特征映射。过滤器上采样相当于在非零滤波器分接头之间插入孔（法语中为“trous”）。这种技术在信号处理方面有着悠久的历史，最初是为了在一种被称为“算法”的方案中对未抽样小波变换进行有效计算而开发的。一个愚蠢的“[15]。我们使用术语atrous卷积作为与上采样滤波器卷积的简写。在[3]，[6]，[16]中，在DCNN的背景下已经使用过这种想法的各种方法。在实践中，我们通过组合atrous卷积来恢复全分辨率特征图，该卷积更加密集地计算特征图，然后是对原始图像尺寸的特征响应的简单双线性插值。该方案提供了一种简单而强大的替代方法，可在密集预测任务中使用去卷积层[13]，[14]。与具有较大滤波器的常规卷积相比，紊乱卷积允许我们有效地扩大滤波器的视野，而不增加参数的数量或计算量。

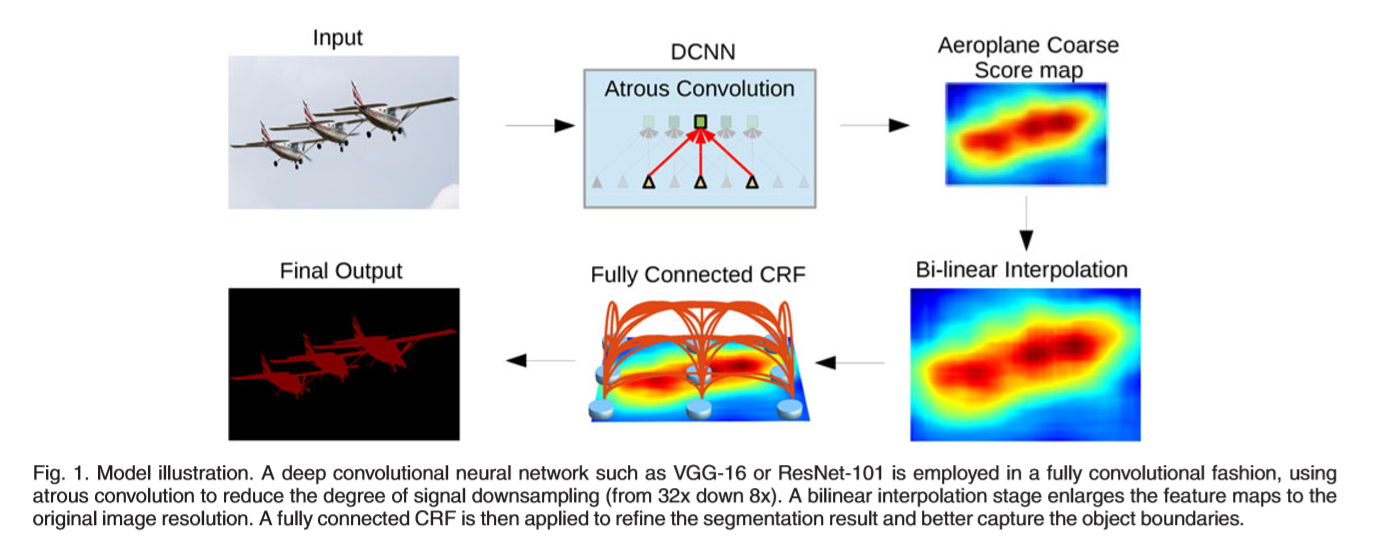
第二个挑战是由多个尺度的物体的存在引起的。解决这个问题的一种标准方法是向DCNN呈现相同图像的重新缩放版本，然后聚合特征或得分图[6]，[17]，[18]。我们表明这种方法确实提高了我们系统的性能，但是以输入图像的多个缩放版本的所有DCNN层计算特征响应为代价。相反，在空间金字塔汇集[19]，[20]的推动下，我们提出了一种计算有效的方案，在卷积之前以多种速率重新采样给定的特征层。这相当于使用具有互补有效视野的多个过滤器探测原始图像，从而在多个尺度上捕获对象以及有用的图像上下文。我们不是实际重新采样特征，而是使用具有不同采样率的多个并行的迂回卷积层来有效地实现这种映射。我们将所提出的技术称为“巨大的空间金字塔池”（ASPP）。

第三个挑战涉及这样一个事实，即以物体为中心的分类器需要空间变换的不变性，固有地限制了DCNN的空间精度。缓解此问题的一种方法是在计算最终分割结果时使用跳过层从多个网络层中提取“超列”特征[14]，[21]。我们的工作探索了一种我们表现出高度有效的替代方法。特别是，我们通过采用完全连通的条件随机场（CRF）来捕捉细节的能力[22]。 CRF已被广泛用于语义分割，将多路分类器计算的类别分数与像素和边缘的局部相互作用所捕获的低级信息结合起来[23]，[24]或其他像素[25]。提出了提高复杂度的技术。模拟层次依赖[26]，[27]，[28]和/或段[29]，[30]，[31]，[32]，[33]的高阶依赖关系，我们使用完全连接由[22]提出的成对CRF，它具有高效的计算能力，能够捕捉细节边缘细节，同时也能满足长距离依赖性。该模型在[22]中显示，以改善基于增强的像素级分类器的性能。在这项工作中，我们证明了当与基于DCNN的像素级分类器结合使用时，它可以产生最先进的结果。

图1显示了所提出的DeepLab模型的高级图示。在图像分类任务中训练的深度卷积神经网络（VGG-16 [4]或ResNet-101 [11]）是通过以下方式实现语义分割的任务：（1）将所有完全连接的层转换为卷积层（即完全卷积网络[14]）和（2）通过不正常的卷积层增加特征分辨率，允许我们每8像素计算特征响应而不是原始网络中的每32个像素。然后，我们采用双线性插值将分数图上采样8倍，以达到原始图像分辨率，从而产生完全连接的CRF [22]的输入，该CRF确定了分割结果。

