两种卷积神经网络的介绍

STUDENT 1 STUDENT 2 STUDENT 3
Student Number 1 Student Number 2 Student Number 3

摘　要： 卷积神经网络是计算机视觉领域最重要的成果，基于卷积神经网络，出现了无数新的变种。我们在本文介绍两种最近提出的卷积神经网络，他们在图像识别领域有着至关重要的开创性成就。首先介绍的是 SAC 模型，它是 DetectoRS 模型下的一个子模型，SAC 模型中有两个卷积网络，他们有着不一样的空洞率，SAC 会训练一个判别器来从两个卷积网络中进行选择，从而从大视野和小事业两个角度来识别图像，对于大型物体偏多的图像，SAC 能取得更好的效果。其次，我们介绍了 SC 模型，它在卷积网络中加上 2D 的偏移，再进行后续的计算，取得了更好的表现。

关键词： SAC；DC；卷积神经网络；计算机视觉；图像检测；实例分割。

目录

1 绪论

1.1 课题研究背景

卷积神经网络是计算机视觉领域最重要的成果，计算机科学家们基于卷积神经网络，提出了很多新的变种，而各种各样的变种在计算机视觉领域的表现更是长江后浪推前浪。研究这些变种网络，对于计算机视觉领域我们在计算机视觉领域的理解，对于卷积神经网络的理解，以及未来的科研思路，有着至关重要的作用。

1.2 本文的研究内容

本文选择了两个当今图像识别领域比较前沿的模型进行了基本原理的介绍，详细说明了各个结构的作用，并结合实验对模型的可靠性进行了评估，最后，说明了该模型的应用场景。

本文研究的主要对象为可转换的空洞卷积网络（Switchable Atrous Convolution）和可变换卷积网络（Deformable Convolution）。

SAC 的提出是为了解决传统图像分割中，使用固定空洞率的的空洞卷积模型会导致图像识别不准确的问题，由于空洞卷积的计算方式类似于棋盘格式，某一层得到的卷积结果，来自上一层的独立的集合，没有相互依赖，因此该层的卷积结果之间没有相关性，即局部信息丢失。同时远距离获取的信息没有相关性，由于空洞卷积稀疏的采样输入信号，使得远距离卷积得到的信息之间没有相关性，影响分类结果。

DC 的提出是为了处理识别物体的几何变换问题，DC 能够学习具有转换不变性的特征，使得能够识别在不同的大小、姿势、视角、部分变形情况下的物体。

2 SAC[1]

2.1 设计思路

SAC（Switchable Atrous Convolution），即可转换的空洞卷积。该方法来源于 DetectoRS 模型中微观部分特征获取。

2.1.1 DetectoRS 模型概述

介绍 SAC 之前， 需要先了解一下它的父级模型——DetectoRS。

许多现代的目标检测模型，都是以“多看多想（looking and thinking twice）”为底层逻辑，表现出了优异的性能。DetectoRS 模型也基于此机制进行了深层次的探讨，旨在进一步提高目标检测领域的算法性能。

DetectoRS 模型从宏观和微观上来看，有两个子模型组成：RFP 和 SAC。

① 循环特征金字塔 RFP（Recursive Feature Pyramid）

1）相关研究：特征金字塔网络 FPN(Feature Pyramid Networks)

FPN 提供了一条自上而下的路径来融合多个尺度特征图的特征，利用高层网络提取的语义信息和低层网络提取的细粒细节特征信息来预测多尺度的目标。FPN 模型如图 2.1 所示。

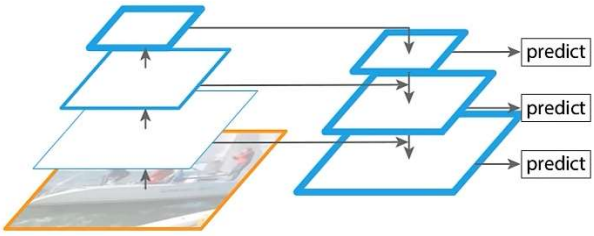


图 2.1 FPN 模型

Figure 2.1: FPN model

从图中可知，原图经过网络的处理，尺寸不断减小，模型对于每个尺寸下图片应该有着不同的预测结果，而 FPN 模型让模型预测之前除了要接受对应层的信息以外，还要额外接受上一层的信息，这就导致模型能更好的融合图像细节与宏观的信息。

