Acta Physica Sinica

用于高效物体检测的深度卷积神经网络的空间金字塔池化层*

陈若愚[1)2)†] 黄锟[1)] 皮峻银[1)] 赵占宇[1)] 王旭薇[1)] 刘宜奇[1)] 刘少文[1)] 郭子沐[1)]

1) (东北大学秦皇岛分校控制工程学院, 秦皇岛 066004)

2) (香港城市大学机械工程学系, 香港特别行政区 999077)

物体检测，旨在定位并识别图像中的物体，是计算机视觉的核心问题之一。现有的深度卷积神经网络（CNN）需要固定大小（例如224×224）的输入图像。该要求是人工的，并且可能降低对任意大小或比例的图像或子图像的识别精度。而目标检测需要对图像的生成区域进行多次的卷积运算，存在较大的物体检测计算复杂度。为了消除上述限制，提出了用于深度卷积神经网络的空间金字塔池化策略。新的组合网络称之为SPP-net，该网络可以生成固定长度的表示向量，而与图像大小或比例无关。在ImageNet分类数据集上，验证了SPP-net可以提高卷积神经网络的准确率。基于R-CNN物体检测算法，与空间金字塔池化结合，提出一种可以输入任意大小图片的高效物体检测方法。使用SPP-net用于物体检测，可以对整个图像只计算一次特征图，然后在任意区域（子图像）中合并特征，以生成固定长度的表示向量以训练检测器，避免了重复计算卷积特征。与R-CNN方法相比，SPP-net将计算速率提高了24-102倍，同时在Pascal VOC 2007数据集上达到了更高的准确率。

关键词: 空间金字塔池化, 卷积神经网络, 物体检测, 高效

PACS: 07.05.Mh, 07.05.Pj

---

* 河北省高等学校科学研究重点项目(批准号: ZD2019305)、国家自然科学基金(批准号：61873307)资助项目

† 通信作者. E-mail: chenruoyu@neuq.edu.cn 电话: 13081868853

1 引　言

物体检测(object detection)，是一种使计算机能够在图像中自动找到既定类别的物体，并判断物体的类别、位置、大小及置信度的技术。我们的视觉，主要是由深度卷积引起的，正在见证我们的传统神经网络[1]和大规模训练数据[2]的快速、革命性的变化。最近，基于卷积神经网络(convolutional neural network, CNN)的方法在物体检测[3][4][5]的技术水平上有了很大的改进。

然而，在CNN的训练和测试中存在一个技术问题：目前流行的CNN需要一个固定的输入图像大小，这限制了输入图像的长宽比和尺度。当应用于任意大小的图像时，当前的方法主要是将输入图像裁剪为固定大小，但裁剪区域可能不包含整个对象，而扭曲的内容可能导致不必要的几何失真。为什么CNN需要一个固定的输入大小呢？CNN主要由两个部分组成：卷积层，以及随后的全连接层。因此，卷积层不需要固定的图像大小，可以生成任意大小的特征图。另一方面，根据定义，全连接层需要有固定大小/长度的输入。因此，固定图像尺寸的约束只来自于全连接层。

如何设计高效的物体检测算法，以减少检测系统整体的计算代价并提高检测性能是物体检测研究的主要问题。本文中，我们提出了一种通过空间金字塔池化层(spatial pyramid pooling)[6][7]来消除卷积神经网络需要固定的输入图片大小的限制，新的网络我们称为SPP-net。我们基于一些已有的分类网络[3][8][9]或物体检测算法的[4]网络，在网络的最后一层卷积层添加了空间金字塔池化层。空间金字塔池化层可以将卷积层输出的任意大小的特征图像池化为固定长度的表示向量，然后与全连接层相连接。基于R-CNN的目标检测方法，我们将SPP-net作为其主网络，提出一个新的高效目标检测方法。除了对输入图像无固定大小的限制，我们仅需要对整幅图像进行一次卷积运算，降低计算复杂度，避免了重复计算卷积特征。

在ImageNet数据集的一系列对照实验中，我们证明了在现有卷积神经网络[3][8][9]，经空间金字塔池化层改进的四种不同的CNN架构，超过了原本的卷积神经网络的准确率。与R-CNN[4]方法相比，SPP-net将计算速率提高了24-102倍，同时在Pascal VOC 2007数据集上达到了更高的准确率。