从实用的角度来看，我们的DeepLab系统的三个主要优点是：（1）速度：由于紊乱的卷积，我们的密集DCNN在NVidia Titan X GPU上以8 FPS运行，而完全连接的CRF的平均场推断需要0.5秒在CPU上。 （2）准确性：我们在几个具有挑战性的数据集上获得了最新的结果，包括PASCAL VOC 2012语义分割基准[34]，PASCAL-Context [35]，PASCAL-PersonPart [36]和Cityscapes [37]。 （3）简单性：我们的系统由两个非常完善的模块，DCNN和CRF组成。

我们在本文中介绍的更新后的DeepLab系统与我们原始会议出版物[38]中报告的第一个版本相比有几项改进。我们的新版本可以通过多尺度输入处理[17]，[39]，[40]或建议的ASPP更好地分割多个尺度的对象。我们通过调整最新的ResNet [11]图像分类DCNN构建了DeepLab的残差网络变体，与基于VGG-16的原始模型相比，实现了更好的语义分割性能[4]。最后，我们对多种模型变体进行了更全面的实验评估，并报告了最新结果，不仅是PASCAL VOC 2012基准，还有其他具有挑战性的任务。我们通过扩展Caffe框架实现了所提出的方法[41]。我们在配套网站http://liangchiehchen.com/projects/ DeepLab.html上分享我们的代码和模型。



2.related work

在过去十年中开发的大多数成功的语义分割系统依赖于手工制作的特征与平面分类器相结合，例如Boosting [24]，[42]，Random Forests [43]或支持向量机[44]。通过整合来自背景[45]和结构化预测技术[22]，[26]，[27]，[46]的更丰富的信息，已经取得了实质性的进展，但这些系统的表现一直受到有限的表达能力的影响。特点。在过去几年中，深度学习在图像分类中的突破很快转移到了语义分割任务。由于此任务涉及分段和分类，因此一个核心问题是如何组合这两个任务。

用于语义分割的第一类基于DCNN的系统通常采用自下而上的图像分割级联，然后是基于DCNN的区域分类。例如，[47]，[48]提供的边界框提议和掩蔽区域在[7]和[49]中用作DCNN的输入，以将形状信息合并到分类过程中。同样，[50]的作者依赖于超像素表示。尽管这些方法可以从良好分割所提供的明显边界中获益，但它们也无法从其任何错误中恢复。

第二类工作依赖于使用卷积计算的DCNN特征进行密集图像标记，并将它们与独立获得的分割结合起来。首先，[39]以多种图像分辨率应用DCNN，然后使用分割树来平滑预测结果。最近，[21]提出使用跳过层并将DCNN内的计算中间特征映射连接起来用于像素分类。此外，[51]建议按区域提案汇集中间特征地图。这些工作仍采用与DCNN分类器结果分离的分段算法，因此有可能承担过早的决策。

第三类作品使用DCNN直接提供密集的类别级像素标签，这使得甚至可以完全丢弃分段。 [14]，[52]的无分割方法以完全卷积的方式直接将DCNN应用于整个图像，将DCNN的最后完全连接的层转换为卷积层。为了处理引言中概述的空间定位问题，[14]对中间特征图的分数进行上采样和连接，而[52]通过将粗略结果传播到另一个DCNN来重新确定从粗到细的预测结果。我们的工作建立在这些工作的基础上，并且如引言中所述通过对特征分辨率施加控制来扩展它们，引入多尺度池技术并将密集连接的CRF [22]集成在DCNN之上。我们证明这会导致显着更好的分割结果，特别是沿着对象边界。 DCNN和CRF的组合当然不是新的，但以前的工作只尝试了本地连接的CRF模型。具体而言，[53]使用CRF作为基于DCNN的重新排名系统的提议机制，而[39]将超像素视为本地成对CRF的节点，并使用图形切割进行离散推理。因此，他们的模型受到超像素计算中的错误或忽略的远程依赖性的限制。相反，我们的方法将每个像素视为由DCNN接收一元电位的CRF节点。至关重要的是，我们采用的完全连接的CRF模型[22]中的高斯CRF电位可以捕获长程依赖性，同时该模型适用于快速平均场推断。我们注意到，对于传统的图像分割任务，已经广泛研究了平均场推断[54]，[55]，[56]，但这些旧模型通常仅限于短距离连接。在独立工作中，[57]使用非常相似的密集连接的CRF模型来重新确定材料分类问题的DCNN结果。然而，[57]的DCNN模块仅通过稀疏点监督而不是在每个像素处进行密集监督来训练。

由于这项工作的第一版已公开发表[38]，语义分割领域已经取得了巨大进展。多个小组取得了重大进展，显着提高了PASCAL VOC 2012语义分段基准的标准，反映了基准排行榜1中的高水平活动[17]，[40]，[58]，[59]，[60] ]，[61]，[62]，[63]。有趣的是，大多数表现良好的方法都采用了我们DeepLab系统的一个或两个关键要素：通过完全连接的CRF进行有效的密集特征提取和原始DCNN分数的改进的Atrous卷积。我们在下面概述了一些最重要和最有趣的进展。

最近在一些相关的工作中探索了结构化预测的端到端培训。虽然我们使用CRF作为后处理方法，[40]，[59]，[62]，[64]，[65]成功地进行了DCNN和CRF的联合学习。特别是，[59]，[65]展开CRF平均场推断步骤，将整个系统转换为端到端可训练的前馈网络，而[62]近似密集CRF平均场推断的一次迭代[22]卷积层与可学习的过滤器。 [40]，[66]追求的另一个富有成效的方向是通过DCNN学习CRF的成对术语，以更重的计算为代价显着提高性能。在不同的方向，[63]用更快的域变换模块[67]取代平均场推理中使用的双边滤波模块，提高速度并降低整个系统的内存需求，而[18]，[68]结合起来边缘检测的语义分割。