2）RFP 介绍

RFP 是在 FPN 的基础上，反复的通过自下而上的骨干结构（backbone）来丰富 FPN 的表征能力，这个过程可以看作是一个递归操作。从本质上来看，这

样的递归操作让模型查看图像两次甚至更多次。RFP 的基本原理如图 2.2 所

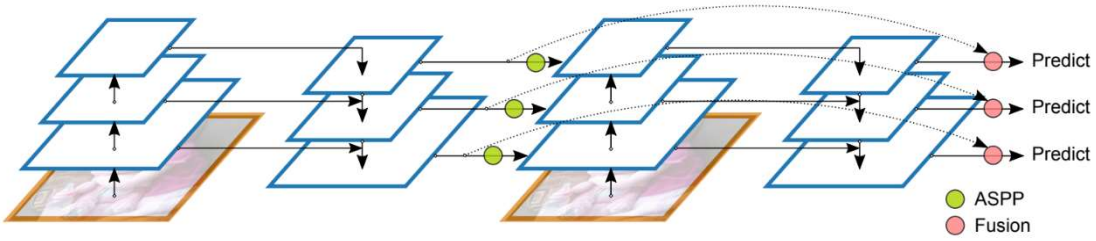


图 2.2 RFP 基本原理

Figure 2.2: Fundamentals of RFP

示。

如图 2.2 所示，以两层递归的 RFP 为例，原图经过 FPN 的处理后，需要作为输入，再次传回 FPN 模型中，再经过 ASPP 和 Fusion 两个子模型，得到最终输出结果。

3）空洞空间金字塔池化模型 ASPP（Atrous Spatial Pyramid Pooling）

ASPP 能将 FPN 处理后的输入转换为后续 RFP 网络中所需要的输入特征。ASPP 中有 4 个并行分支，也就是说原输入会根据其通道数分为 4 份以进入 4 个分支进行运算，运算结束后需要重新拼接回一个整体。具体的运算规则不是本文的重点，就不再介绍。

4）聚合模型（Fusion）

这个模型需要将 FPN 的输出特征f_i^t与 RFP 的输出特征f_i^{t+1}进行结合，会通过一个类似注意力计算的过程来为两次输出特征分配权重，最后组合在一起。具体的运算规则不是本文的重点，就不再介绍。

② 可转换的空洞卷积 SAC（Switchable Atrous Convolution）

在微观层面，SAC 模型以不同的空洞率（atrous rates）对相同的输入特征进行卷积，然后合并卷积后的结果。相关内容将在后文进行详细介绍。

2.1.2 SAC 模型的设计思路

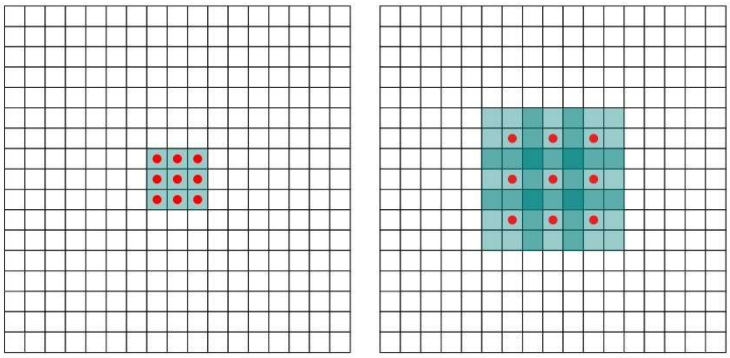
① 空洞卷积 AC

空洞卷积（atrous convolution）是针对图像语义分割问题中下采样会降低图像分辨率、丢失信息而提出的一种卷积思路。利用添加空洞扩大感受野，让原本 3×3 的卷积核，在相同参数量和计算量下拥有 5×5 或者更大的感受野。如果使用 5×5 的卷积核也能得到更大的感受野，但是会出现更多的参数，影响算法执行速度。空洞卷积通过一个超参数“扩张率（dilation rate）”来衡量。它的数学表达式如下所示：

