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