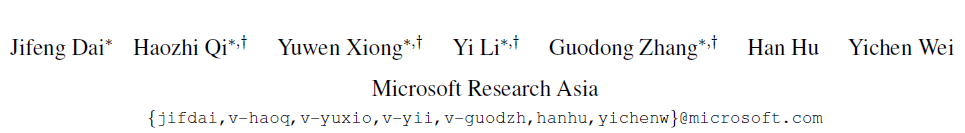
Deformable Convolutional Networks

可变形的卷积网络



摘要

卷积神经网络（CNN）由于其扩展模块中的固定几何结构而固有地局限于模型几何变换。在这项工作中，我们引入了两个新模块来增强CNN的转换建模能力，即可变形卷积和可变形RoI池。两者都是基于增加模块中的空间采样位置的想法，具有额外的偏移量，并且从目标任务中学习偏移量，而无需额外的监督。新的模块可以很容易地替代现有CNN中的普通对等体，并且可以通过标准反向传播轻松地进行端到端的训练，从而产生可变形的卷积网络。广泛的实验验证了我们的方法的性能。我们首次展示，深入CNN中学习密集空间变换对于复杂的视觉任务（如对象检测和语义分割）是有效的。该代码在https://github.com/msracver/Deformable-ConvNets发布。

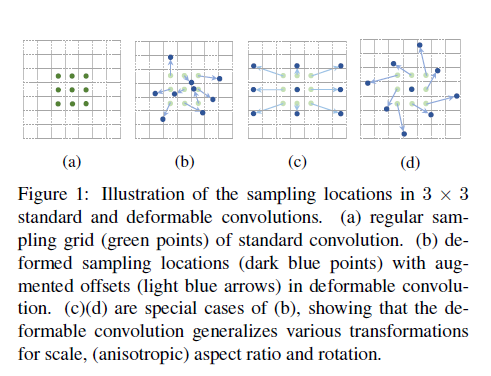
1 简介

视觉识别中的一个关键挑战是如何适应对象尺度，姿态，视点和部件变形中的几何变化或模型几何变换。 一般来说，有两种方法。 第一个是建立具有足够的期望变化的训练数据集。 这通常通过增加现有数据样本来实现，例如通过仿射变换。 可以从数据中获取可靠的表示，但通常是以昂贵的训练和复杂模型参数为代价。 二是使用变换不变特征和算法。 该类别包含许多众所周知的技术，例如SIFT（尺度不变特征变换）[42]和基于对象检测范例的滑动窗口。

上面的方法有两个缺陷。首先，假定几何变换是固定的和已知的。这种使用先验知识增加数据，并设计特征和算法。 这个假设阻止了对具有未知几何变换的新任务的泛化，这些变换没有被适当的建模。 第二，即使知道，过分复杂的变换，不变特征和算法的手工设计可能也很困难或不可行。

近来，卷积神经网络（CNN）[35]在图像分类[31]，语义分割[41]和对象检测[16]等视觉识别任务中取得了显着成功。 然而，他们仍然有两个缺点。 他们对几何变换建模的能力主要来自广泛的数据增加，大型模型容量和一些简单的手工模块（例如，用于小型平移不变性的最大池[1]）。

简而言之，CNNs本质上限于模拟大的未知转换。限制源自CNN模块的固定几何结构：卷积单元在固定位置对输入特征图进行采样;池化层以固定比例降低空间分辨率; RoI（感兴趣区）池化层将RoI分为固定空间箱等。它们缺少处理几何变换的内部机制。这会引起明显的问题。例如，相同CNN层中的所有激活单元的感受野大小相同。这对于通过空间位置编码语义的高级CNN层是不可取的。因为不同的位置可能对应于具有不同尺度或变形的物体，因此对于具有精细定位的视觉识别（例如，使用完全卷积网络的语义分割）需要自适应确定尺度或感受野大小[41]。对于另一个例子，尽管目标检测已经取得了显着和快速的进展[16,52,15,47,46,40,7]最近，所有的方法仍然依赖于基于原始边界框的特征提取。这显然是次优的，特别是对于非刚性物体。



在这项工作中，我们引入了两个新模块，大大增强了CNN的几何变换建模能力。第一个是可变形卷积。 它在标准卷积中将2D偏移添加到常规网格采样位置。 它使得采样网格可以自由形状变形。 它在图1中示出。通过附加卷积层从前面的特征图学习偏移量。 因此，变形以局部，密集和自适应的方式对输入特征进行调节。