$$F_i = \left(2^{i+1} - 1\right) \times \left(2^{i+1} - 1\right) \tag{2.1}$$

式中F_i表示扩张率为i的 3×3 卷积核的感知野大小。如$i = 1$时，此时卷积核的感知野大小为 3×3，即常规的卷积操作；$i = 2$时，此时卷积核的感知野大小为 7×7。

下图 2.3 为空洞卷积公式的可视化呈现。(a)图对应的扩张率为 1，和普通的卷积操作一样；(b)图对应的扩张率为 2，实际的卷积核大小还是 3×3，但是他能获取到 7×7 大小的信息，此时空洞的位置需要全部填入 0，以参与下一层网络的运算。



(a) (b)

图 2.3 空洞卷积

Figure 2.3: atrous convolution

② 可转换的空洞卷积 SAC

SAC 模型有三个组成部分：两个全局上下文组件、一个 SAC 组件。如图 2.4 所示。

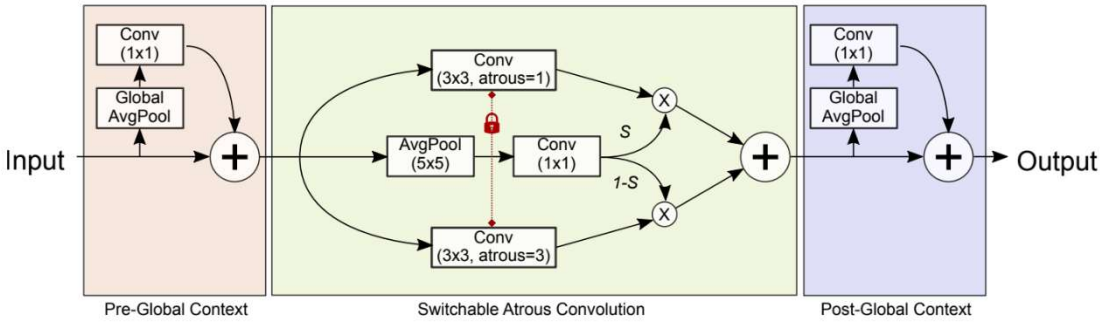


图 2.4 可转换的空洞卷积

Figure 2.4: Switchable Atrous Convolution (SAC).

1）全局上下文组件

全局上下文模块中，输入的特征值要首先经过一个全局平均池化层压缩，随后经过一个卷积核为 1×1 大小的卷积层，最后将输出和原输入进行求和操作。作者通过实验证实了全局上下文组件对 SAC 组件有着正向作用，这在后文的实验中也得到了证实。

作者并没有给出严格的证明或者推导，而是更多的将这个组件当成模型训

练过程中的一个特殊方法（trick）。因为作者也只给出了自己的推测。

2）SAC 组件

原模型中的全部卷积层都是按照 SAC 来定义的。作者使用了如下表达式来定义卷积操作：

$$S(x) \cdot Conv(x, w, 1) + \big(1 - S(x)\big) \cdot Conv(x, w + \Delta w, r) \tag{2.2}$$

式中$Conv(x, w, 1)$是卷积操作，且空洞率为 1，即无空洞操作；w是预训练模型的参数权重；r是空洞率；Δw是一个可训练的参数矩阵；$S(x)$是转换函数，用来判定这一层卷积核应该重视宏观角度还是重视微观角度，$S(x)$由一个 5×5 的平均池化层和 1×1 的卷积层组成。

作者提出一种锁机制，也就是将常规卷积网络的权重设为w，将带有空洞率的卷积网络的权重设为$w + \Delta w$，且Δw初始设置为 0。如果不对带有空洞率的卷积网络的权重进行初始化，那么训练的效果肯定是会受到影响的。如此设置之后，两个网络的初始权重将是完全一致的，在相同的初始状态下以不同的空洞率粗略地检测不同大小的目标，作者认为如此设置更加合理。