在许多论文中都进行了较弱的监督，放宽了对整个训练集可用的像素级语义注释的假设[58]，[69]，[70]，[71]，得到的结果明显好于弱 - 监督前DCNN系统，如[72]。在另一个研究领域，[49]，[73]追求实例分割，共同解决对象检测和语义分割。

我们在这里所谓的无味卷积最初是为了在[15]的“算法？trous”方案中对未抽取小波变换的有效计算而开发的。我们将感兴趣的读者引用到[74]以获得小波文献的早期参考。 Atrous卷积也与多速率信号处理中的“高贵身份”密切相关，它建立在输入信号和滤波器采样率的相同相互作用的基础上[75]。 Atrous卷积是我们在[6]中首次使用的术语。同样的操作后来被称为扩张卷积[76]，他们创造的一个术语是由于操作对应于常规卷积与上采样（或在[15]的术语中）扩散的事实。为了在DCNN中进行更密集的特征提取，各种作者使用了相同的操作[3]，[6]，[16]。除了单纯的分辨率增强之外，激烈的卷积使我们能够扩大滤波器的视野范围，以结合更大的背景，我们在[38]中已经证明了它是有益的。 [76]进一步推行了这种方法，他们利用一系列不稳定的卷积层，以更高的速率聚合多尺度环境。这里提出的用于捕获多尺度对象和上下文的浮动空间金字塔池方案也采用具有不同采样率的多个迂回卷积层，然而我们并行地而不是串行地布置。有趣的是，动态卷积技术也被用于更广泛的任务，例如物体检测[12]，[77]，实例级分割[78]，视觉问答[79]和光学流[80]。 。

我们还表明，正如预期的那样，集成到DeepLab中的更先进的图像分类DCNN，例如[11]的残差网络可以产生更好的结果。 [81]也独立观察到了这一点。

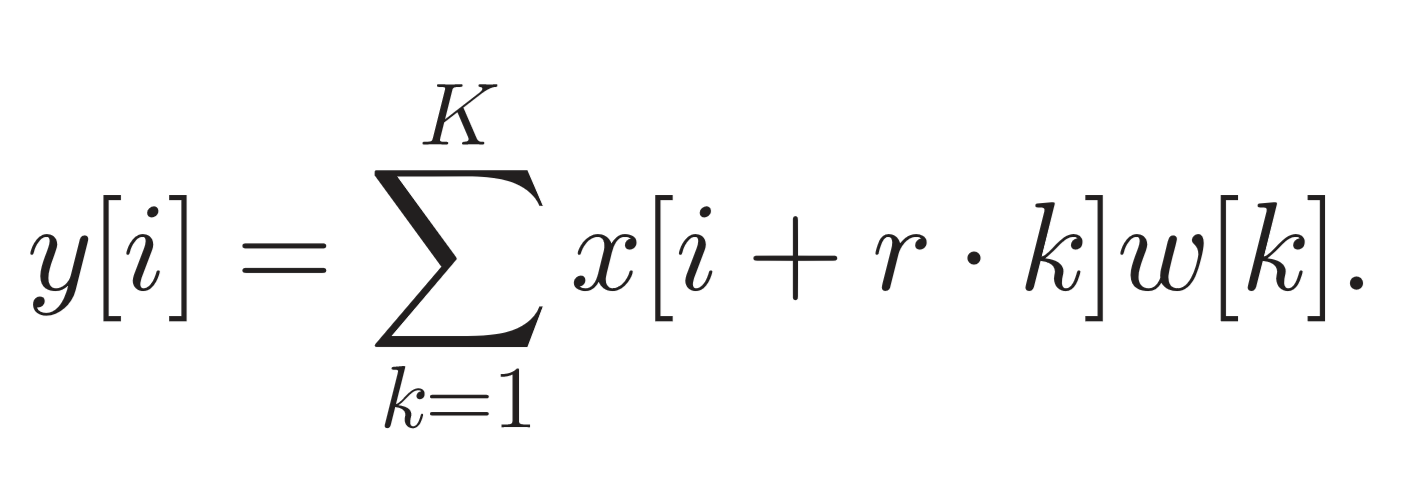
3 METHODS

3.1 Atrous Convolution for Dense Feature Extraction and Field-of-View Enlargement

通过以完全卷积的方式部署DCNN，已经证明DCNN用于语义分段或其他密集预测任务是简单而成功地解决的[3]，[14]。然而，在这些网络的连续层处的最大池和跨步的重复组合显着地降低了所得特征图的空间分辨率，通常在最近的DCNN中的每个方向上的因子为32倍。部分补救措施是使用[14]中的“反卷积”层，但这需要额外的内存和时间。

我们提倡使用最初为[15]的“算法？trous”方案中的未抽取小波变换的有效计算而开发的atrous卷积，并且在DCNN背景中使用[3]，[6]，[ 16。该算法允许我们以任何所需的分辨率计算任何层的响应。一旦网络经过培训，它就可以在事后应用，但也可以与培训无缝集成。