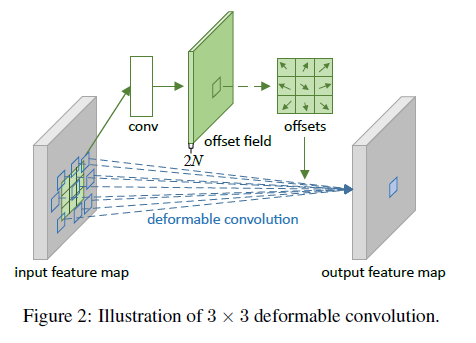
第二个是可变形的RoI池。它为先前RoI池[15,7]的常规箱分区中的每个箱添加一个偏移量。类似地，从先前的特征图和RoIs学习偏移，使得具有不同形状的对象的自适应部分定位成为可能。

两个模块都是轻量的。它们为偏移学习添加了少量参数和计算。它们可以很容易地替代它们在深度卷积网络中的普通对等体，并且可以通过标准反向传播轻松地进行端到端的训练。所得到的CNN称为可变形卷积网络，或可变形ConvNets。

我们的方法与空间变换网络[26]和可变部件模型[11]具有类似的高级精髓。它们都具有内部变换参数，纯粹从数据中学习这些参数。与可变形ConvNets的一个关键区别是它们以简单，高效，深入和端对端的方式处理密集的空间变换。在3.1节中，我们详细讨论了我们的工作与以前的作品的关系，并分析了可变形ConvNets的优越性。

2 可变形的卷积网络

CNN中的特征映射和卷积是3D。 可变形卷积和RoI池模块都在2D空间域上运行。 通道维度的操作保持不变。 在不损失一般性的情况下，这些模块在2D中描述可以表示清晰。扩展到3D是直接的。



2.1 可变卷积

2D卷积由两个步骤组成：1）使用输入特征图x上的规则网格R进行采样; 2）由w加权的采样值的求和。 网格R定义了感受野的大小和扩张。 例如，

R = {(-1,-1),(-1,0),… ,(0,1),(1,1)}

定义了一个3\*3膨胀值为1的核

对于输出特征图y的每一个位置p0，我们有：

, (1)

这里的pn枚举了R中的位置

在可变形的卷积中，规则网格R以偏移量{}增加,这里N=|R|. 方程（1）变为：

. (2)

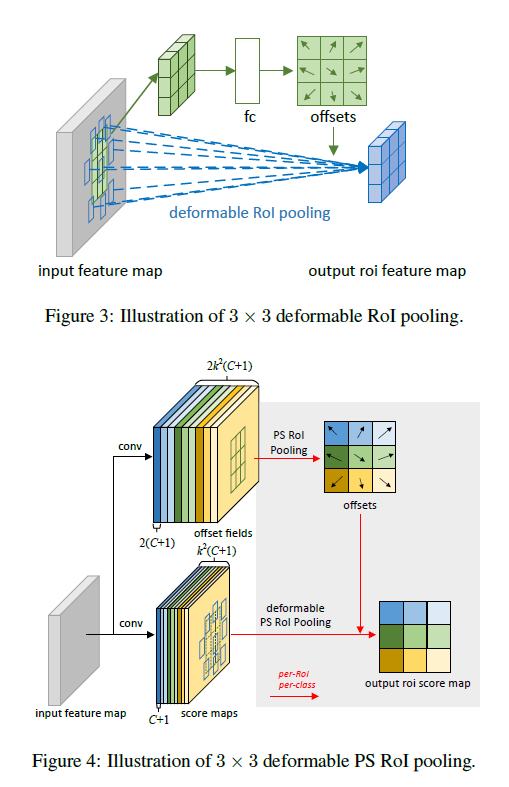
现在，在不规则和偏移位置,由于偏移通常是极小的，方程（2）通过双向线性插值实现为：

, (3)

这里p表示一个任意的（极小的）位置（），q枚举所有特征图x中可积分的空间位置，此外G（.,.）是双向线性插值核。注意G是二维的。它被分为两个一维核如下：

G(q,p) = g()\*g(), (4)

这里G(a,b)=max(0,1-|a –b|),方程式（3）快速被计算因为G（q，p）仅对于少数的q是非零的。



如图2所示，通过在相同的输入特征图上应用卷积层来获得偏移。 卷积核具有与当前卷积层相同的空间分辨率和膨胀（例如，图2中具有膨胀1的3\*3）。 输出偏移字段与输入特征图具有相同的空间分辨率。 通道维度2N对应于N 2D偏移。 在训练期间，同时学习用于产生输出特征和偏移量的卷积内核。 为了学习偏移量，通过方程式（3）和方程式（4）中的双向线性运算来后向传播梯度。 详见附录A.