3） 作用

SAC 可以给模型赋予自适应观察宏观或微观的能力。如果图像中有大型物体，SAC 的表现性能将会更好。

2.2 应用

作者对于 DetectoRS 模型进行了实验，实验结果如下图 2.5 所示。

图 2.5 DetectoRS 模型的实验

Figure 2.5: From left to right: visualization of the detection results by HTC, ‘HTC + RFP’, ‘HTC + SAC’ and the ground truth.

实验结果表明，应用了 RFP 模型的算法，更类似于人类的视觉感知，它能够选择性地增强或抑制神经元的激活，可以更容易地找到被遮挡的物体。

而应用了 SAC 模型的算法，因为它能够根据需要增加感受野，因此它更有能力检测图像中的大型物体。

空洞卷积最初的提出是为了解决图像分割的问题而提出的，常见的图像分割算法通常使用池化层和卷积层来增加感受野，同时也缩小了特征图尺寸，然后再利用上采样还原图像尺寸，特征图缩小再放大的过程造成了精度上的损失，因此需要一种操作可以在增加感受野的同时保持特征图的尺寸不变，从而代替下采样和上采样操作，在这种需求下，空洞卷积就诞生了。但是我们不能对所有的场景都采用空洞卷积，对于小物体还使用空洞卷积只会适得其反，因此，可变换的空洞卷积应运而生，事实也证明可变换的空洞卷积确实有着非常优秀的表现。

作者证明了有着 SAC 的 DetectoRS 模型在图像检测、实例分割、全景分割中都有着优异的性能。

3 Deformable Convolution[2]

3.1 设计思路

通过增加少量的参数和计算量在训练过程中得到偏移量，使得模型能够学习具有变换不变性的特征。

3.3.1 可变换卷积

基本思想是在卷积网络中加上 2D 的偏移，如图 1 所示：

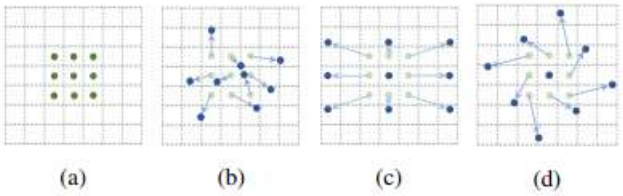


图 3.1 可变换卷积

Figure 3.1: Deformable Convolution

图（a）是标准的卷积采样的网格，图（b）是应用了偏移的卷积采样的网格，也就是可变换卷积采样的网格，图（c）（d）是可变换卷积采样网格的不同变体，分别表示物体大小的放缩和物体的旋转。卷积的偏移通过输入特征图进行学习，实现根据输入的特征图确定具体的变换方式来自适应提取特征。

首先定义一个卷积核的计算，R为卷积核所要计算的相对位置集合，假设一个3×3 的卷积核，那么R可定义为$R = \{(-1,1),(-1,0),\dots,(0,1),(1,1)\}$。设$x$为输入特征图，$y$为输出特征图，$p_{ij}$表示为$i$行$j$列$(i,j)$, w_{ij}为卷积核对应的权重，那么标准的卷积计算可表示为公式(3.1)：

$$y(p_{ij}) = \Sigma_{p_n \in R} w(p_n) \cdot x(p_{ij} + p_n), \tag{3.1}$$

而对于可变换卷积网络，计算可表示为公式(3.2)：

$$y(p_{ij}) = \Sigma_{p_n \in R} w(p_n) \cdot x(p_{ij} + p_n + \Delta p_n), \tag{3.2}$$