2 嵌入空间金字塔池化的卷积神经网络

2.1 卷积层与特征图像

我们首先考虑一个拥有七层的分类网络：Alexnet[8]。该网络由5个卷积层和2个全连接层组成。每个卷

积层后附加一个池化层，用于减半图片的感受野。最后一层卷积层将由张量重塑为特征向量，并连接全连接层。全连接层包含两层隐藏层，每个隐藏层包含2048个节点，输出层包含N个节点，其中N由分类的类别数决定。除了输出层以softmax函数激活，每一层卷积层或全连接层后以relu函数激活，该网络的结构如图1所示。该深度卷积神经网络需要固定大小的输入图像，原因是全连接层依赖于长度固定的特征向量。而卷积层采用滑动窗口的方式处理图像，输入图片大小不同不影响卷积的运算，且输出图像与输出图像的长宽比保持大致相同。这里输出的图像称为特征图像[1]，这些图像不仅包括了响应的强度也包括了空间位置信息。

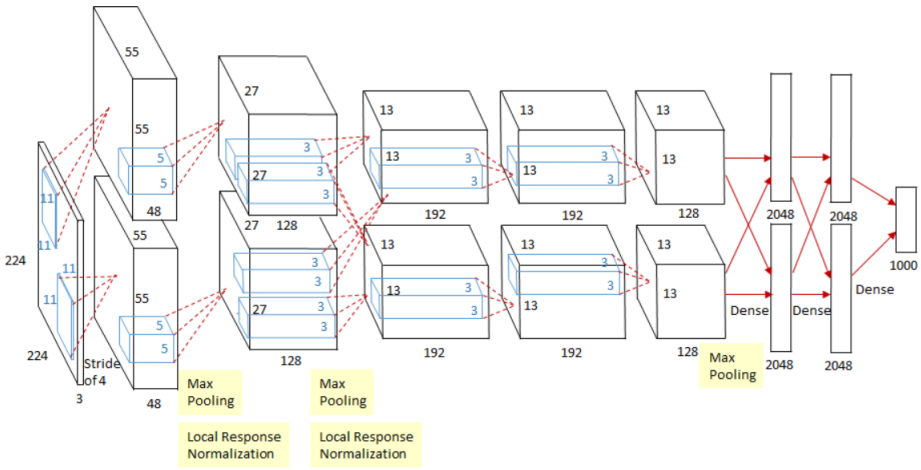

图 1: 用于实验的Alexnet的结构，网络的输入大小固定为 224×224×3

Fig 1: The structure of alexnet for experiment, the input size of network is fixed as 224 × 224 × 3.

我们以在Imagenet数据集[2]上官方训练好权重初始化网络的卷积层部分的参数，以非固定大小的图像作为输入并得到特征图像。在图2中，我们以可视化的形式将特征图像展示出来，它们是第五层卷积层中滤波器的卷积运算的输出结果。图2（c）展示了在ImageNet数据集中目标区域响应最大的特征图片的单通道图片。我们可以看到不同的滤波器可以被特定的语义激活，比如圆形物体激活，三角形物体激活，倒三角形物体激活。这些输入图像中的形状会在特征图像上相应位置被激活。

值得一提的是我们在生成图2的特征图像时没有固定输入图像的大小，深度卷积层生成这些特征图像的过程就像传统方法产生特征图像一样[10][11]，在传统方法中，尺度不变特征变换（SIFT）特征[12]或者图像块被密集提取（densely extract）并编码。编码的方式包括矢量量化（vector quantization）[14][15]，稀疏编码（sparse coding）[16][17]。这些编码的特征包括图像特征，然后由词袋模型（Bag-of-Words，BoW）或者空间金字塔池化[14][18]。所以，深度卷积特征也可以通过相似的方法池化。

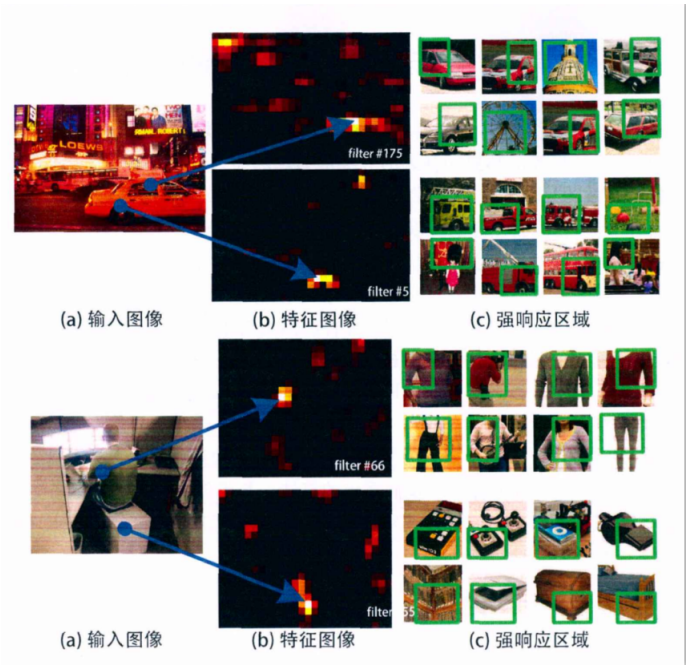

图 2: 可视化特征图像。（a）PASCAL VOL 2007数据集中的两张图像。（b）conv5层的某些特征图像。箭头指向区域为图像中最强相应和他们的位置。（c）ImageNet数据集中对这些滤波器产生强相应的图像。绿色框标记了产生最大响应的感受区域