2.2 可变形的RoI池化

RoI pooling用于所有基于区域推荐的对象检测方法[16,15,47,7]。 它将任意大小的输入矩形区域转换为固定大小的特征。

RoI pooling 指定输入特征图x和一个大小为w\*h的RoI以及左上角为，RoI pooling将RoI分为一个k\*k（k是一个自由参数）箱子并且输出一个k\*k的特征图y，对于第（i，j）个箱子（0）,我们有：

, (5)

这里是箱子中像素数，第（i，j）个箱子跨度在和

类似于方程（2）中，在可变形RoI池化层，偏移量被添加到空间分箱位置，方程（5）变为：

. (6)

通常，是极小的，方程式（6）通过方程式（3）和方程式（4）的

双向线性插值来实现。

图3展示了如何获得偏移量，首先，RoI池化层生成池化特征映射，从映射中,一个全连接层生成标准的偏移,然后通过方程（6）中逐个元素与RoI的宽与高相乘转化为偏移量，因为这里是一个预定义的标量来调整偏移量。 经验设定为r = 0.1。 偏移归一化是使偏移学习必须相对于RoI大小保持不变。 fc层通过反向传播学习，详见附录A。

位置敏感（PS）RoI池 [7]它是完全卷积的，不同于RoI池。通过一个卷积层，所有的输入特征图首先被转换为每个对象类的k2分数图（C对象总共C + 1类），如图4中的底部分支所示。无需区分类，这样的分数映射表示为{}; 其中（i,j）枚举所有分箱。池化在这些分数图上执行。 对于第（i，j）个箱子的输出值由与该箱子相对应的一个分数图求和来获得。简而言之，与方程（5）中的RoI池化的区别在于一般特征图x被特定的正敏感分数图代替。

在可变形PS RoI池中，方程式(6)的唯一变化是x也被修改为。 然而，偏移学习是不同的。 它遵循[7]中的“完全卷积”思想，如图4所示。在顶部分支中，卷积层生成完整的空间分辨率偏移区域。 对于每个RoI（也对于每个类），PS RoI池化被应用于这样的区域以获得标准化偏移量，然后将它们以与上述可变形RoI池相同的方式转换为实际偏移量。

2.3 可变形卷积网络

可变卷积和RoI池模块都具有与其简单版本相同的输入和输出。因此，他们可以很容易地替代现有CNN中的普通部分。 在训练中，这些添加的用于偏移学习的conv和fc层被初始化为零权重。 他们的学习率设置为现有层学习速率的（默认情况下=1，而faster R-CNN中的fc层）。 他们通过反向传播训练由方程式（3）和方程式（4）中的双线性插值运算。所得到的CNN称为可变形卷积网络。

为了将可变形ConvNets与现有的CNN架构相结合，我们注意到这些架构分为两个阶段。首先，深度完全卷积网络在整个输入图像上生成特征图。第二，浅层任务特定网络从特征图生成结果。我们详细说明了以下两个步骤。

特征提取的变形卷积 我们采用两种最先进的特征提取架构：ResNet-101 [22]和Inception-ResNet [51]的修改版本。两者都在ImageNet [8]分类数据集上进行了预训练。

原始的Inception-ResNet专为图像识别而设计。它具有特征不对齐问题，并且对于密集预测任务是有问题的。它被修改修复对齐问题[20]。修改版本被称为“对齐 - 初始 - ResNet”，详见附录B.

这两个模型由几个卷积块组成，一个平均池和一个1000路图像的ImageNet分类。平均池和fc层被去除。一个随机初始化的1\*1卷积被添加，最后将通道尺寸减小到1024.如常见的做法[4,7]，最后卷积块中的有效步长从32像素减少到16像素，以增加特征映射分辨率。具体来说，在最后一个块的开头，stride从2变为1（ResNet-101和Aligned-Inception-ResNet都为“conv5”）。为了补偿，该块中的所有卷积滤波器（内核大小> 1）的扩展从1变为2。