首先考虑一维信号，将具有长度为K的滤波器w [k]的1-D输入信号x [i]的迂回卷积2的输出y [i]定义为



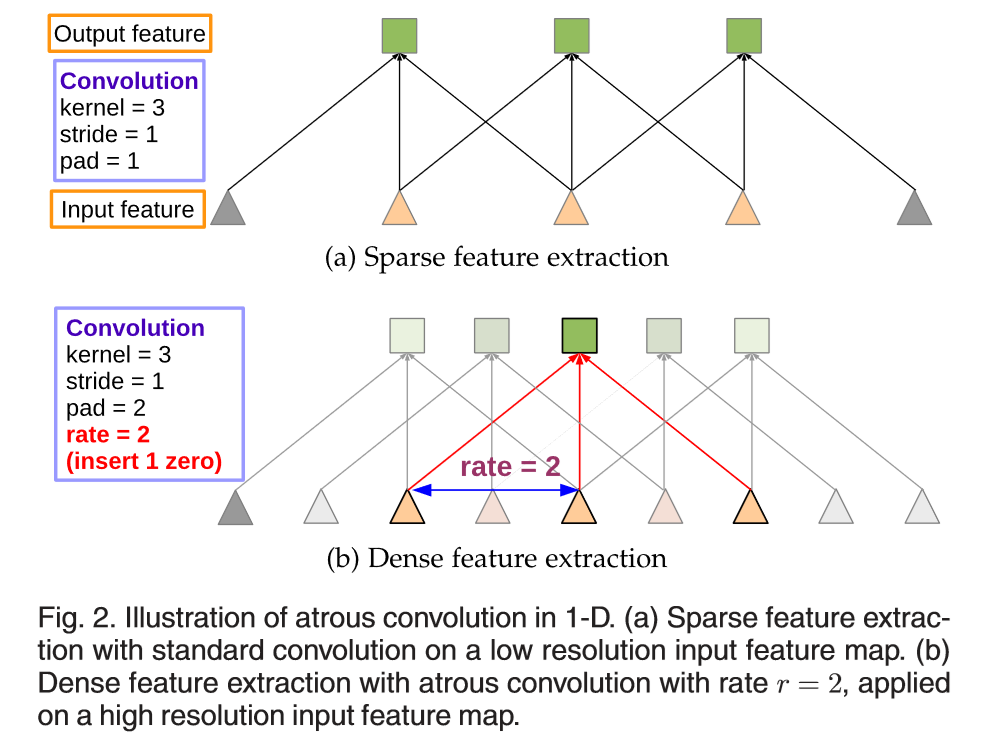
速率参数r对应于我们对输入信号进行采样的步幅。标准卷积是速率r = 1的特殊情况。参见图2以进行说明。

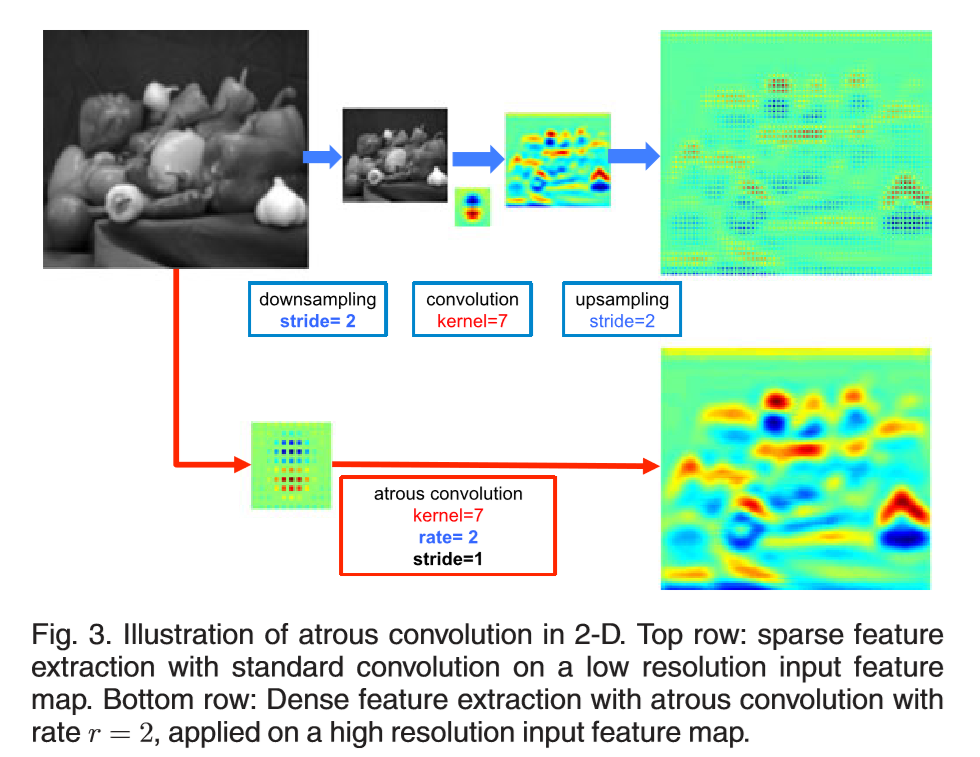
我们通过图3中的一个简单示例来说明算法在2-D中的操作：给定图像，我们假设我们首先进行下采样操作，将分辨率降低2倍，然后使用内核执行卷积 - 这里，垂直高斯导数。如果在原始图像坐标中植入生成的特征图，我们意识到我们仅在图像位置的1/4处获得了响应。相反，我们可以计算所有图像位置的响应，如果我们将全分辨率图像与过滤器'带孔'进行卷积，其中我们将原始滤波器上采样2倍，并在滤波器值之间引入零。虽然有效滤波器尺寸增加，但我们只需要考虑非零滤波器值，因此滤波器参数的数量和每个位置的操作数量保持不变。由此产生的方案使我们能够轻松，明确地控制神经网络特征响应的空间分辨率。