其中Δp_n为卷积每个相对位置的偏移量，满足条件$\{\Delta p_n | n = 1,\dots,N\}$，$N = |\mathrm{R}| = 9$。而由于经过$p_{ij} + p_n + \Delta p_n$计算后的位置并不是整数，作者采用双线性插值的方法来计算$x(p_{ij} + p_n + \Delta p_n)$的值。设$p_{target} = p_{ij} + p_n + \Delta p_n$，计算方法表示为公式(3.3)：

$$x(p_{target}) = \Sigma_{x(q) \in \mathrm{X}} G(x(p_{target}), x(q)) \cdot x(q), \tag{3.3}$$

其中X为输入特征图，$G(\cdot,\cdot)$为二维双线性插值计算，计算可表示为公式(3.4)：

$$G(A,B) = g(A_x, B_x) \cdot g(A_y, B_y), \tag{3.4}$$

$g(\cdot,\cdot)$为一维双线性插值计算，计算可表示为公式(3.5)：

$$g(a,b) = \max(0,1-|a-b|), \tag{3.5}$$

因为大部分输入特征图的点在公式(3.4)中计算为 0，所以计算速度很快。整体的计算流程如图 3.2 所示。

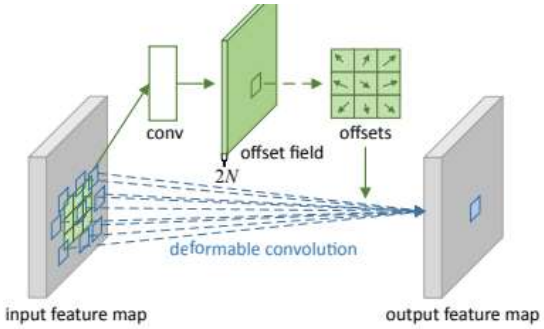


图 3.1 3×3 可变换卷积

Figure 3.2 Deformable Convolution

3.3.2 可变换卷积 RoI 池化层

可变换 RoI 池化层与可变换卷积采用的方法相同，也是设置一个可学习的偏移。根据实际输入的特征图确定每一个块提取特征的位置，使得能够自适应定位不同形状的物体的部分特征。RoI 池化层的计算可表示为公式(3.6)：

$$y(p_{ij}) = \Sigma_{p_n \in \mathrm{bin}(\mathrm{p_{ij}})} x(p_{ij}+p_n)/n_{bin(p_{ij})}, \tag{3.6}$$

其中$bin(p_{ij})$为位置p_{ij}块所需要的输入特征图的位置，$n_{bin(p_{ij})}$为需要特征图位置的个数。而对于可变换 RoI 池化层的计算可表示为公式(3.7)：

$$y(p_{ij}) = \Sigma_{p_n \in \mathrm{bin}(\mathrm{p_{ij}})} x(p_{ij}+p_n+\Delta p_n)/n_{bin(p_{ij})}, \tag{3.7}$$