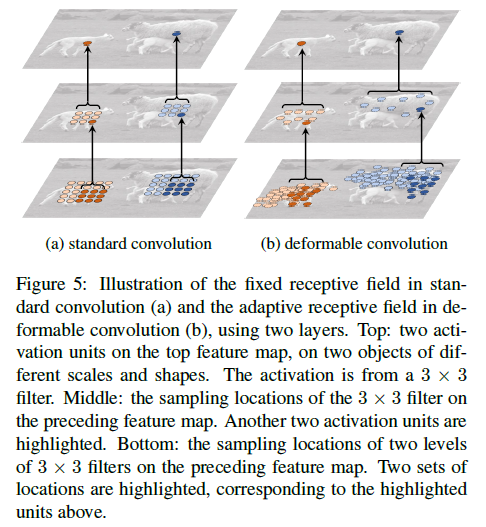
可选地，可变形卷积被施加到最后几个卷积层（核大小> 1）。我们尝试了不同数量的这样的层，发现3是不同任务的良好权衡，如表1所示。

分割和检测网络 任务专用网络建立在上述特征提取网络的输出特征图上。

在下面，C表示对象类的数量。

DeepLab [5]是语义分割的最先进的方法。它增加了1\*1卷积层上的特征映射以生成（C + 1）映射，表示每像素分类得分。随后的softmax层输出每像素概率。

类别感知RPN与[47]中的区域提案网络几乎相同，只是将2类（对象或非对象）卷积分类器替换为（C + 1）类卷积分类器。



它可以被认为是SSD的简化版本[40]。

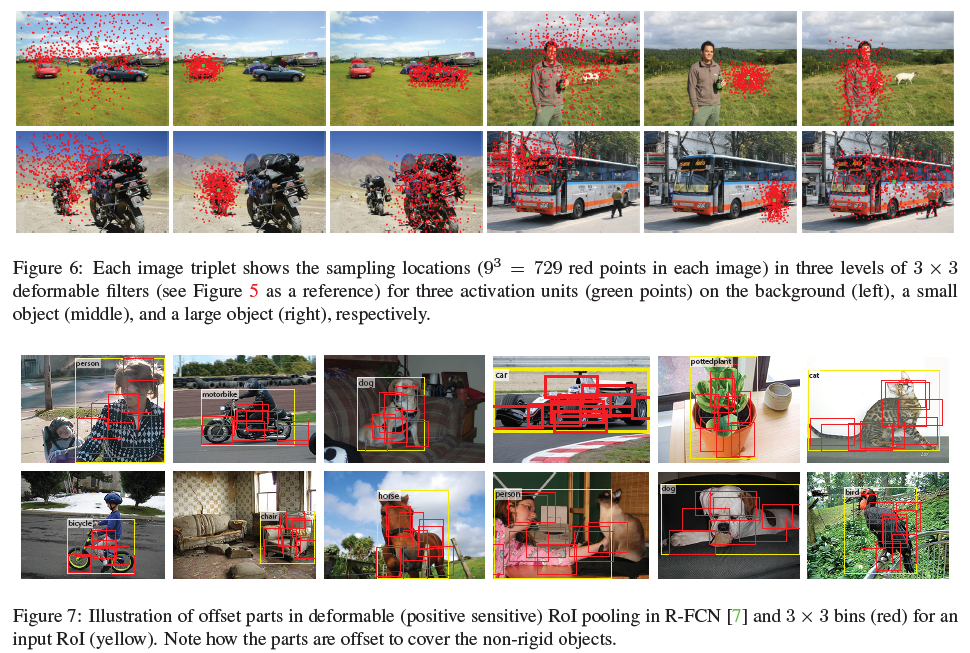
更快的R-CNN [47]是最先进的检测器。在我们的实现中，RPN分支被添加在conv4块的顶部，遵循[47]。在以前的做法[22,24]中，RoI池化层被插入到ResNet-101中的conv4和conv5块之间，为每个RoI留下10层。该设计实现了良好的精度，但是具有高的每个RoI计算。相反，我们采用如[38]中的简化设计。最后添加了RoI池化层1。在池化的RoI特征之上，添加了维度为1024的两个fc层，接着是边界框回归和分类分支。虽然这样的简化（从10层conv5块到2fc层）会略微降低准确度，但是它仍然是一个足够强大的基线，在这项工作中并不是一个问题。

可选地，可以将RoI池化层改为可变形的RoI池。

R-FCN [7]是另一种最先进的检测器。它具有可忽略的每个RoI计算成本。我们遵循原来的实现。可选地，其RoI池化层可以改变为可变形的位置敏感的RoI池化。

3 理解可变形卷积网络

这项工作是建立在在卷积中增加空间采样位置和带有额外偏移量的RoI池化的想法之上的，并且从目标任务中学习偏移量。



当可变形卷积堆叠时，复合变形的影响是深刻的。 这在图5中有示例。标准卷积中的感受野和采样位置在顶部特征图上是固定的（左）。 它们根据物体在可变形卷积中的尺度和形状进行自适应调整（右）。 更多的例子如图6所示。表2提供了这种自适应变形的定量证据。