在DCNN的上下文中，可以在层链中使用迂回卷积，有效地允许我们以任意高的分辨率计算最终的DCNN网络响应。例如，为了使VGG-16或ResNet-101网络中计算出的特征响应的空间密度加倍，我们找到了降低分辨率的最后一个汇集或卷积层（分别为“pool5”或“conv5\_1”），设置了它的步幅为了避免信号抽取，并用速率为r = 2的迂回卷积层替换所有后续卷积层。在整个网络中一直推动这种方法可以让我们以原始图像分辨率计算特征响应，但这最终成为太贵了。我们采用了混合方法来实现良好的效率/准确性权衡，使用迂回卷积将计算特征图的密度增加4倍，然后通过额外因子8快速双线性插值来恢复特征图在原始图像分辨率。双线性插值在此设置中是足够的，因为类别得分图（对应于对数概率）非常平滑，如图5所示。与[14]采用的去卷积方法不同，所提出的方法将图像分类网络转换为密集特征提取器无需学习任何额外参数，从而在实践中实现更快的DCNN培训。

Atrous卷积还允许我们在任何DCNN层任意放大滤波器的视野。现有技术的DCNN通常采用空间上小的卷积核（通常为3×3），以便保持计算和包含的参数数量。利率r的剧烈卷积在连续滤波器值之间引入r-1个零，有效地将ak×k滤波器的内核大小扩大到ke = k +（k-1）×（r-1），而不增加参数的数量或者计算。因此，它提供了一种有效的机制来控制视野，并在精确定位（小视场）和上下文同化（大视场）之间找到最佳平衡点。我们已经成功地尝试了这种技术：我们的DeepLab-LargeFOV模型变体[38]采用了萎缩卷积，在VGG-16'fc6'层中速率r = 12，具有显着的性能增益，详见第4节。

转向实现方面，有两种有效的方法来执行痛苦的卷积。首先是通过插入孔（零）来隐式地对过滤器进行上采样，或者等效地稀疏地对输入特征图进行采样[15]。我们在早期的工作[6]，[38]中实现了这一点，然后在[76]中，在Caffe框架[41]中通过添加到im2col函数（它从多通道特征映射中提取矢量化补丁）稀疏地实现了这个选项。对基础要素图进行采样。最初由[82]提出并在[3]，[16]中使用的第二种方法是通过等于迂回卷积率r的因子对输入特征图进行子采样，对其进行解交织以产生r2降低分辨率的映射，每个映射一个r×r可能的变化。然后将标准卷积应用于这些中间特征图并将它们重新交换为原始图像分辨率。通过将atrous卷积减少为常规卷积，它允许我们使用现成的高度优化的卷积例程。我们已经在TensorFlow框架中实现了第二种方法[83]。