具体的计算过程与可变换卷积计算相同。整体计算流程如图 3 所示：

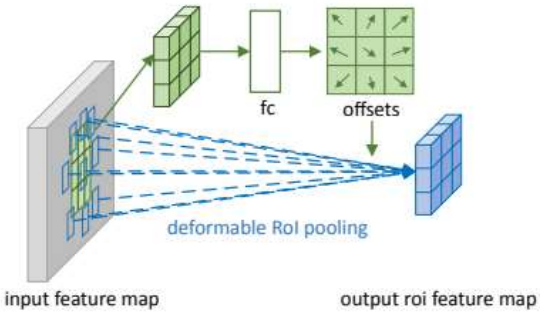


图 3.2 可变换 RoI 池化层

Figure 3.3: Deformable RoI pooling

3.2 实验

作者将可变换卷积与空洞卷积[3]在目标检测任务上进行比较，由表 1 可

知，可变换卷积效果相比于空洞卷积略微提升。

表 3.1 可变换卷积与空洞卷积

Table 3.1: Deformable Convolution & Dilated Convolution

deformation models	DeepLab mIoU@V/@C	Class-aware RPN mAP@0.5/@0.7	Faster R-CNN mAP@0.5/@0.7	R-FCN mAP@0.5/@0.7
atrous conv(2,2,2)	69.7/70.4	68.0/44.9	78.1/62.1	80.0/61.8
atrous conv(4,4,4)	73.1/71.9	72.8/53.1	78.6/63.1	80.5/63.0
atrous conv(6,6,6)	73.6/72.7	73.6/55.2	78.5/62.3	80.2/63.5
atrous conv(8,8,8)	73.2/72.4	73.2/55.1	77.8/61.8	80.3/63.2
deformable conv	75.3/75.2	74.5/57.2	78.6/63.3	81.4/64.7
deformable RoI pooling	NA	NA	78.3/66.6	81.2/65.0
deformable conv& RoI pooling	NA	NA	79.3/66.9	82.6/68.5

此外作者还比较加上可变换卷积后模型传播所用的时间，由表 2 的实验结果可知，可变换卷积增加的计算量较小。

表 3.2 可变换卷积计算量

Table 3.2: Computation of Deformable Convolution

Method	#params	Net forward (sec)	Runtime (sec)
Deeplab@C	46.0M	0.610	0.650
Ours	46.1M	0.656	0.696
Deeplab@V	46.0M	0.084	0.094
Ours	46.1M	0.088	0.098
Class-aware RPN	46.0M	0.142	0.323
Ours	46.1M	0.152	0.334
Faster R-CNN	58.3M	0.147	0.190
Ours	59.9M	0.192	0.234
R-FCN	47.1M	0.143	0.170
Ours	49.5M	0.169	0.193

最后，作者进行消融实验，比较传统的 Backbone 与在传统 Backbone 基础上加上可变换卷积后的效果。由表 3 可知，可变换卷积在一定程度上学习到物体的变换不变性的特征，效果得到提升。

表 3.3 可变换卷积消融实验

Table 3.3 Ablation Experiment of Deformable Convolution

Method	Backbone	mAP@[0.5:0.95]	mAP@0.5	mAP@[0.5:0.95] (small)	mAP@[0.5:0.95] (mid)	mAP@[0.5:0.95] (large)
Class-aware RPN	ResNet-101	23.2	42.6	6.9	27.1	32.1
Ours		25.8	45.9	7.2	28.3	40.7
Faster RCNN	ResNet-101	29.4	48.0	9.0	30.5	47.1
Ours		33.1	50.3	11.6	34.9	51.2
R-FCN	ResNet-101	30.8	52.6	11.8	33.9	44.8
Ours		34.5	55.0	14.0	37.7	50.3
Faster RCNN	Aligned-Inception-ResNet	30.8	49.6	9.6	32.5	49.0
Ours		34.1	51.1	12.2	36.5	52.4
R-FCN	Aligned-Inception-ResNet	32.9	54.5	12.5	36.3	48.3
Ours		36.1	56.7	14.8	39.8	52.2

3.3 应用

Deformable Convolution 能够学习具有转换不变性的特征，使得能够识别在不同的大小、姿势、视角、部分变形情况下的物体。

传统卷积神经网络等方法由于其本身的几何结构，受限于模型的几何变换能力，例如卷积神经网络 CNN 是通过卷积单元在固定的位置上采样输入的特征图来提取特征、RoI 池化层则是将特征图划分到固定的空间块中。为了处理识别物体的几何变换问题，传统的解决方法分为两种：

a）通过数据增强丰富数据集，模型学习丰富的特征以此保证鲁棒性；

b）通过学习转换不变性的特征，使得模型具有鲁棒性，例如 SIFT 等。

虽然这两种方法起到一定的效果，但这两种方法存在以下问题：

a）通过数据增强使得模型具有鲁棒性的方法需要大量的训练和更加复杂的模型来学习丰富语义的特征。

b）以往的方法假设数据几何变换是固定已知的，通过先验知识进行数据增

强或设计具有特征不变性的特征学习方法。而对于实际数据的几何变换，假设可能是错误的，削弱了模型的生成能力。

c）以往设计的具有特征不变性的特征学习方法是人工的，而实际数据是未知的，这会导致难以设计合理的特征学习方法。

参考文献

[1] Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10213-10224.

[2] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

[3] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.