可变形RoI池的效果是相似的，如图7所示。标准RoI池中网格结构的规则性不再成立。 相反，零件偏离RoI箱移动到附近的物体前景区域。 增强了定位能力，特别是对于非刚性物体。

3.1 相关作品背景

我们的工作与以往不同方面的工作有关，我们详细讨论了它们之间的关系和差异。

空间变换网络（STN）[26] 这是在深度学习框架中从数据学习空间转换的第一个工作。 它通过全局参数变换（如仿射变换）来扭曲特征图。

这种扭曲是昂贵的，并且学习变换参数是困难的。 STN在小规模图像分类问题上取得了成功。逆STN方法[37]通过有效的变换参数传播来代替昂贵的特征扭曲。

可变卷积的偏移学习可以被认为是STN中极轻量级的空间变换器[26]。然而，可变形卷积不采用全局参数变换和特征扭曲。相反，它以局部和密集的方式对特征图进行采样。为了生成新的特征图，它具有加权求和步骤，STN中不存在。

可变形卷积易于集成到任何CNN架构中。它的训练很容易。显示出对需要密集（例如语义分割）或半密集（例如，对象检测）预测的复杂视觉任务是有效的。这些任务对于STN来说是困难的（如果可行的话）[26,37]。

主动卷积[27] 这项工作是当代的。它还通过偏移来增加卷积中的采样位置，并通过端对端反向传播来学习偏移量。在图像分类任务中显示有效。

与可变形卷积的两个关键差异使得这种工作不那么普遍和适应。 首先，它在不同的空间位置上共享偏移量。 其次，偏移是每个任务或每个训练学习的静态模型参数。 相比之下，可变形卷积中的偏移量是每个图像位置变化的动态模型输出。 他们对图像中的密集空间变换进行建模，并对（半）密集预测任务（如对象检测和语义分割）有效。

有效的感受野[43] 研究发现，感受野中的所有像素并不是同等地对输出响应做出贡献。中心附近的像素具有更大的影响。有效感受野仅占据理论感受野的一小部分，并具有高斯分布。虽然理论感受野尺寸随着卷积层数的增加呈线性增加，但令人惊讶的结果是，有效感受野尺寸随着数字的平方根而线性增加，因此它以比我们将要预期的速度慢得多的方式增加。

这个发现表明，即使顶层的深层CNN中的单元可能没有足够大的感受野。这部分解释了为什么atrous卷积[23]被广泛应用于视觉任务（见下文）。它表明了适应性感受野学习的必要性。

可变形卷积能够自适应地学习感受野，如图5，图6和表2所示。

Atrous卷积[23]将正常滤波器的步幅增加到大于1，并将原始权重保持在稀疏的采样位置。这增加了感受野大小，并保持了相同的参数和计算复杂度。它被广泛用于语义分割[41,5,54]（也称为[54]中的扩张卷积），对象检测[7]和图像分类[55]。

可变形卷积是非线性卷积的泛化，如图1（c）所示。与曲线卷积的广泛比较见表3。

可变形部件模型（DPM）[11]可变形RoI池类似于DPM，因为两种方法都可以学习对象部件的空间变形以最大化分类分数。由于不考虑部件之间的空间关系，因此可变形RoI池化更简单。

DPM是一个浅层模型，模型变形能力有限。虽然其推理算法可以通过将距离变换视为特殊的池化操作来转换为CNN [17]，但是其训练不是端到端的，并且涉及启发式选择，例如组件的选择和部件大小。相比之下，可变形的ConvNets是深度的，并进行端到端的训练。当多个可变形模块堆叠时，建模变形的能力变得更强。

DeepID-Net 它引入了一个可变形约束池化层，也考虑了物体检测的部件变形。 因此，它具有与可变形的RoI池化类似的精髓，但复杂得多。这项工作是高度工程化的和基于RCNN [16]。目前还不清楚如何以端对端的方式让它适应最近的先进的对象检测方法 [47,7]。