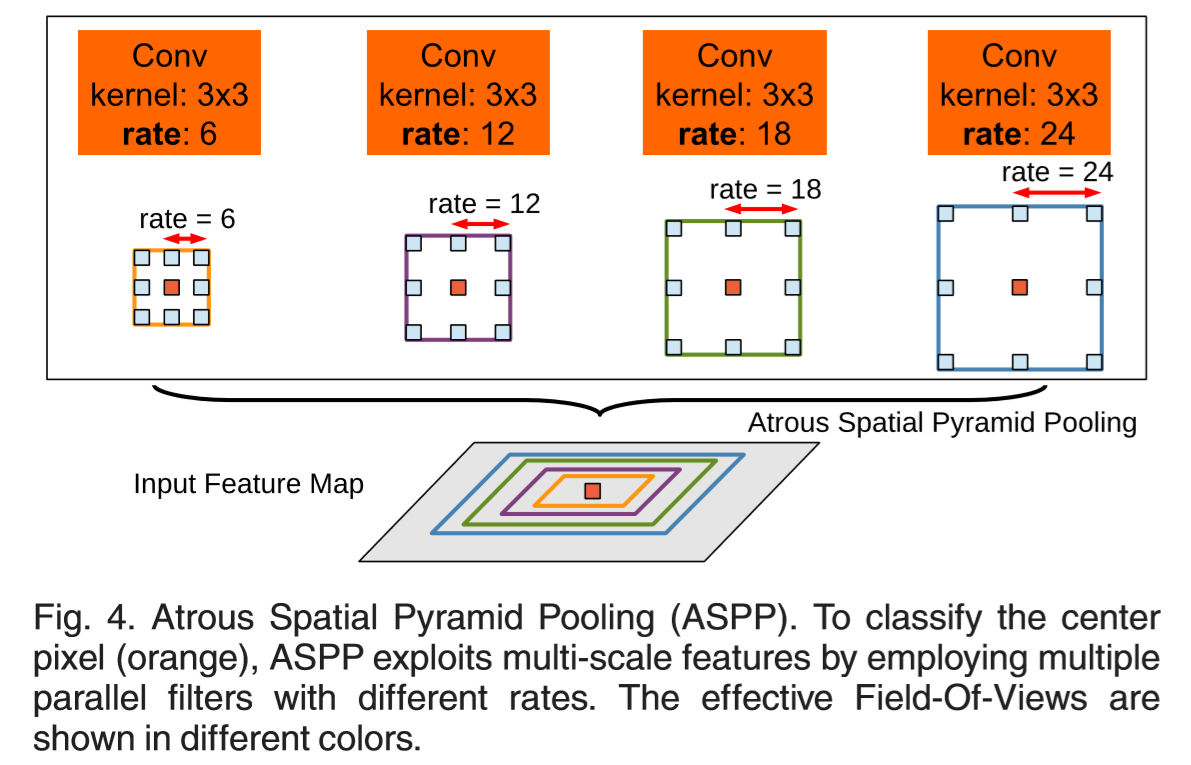


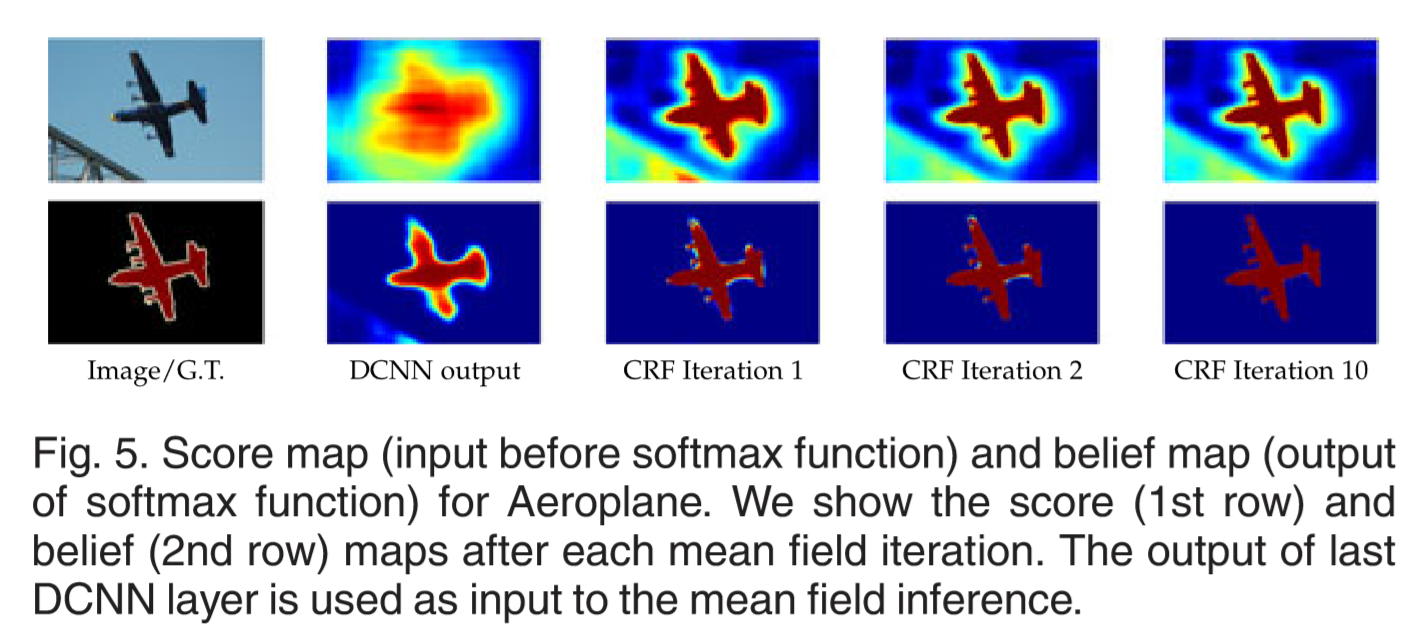
3.2 Multiscale Image Representations Using Atrous Spatial Pyramid Pooling

DCNN已显示出显着的隐式表示比例的能力，只需在包含不同大小的对象的数据集上进行训练即可。尽管如此，明确考虑对象规模可以提高DCNN成功处理大型和小型对象的能力[6]。

我们已经尝试了两种处理语义分割中的尺度可变性的方法。第一种方法相当于标准的多尺度处理[17]，[18]。我们使用共享相同参数的并行DCNN分支从多个（在我们的实验中）三个重新缩放版本的原始图像中提取DCNN得分图。为了产生最终结果，我们将并行DCNN分支的特征映射双线性插值到原始图像分辨率并融合它们，通过在每个位置获取不同尺度上的最大响应。我们在培训和测试期间都这样做。多尺度处理显着提高了性能，但代价是计算所有DCNN层的特征响应，以满足多种输入规模。

第二种方法的灵感来自于[20]的R-CNN空间金字塔合并方法的成功，该方法表明，通过重新采样在单一尺度上提取的卷积特征，可以准确且有效地对任意尺度的区域进行分类。我们已经实现了他们的方案的变体，其使用具有不同采样率的多个并行的atrous卷积层。为每个采样率提取的特征在不同的分支中进一步处理并融合以产生最终结果。提出的“atrous空间金字塔池”（DeepLabASPP）方法概括了我们的DeepLab-LargeFOV变体，如图4所示。





3.3 Structured Prediction with Fully-Connected Conditional Random Fields for Accurate Boundary Recovery

在定位精度和分类性能之间的权衡似乎是DCNN中固有的：具有多个最大池层的更深层模型已经证明在分类任务中最成功，但是增加的不变性和顶级节点的大型接收场只能产生平滑的响应。如图5所示，DCNN得分图可以预测对象的存在和粗略位置，但不能真正描绘其边界。

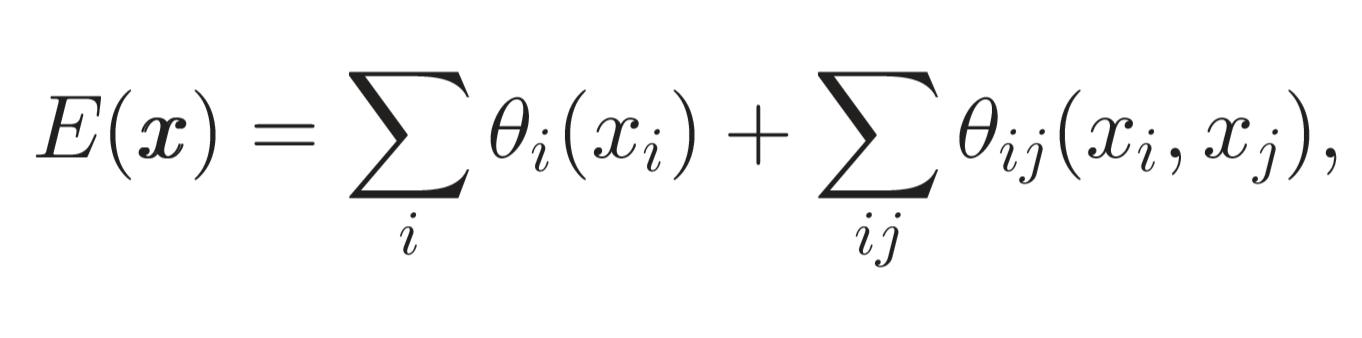
以前的工作追求两个方向来解决这一本地化挑战。第一种方法是利用卷积网络中多层的信息，以便更好地估计物体边界[14]，[21]，[52]。第二种是采用超像素表示，基本上将本地化任务委托给低级分割方法[50]。