Fig 2: Visual feature image. (a) Two images in the Pascal Vol 2007 dataset. (b) Some characteristic images of conv5 layer. The arrow pointing area is the strongest corresponding and their position in the image. (c) The Imagenet dataset produces strong corresponding images for these filters. The green box marks the area of perception that produces the greatest response.

2.2 空间金字塔池化层

卷积层接收任意大小的输入，并可以产生变大小的输出，而全连接层需要固定长度的特征向量。分类器（SVM/softmax）或全连接层需要固定大小的输入向量。词袋模型可以通过将特征池化到一起的方法产生固定长度的特征向量。相比于词袋模型，空间金字塔池化方法可以保留更多的空间信息，因此具有更好的性能。空间金字塔池化的每个区域的大小与图像大小成固定的比例，因此池化区域的数量是固定的，与图像大小无关。这种新方法与以往的滑窗型池化方式不同——滑窗的数量取决于输入尺寸。

为了使深度卷积神经网络可以接收任意大小的输入，我们将最后一个池化层（最后一个卷积层之后，例如Alexnet的第五层卷积层）替换为空间金字塔池化层，图3介绍了我们的方法。在每个空间区域之后，我们将每个滤波器得到的结果进行最大池化。空间金字塔池化的输出结果是$M \times K$维的向量，其中M是空间区域的总数，最后一层卷积层滤波器个数为K。这个固定维数的向量可以被用作全连接层的输入。

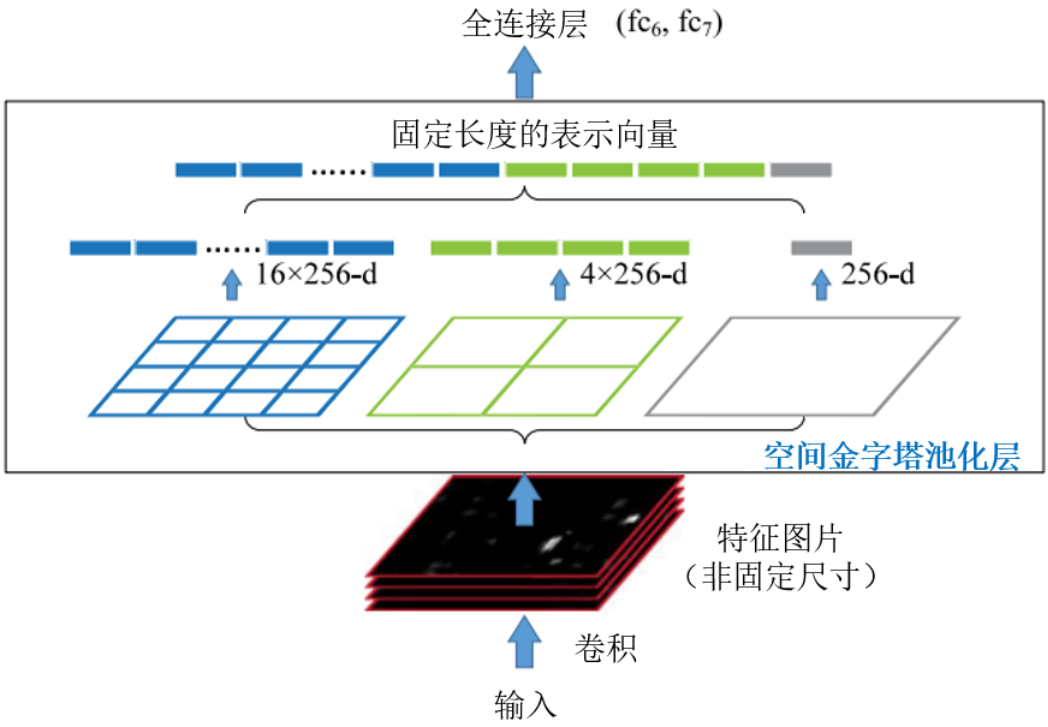

图 3: 包含空间金字塔池化的卷积结构层的网络结构，其中输入图片非固定大小，最后一层卷积层包含256个滤波器

Fig 3: The network structure including convolution structure layer of spatial pyramid pooling, in which the input image is not fixed size, and the last convolution layer contains 256 filters.