在RoI池中的空间操纵 空间金字塔池[34]在尺度上使用手工制作的池区域，它是计算机视觉中的主要方法，也用于基于深度学习的对象检测[21,15]。

学习池化区域的空间布局取得较少研究。 在[28]中的工作从一个大的完整集合中学习了池化区域的一个稀疏子集。大集合是手工设计的，也不是端对端的学习。

可变形的RoI池是第一个在CNN中端到端学习池化区域的。虽然目前这些区域的大小相同，但是在空间金字塔池中扩展到多个大小[34]是直接的。

转换不变特征及其学习 设计变换不变特征已经有了巨大的努力。值得注意的例子包括尺度不变特征变换（SIFT）[42]和ORB [49]（O

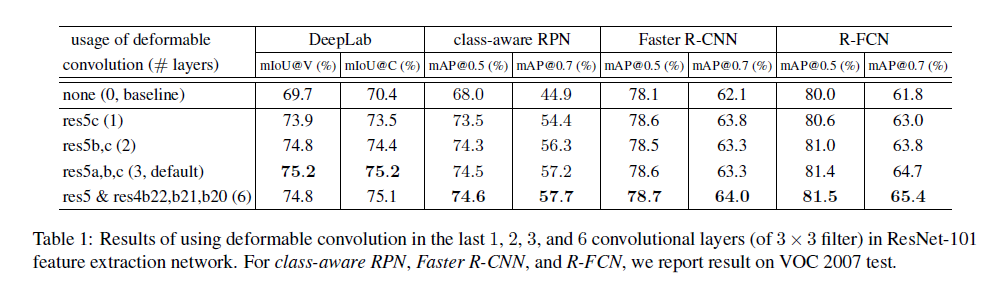
为方向）。在CNN的背景下有大量这样的作品。 [36]研究了CNN表示对图像变换的不变性和等价性。有些作品针对不同类型的变换，例如[50]，散射网络[3]，卷积丛林[32]和TIpooling [33]，学习不变的CNN表示。一些作品专门用于具体的转换，如对称性[13,9]，尺度[29]和旋转[53]。

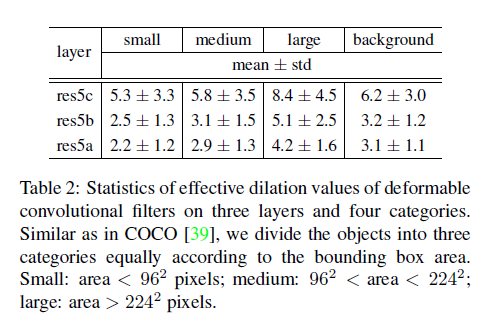
如第1节所述，在这些作品中，这些变换是先验已知的。知识（如参数化）用于手工制作特征提取算法的结构，或者固定在SIFT中，或者使用诸如基于CNN的可学习参数。他们无法处理新任务中的未知转换。

相反，我们的可变形模块总结了各种变换（见图1）。从目标任务中学习转换不变性。

动态滤波器[2] 与可变形卷积类似，动态滤波器也适用于输入特征和样本变化。不同的是，只有过滤器的权重被学习，而不是像我们这样的抽样位置。这项工作适用于视频和立体声预测。

低电平滤波器的组合 高斯滤波器及其平滑导数[30]被广泛用于提取低级图像结构，如角，边缘，T形结等。在某些条件下，这种滤波器形成一组基础，它们的线性组合在同一组几何变换中形成新的过滤器，例如可转向过滤器[12]中的多个方向和[45]中的多个尺度。我们注意到，虽然在[45]中使用了可变形核这个术语，但它的含义与我们在这项工作中是不同的。





大多数CNN从头开始学习所有的卷积过滤器。 最近的工作[25]表明这可能是不必要的。 它通过低级滤波器（高斯导数高达4阶）的加权组合来代替自由形式滤波器，并学习权重系数。 过滤器函数空间的正则化表明当训练数据较小时的提升了泛化能力。

以上工作与我们的相关，因为当多个滤波器，特别是具有不同尺度的滤波器被组合时，所得到的滤波器可以具有复杂的权重并且类似于我们的可变形卷积滤波器。 然而，可变形卷积学习采样位置而不是过滤器权重。

4 实验

4.1 实现和实现

语义分割我们使用PASCAL VOC [10]和CityScapes [6]。 对于PASCAL VOC，有20个语义类别。 遵循[19,41,4]中的协议，我们在[18]中使用VOC 2012数据集和附加掩码注释。 训练集包括10582张图像。 在验证集中对1449个图像进行评估。 对于CityScapes，按照[5]中的协议，对训练集中的2975个图像和验证集中的500个图像分别进行训练和评估。 有19个语义类别加上背景类别。