我们在DCNN的识别能力与完全连接的CRF的细粒度定位精度的耦合基础上寻求替代方向，并表明它在解决定位挑战方面非常成功，产生了准确的语义分割结果并在一定程度上恢复了对象边界。细节远远超出了现有方法的范围。几个后续文件[17]，[40]，[58]，[59]，[60]，[61]，[62]，[63]，[65]已经扩展了这个方向，因为第一个我们的工作版本已发表[38]。

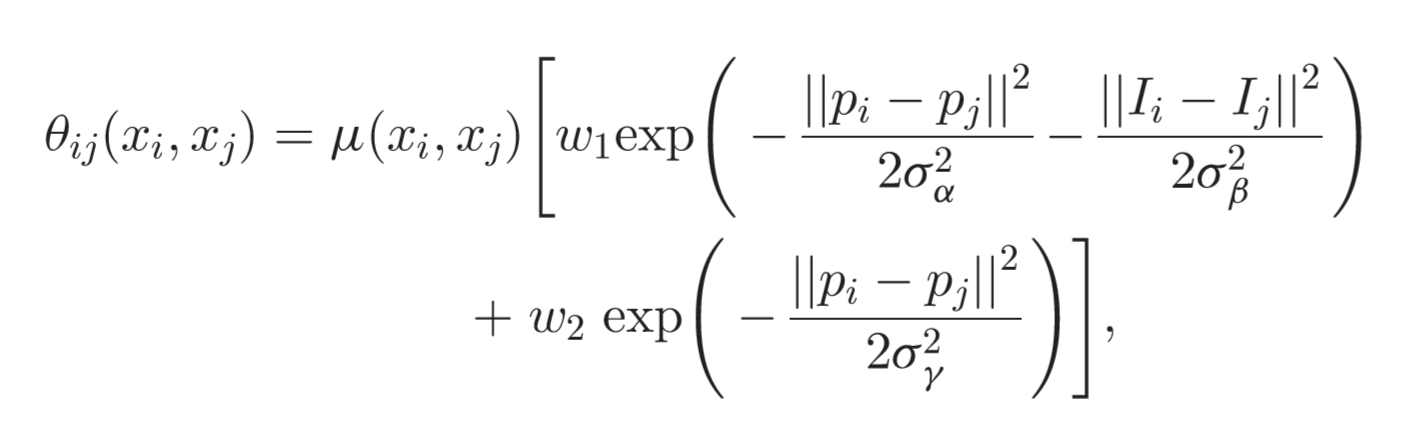
传统上，条件随机场（CRF）已被用于平滑噪声分割图[23]，[31]。通常，这些模型耦合相邻节点，有利于相同标签分配到空间上近端像素。定性地说，这些短程CRF的主要功能是清理基于本地手工设计功能构建的弱分类器的虚假预测。

与这些较弱的分类器相比，现代DCNN架构（例如我们在此工作中使用的架构）产生的得分图和语义标签预测在质量上是不同的。如图5所示，得分图通常非常平滑并产生均匀的分类结果。在这种情况下，使用短程CRF可能是有害的，因为我们的目标应该是恢复详细的局部结构而不是进一步平滑它。将对比敏感电位[23]与局部范围CRF结合使用可以潜在地改善定位，但仍然缺少薄结构，并且通常需要解决昂贵的离散优化问题。

为了克服短距​​离CRF的这些限制，我们将[22]的完全连接的CRF模型集成到我们的系统中。该模型采用能量函数



其中x是像素的标签分配。 我们使用一元势o（xi）= - logP（xi），其中Pðxi是由DCNN计算的像素i处的标签分配概率。 成对电位具有允许在使用完全连接的图形时进行有效推断的形式，即，当连接所有图像像素对时，i; j。 特别是，如[22]中所述，我们使用以下表达式



其中u（xi; xj）= 1，如果xi！= xj，否则为零，如Potts模型中所示，这意味着只有具有不同标签的节点才会受到惩罚。剩下的表达式在不同的特征空间中使用两个高斯核;第一个“双边”内核取决于像素位置（表示为p）和RGB颜色（表示为I），第二个内核仅取决于像素位置。超参数sa，sb和sg控制高斯核的尺度。第一个内核强制具有相似颜色和位置的像素具有相似的标签，而第二个内核仅在强制平滑时考虑空间接近度。

至关重要的是，该模型适用于有效的近似概率推理[22]。在完全可分解的平均场近似bx =累乘bi（xi）下传递更新的消息可以表示为双边空间中的高斯卷积。高维滤波算法[84]显着加速了这种计算，导致算法在实践中非常快，使用[22]的公开实现，Pascal VOC图像平均需要少于0.5秒。

4 EXPERIMENTAL RESULTS

我们按照[14]的程序，通过Imagenet预训练的VGG-16或ResNet-101网络的模型权重，以简单的方式使它们适应语义分割任务。我们用一个分类器替换最后一层中的1000路Imagenet分类器，该分类器具有与我们任务的语义类数量（包括背景，如果适用）一样多的目标。我们的损失函数是CNN输出图中每个空间位置的交叉熵项的总和（与原始图像相比，由8个子采样）。所有位置和标签在整体损失函数中均等加权（忽略的未标记像素除外）。我们的目标是基本事实标签（由8个子样本）。我们通过[2]的标准SGD程序针对所有网络层的权重优化目标函数。我们解耦DCNN和CRF训练阶段，假设在设置CRF参数时固定DCNN一元项。

我们在四个具有挑战性的数据集上评估所提出的模型：PASCAL VOC 2012，PASCAL-Context，PASCALPerson-Part和Cityscapes。我们首先报告了关于PASCAL VOC 2012的会议版本[38]的主要结果，并向前推进了所有数据集的最新结果。

4.1 PASCAL VOC 2012

数据集。 PASCAL VOC 2012分割基准[34]涉及20个前景对象类和一个背景类。 原始数据集包含分别用于训练，验证和测试的1,464（训练），1,449（val）和1,456（测试）像素级标记图像。 数据集由[85]提供的额外注释增强，产生10,582（trainaug）训练图像。 性能是根据21个类别的平均像素交叉（IOU）来衡量的。

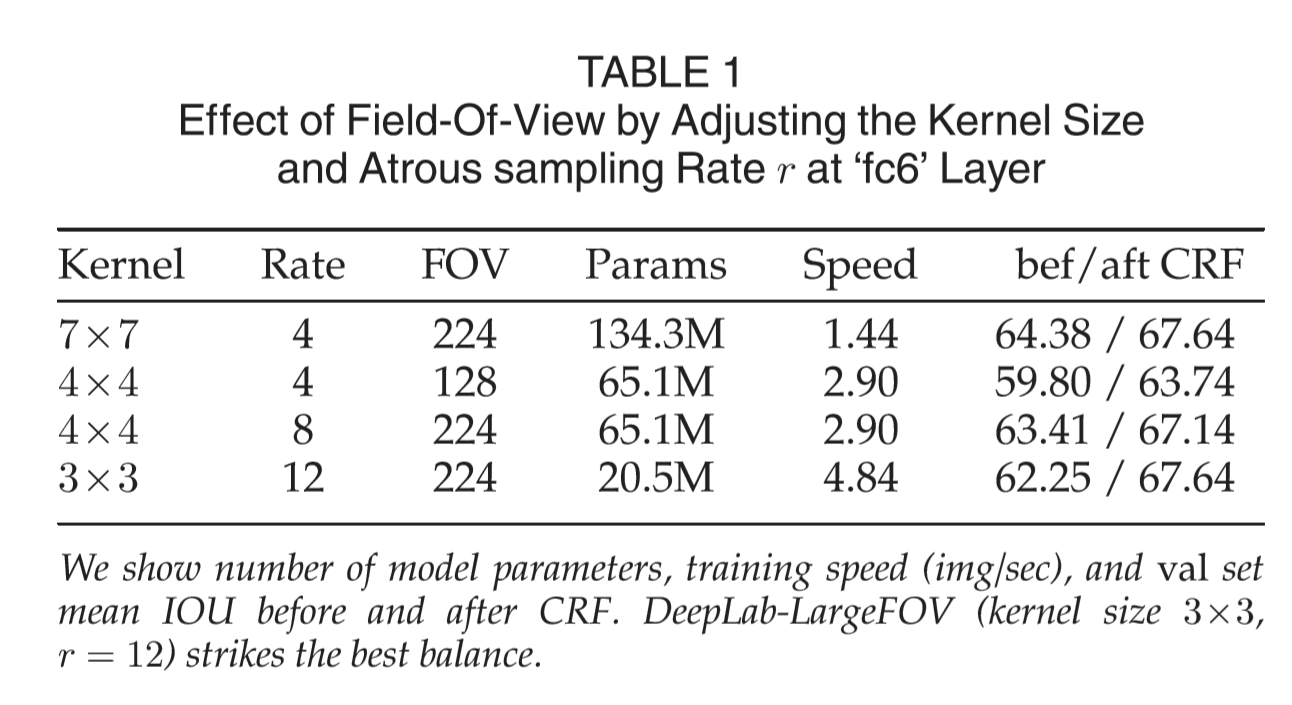
4.1.1 Results from Our Conference Version

我们使用在Imagenet上预训练的VGG-16网络，适用于3.1节中描述的语义分割。我们使用20个图像的小批量和0.001的初始学习率（最终分类层为0.01），每2000次迭代将学习率乘以0.1。我们使用0.9的动量和0.0005的重量衰减。

在DCNN已经在trainaug上进行了调整后，我们沿着[22]的线交叉验证CRF参数。我们使用w2 = 3和sg = 3的默认值，我们通过对来自val的100个图像的交叉验证来搜索w1，sa和sb的最佳值。我们采用粗略搜索方案。参数的初始搜索范围是w1 [3：6] ?, sa [30：10：100]和sb [3：6]（MATLAB表示法），然后我们围绕第一轮的最佳值重新确定搜索步长。我们采用10次平均场迭代。

视野和CRF。在表1中，我们报告使用DeepLab模型变体的实验，这些变体使用不同的视野尺寸，通过调整'fc6'层中的内核大小和不稳定采样率r获得，如3.1节所述。我们首先直接调整VGG16网络，使用原始的7×7内核大小和r = 4（因为我们对最后两个最大池化层没有任何步幅）。该模型在CRF后产生67.64％的性能，但相对较慢（在训练期间每秒1.44幅图像）。通过将内核大小减小到4×4，我们将模型速度提高到每秒2.9个图像。我们已经尝试了两种具有较小（r = 4）和较大（r = 8）FOV大小的网络变体;后者表现更好。最后，我们采用内核大小3×3甚至更大的采样率（r = 12），通过在层'fc6'和'fc7'中保留4,096个滤波器中的1,024个随机子集，使网络更薄。得到的模型DeepLab-CRF-LargeFOV与直接VGG-16自适应（7×7内核大小，r = 4）的性能相匹配。与此同时，DeepLab-LargeFOV的速度提高了3.36倍，参数显着减少（20.5 M而不是134.3 M）。

CRF大大提升了所有型号的性能，平均IOU绝对增加了3-5％。



4.1.2 Improvements after Conference Version of This Work