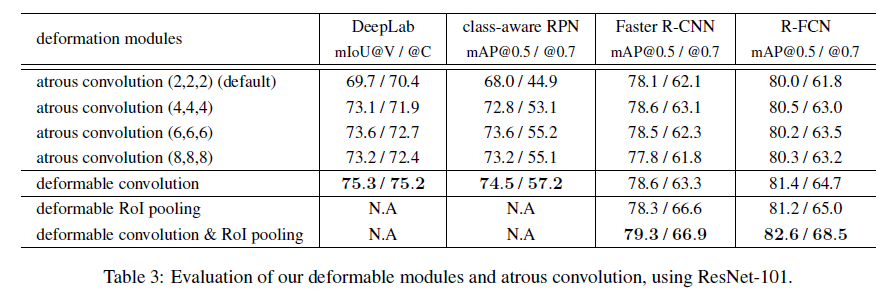
对于评估，我们使用按照标准协议[10,6]在图像像素上定义的mean intersection-over-union（mIoU）度量。我们分别为PASCA1 VOC和Cityscapes使用mIoU @ V和mIoU @ C。

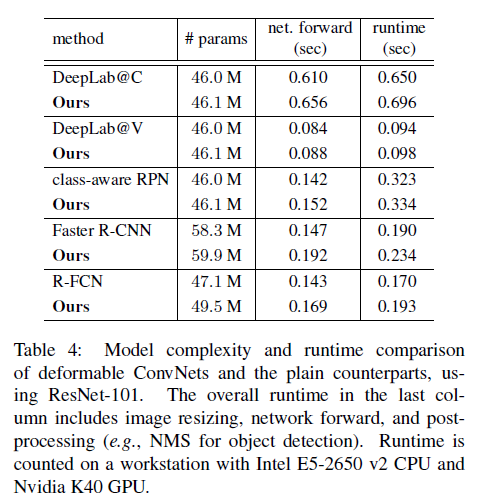
在训练和推理中，图像被调整为具有360像素的较短边，用于PASCAL VOC和1024像素的Cityscapes。在SGD训练中，每个mini-batch随机抽取一张图像。分别为PASCAL VOC和Cityscapes执行了30k和45k的迭代，使用8个GPU和每一个mini-batch。在前2/3次和最后1/3次迭代中，学习率分别为10-3和10-4。

对象检测我们使用PASCAL VOC和COCO [39]数据集。对于PASCAL VOC，按照[15]中的协议，训练是在 VOC 2007 trainval和VOC 2012 trainval进行并集上进行。评估是在VOC 2007上测试。对于COCO，按照标准协议[39]，对trainval中120k图片和test-dev中20k图片分别进行训练和评估。

对于评估，我们使用标准平均精度（mAP）得分[10,39]。对于PASCAL VOC，我们使用IoU阈值在0.5和0.7报告mAP评分。对于COCO，我们使用mAP @ [0.5：0.95]的标准COCO度量以及mAP@0.5。

在训练和推理中，图像被调整为具有600像素的较短边。在SGD训练中，每个小批量随机抽取一张图像。对于类感知RPN，从图像中抽取256个RoI。对于faster RCNN和R-FCN，256和128个RoI分别针对区域候选和对象检测网络进行采样。在RoI池中采用7\*7的箱。为了促进VOC的剥离实验，我们遵循[38]并利用预训练和固定的RPN候选来训练faster R-CNN和R-FCN，而不是区域候选和对象检测网络之间的特征共享。 RPN网络按照[47]的程序第一阶段进行了分开训练。对于COCO，执行[48]中的联合训练，并在训练中进行特征共享。在8个GPU上分别为PASCAL VOC和COCO执行了30k和240k次的迭代。学习率分别在前2/3次和最后1/3次迭代中设为10-3和10-4。





4.2消融实验

进行广泛的消融研究以验证我们的方法的功效和效率。

可变卷积 表1使用ResNet-101特征提取网络评估可变形卷积的效果。当使用更多的可变形卷积层时，精度稳步提高，特别是对于DeepLab和类感知RPN。当DeepLab使用3个可变形层时，其他 6个改进饱和。在剩余的实验中，我们在特征提取网络中使用3。

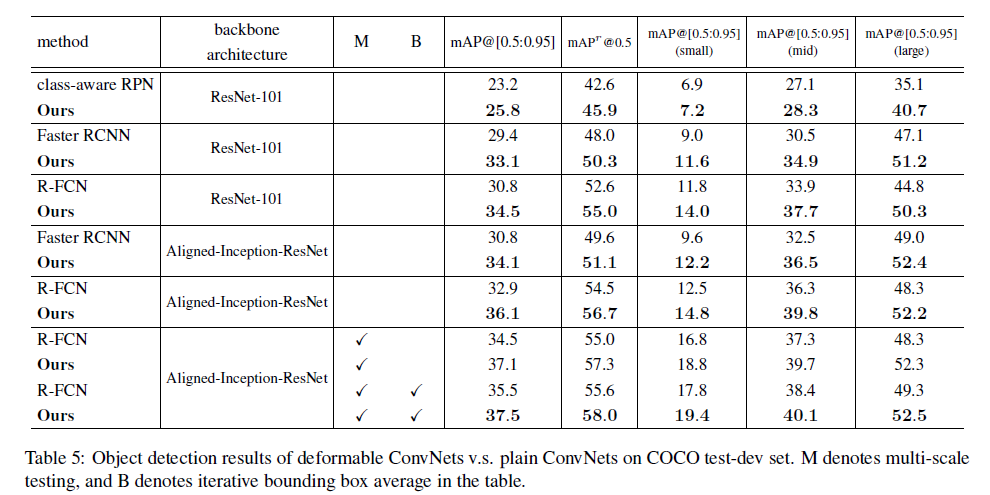
我们经验地观察到，可变形卷积层中的学习偏移量对于图像内容具有高度自适应性，如图5和图6所示。为了更好地理解可变形卷积的机制，我们定义了一种称为可变形卷积滤波器的有效扩张的度量。这是滤波器中所有相邻取样位置对之间距离的平均值。这是过滤器的感受野尺寸的粗略测量。

我们在VOC 2007测试图像上应用具有3个可变形层（如表1所示）的R-FCN网络。根据地面真实边界框注释和过滤器中心位置，我们将可变形卷积滤波器分为四类：小，中，大和背景。表2报告了有效扩张值的统计（平均值和标准差）。清楚地表明：1）可变形滤波器的感受野尺寸与物体尺寸相关，表明从图像内容中有效地学习变形; 2）背景区域上的滤波器尺寸在中等和大对象之间的尺寸之间，表明需要相对较大的感受野来识别背景区域。这些观察结果在不同层次上是一致的。

默认的ResNet-101模型在最后三个3中使用了扩张2的非线性卷积。 3个卷积层（见2.3节）。我们进一步尝试了扩张值4,6和8，并将结果报告在表3中。它显示：1）使用较大扩张值时所有任务的精度均有所提高，表明默认网络的接受范围太小; 2）最佳扩张值对于不同的任务而言是不同的，例如对于DeepLab为6，而对于faster R-CNN则为4。 3）可变形卷积具有最佳精度。这些观察结果验证了滤波器变形的自适应学习是有效的和必要的。

可变形RoI池 它适用于faster RCNN和R-FCN。如表3所示，仅使用它已经产生明显的性能提升，特别是在严格的mAP@0.7度量下。当使用可变形卷积和RoI Pooling两者时，可以获得显着的精度提高。

模型复杂度和运行时间 表4报告了所提出的可变形ConvNets及其简单版本的模型复杂性和运行时间。可变形ConvNets只增加了模型参数和计算的小开销。这表明，除了增加模型参数之外，显著的性能改进来自于几何变换建模的能力。



4.3 在COCO上的对象检测

在表5中，我们对可变形ConvNets和普通ConvNets在COCO测试开发集上的对象检测进行了广泛的比较。我们首先使用ResNet-101模型进行实验。类别感知RPN，faster R-CNN和R-FCN的可变形版本在mAP @ [0.5：0.95]上分别达到25.8％，33.1％和34.5％，它们分别相对高于它们的普通ConvNets值的11％，13％和12％。通过在faster R-CNN和RFCN中通过将Aligned-Inception-ResNet替换ResNet-101，它们的普通ConvNet基线由于更强大的特征表示而得到改进。而可变形ConvNets带来的有效的性能提升也是如此。通过对多个图像尺度的进一步测试（图像短边在[480,576,688,864,1200,1400]），并执行迭代边框平均[14]，在mAP @ [0.5：0.95]得分相对于R-FCN的可变形版本增加到37.5％。请注意，可变形ConvNets的性能增益与这些铃声和口哨是互补的。

5 总结

本文提出了可变形ConvNets，它是一种简单，有效，深入和端到端的解决方案，用于模拟密集空间转换。我们首次表明，对于复杂的视觉任务，如对象检测和语义分割，在CNN中学习密集空间变换是可行的和有效的。