当使用空间金字塔池化时，输入的图像可以是任意尺寸的。我们可以将输入图像伸缩到任意尺度，其中*min*(长, 宽)= 180, 224等,并使用同一个深度网络。当输入的图像是不同尺度时，深度网络会使用相同大小的滤波器在提取到不同尺度的特征。

有趣的是，最大颗粒的金字塔级别使用一个区域覆盖整个图像，这本质上是一种“全局池化”操作，这些也被一些同期工作所研究。使用全局均值池化来减小模型的大小及防止过拟合现象[19][20]。在测试阶段所有的全连接层之后使用全局平均池化来提高精度[21]；在弱监督物体识别任务中使用全局最大池化操作[22]。这种全局池化的操作与传统方法中的词袋方法有类似的效果。

2.3 空间金字塔池化网络

空间金字塔池化的优点独立于使用不同的卷积神经网络结构。我们研究了现有的四种不同的神经网络结构[23]。这些卷积神经网络结构如表1所示。

表 1. 基础网络结构：滤波器数量×滤波器尺寸（例如96×72），滤波器移动步长（例如str 2），池化窗口尺寸（例如Pool 32），输出的特征图像尺寸（例如map size 55×55）

Table 1. Basic network structure: filter number × filter size (e.g. 96 × 72), filter moving step size (e.g. STR 2), pool window size (e.g. pool 32), output characteristic image size (e.g. map size 55 × 55.)

model	$conv_1$	$conv_2$	$conv_3$	$conv_4$	$conv_5$	$conv_6$	$conv_7$
ZF-5	96×7^2, str 2 LRN, pool 3^2, str 2 map size 55×55	256×5^2, str 2 LRN, pool 3^2, str 2 27×27	384×3^2 13×13	384×3^2 13×13	256×3^2 13×13	-	-
Convnet*-5	96×11^2, str 4 LRN, map size 55×55	256×5^2 LRN, pool 3^2, str 2 27×27	384×3^2 pool 3^2, 2 13×13	384×3^2 13×13	256×3^2 13×13	-	-
Overfeat-5/7	96×7^2, str 2 pool 3^2, str 3, LRN map size 36×36	256×5^2 pool 2^2, str 2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18	512×3^2 18×18

我们以这些卷积神经网络的结构为基础，加入空间金字塔池化层构成空间金字塔网络SPP-net。我们在最后一个卷积层之后的池化层生成6×6的特征图像，并接上两个含有4096个节点的全连接层和一个含有1000个节点的输出层，并以softmax激活函数激活。实验中我们采取的金字塔池化层对特征图进行3次空间金字塔池化，每次池化均匀地将特征图分为16、4、1块窗口并进行平均池化，最后将3次池化的结果连接成特征向量，特征图的通道数C决定特征向量的长度 l，为 $l = 21 \times C$。

我们将我们提出的SPP-net网络在ImageNet2012数据集上进行测试，表2为四种网络在不同条件下的预测误差，可看出空间金字塔池化方法提高了这些卷积神经网络的预测准确率，证明了我们提出的方法可以提高卷积神经网络的准确率。

表 2.数据集ImageNet2012的错误率

Table 2. Error rates in the validation set of ImageNet2012.

		top-1 error (%)			
		ZF-5	Convnet*-5	Overfeat-5	Overfeat-7
(a)	no SPP	35.99	34.93	34.13	32.01
(b)	SPP single-size trained	34.98 (1.01)	34.38 (0.55)	32.87 (1.26)	30.36 (1.65)
(c)	SPP multi-size trained	34.60 (1.39)	33.94 (0.99)	32.26 (1.87)	29.68 (2.33)

		top-5 error (%)			
		ZF-5	Convnet*-5	Overfeat-5	Overfeat-7
(a)	no SPP	14.76	13.92	13.52	11.97
(b)	SPP single-size trained	14.14 (0.62)	13.54 (0.38)	12.80 (0.72)	11.12 (0.85)
(c)	SPP multi-size trained	13.64 (1.12)	13.33 (0.59)	12.33 (1.19)	10.95 (1.02)

2.4 训练网络

理论上，上述的深度卷积网络可以通过标准的反向传播训练，而与输入图像尺寸关系不大。但在实际情况中用GPU实现的时候更倾向于在固定尺寸的输入图像上运行。接下来我们描述我们的训练方案，可以利用GPU处理单一尺寸的优势，同时保留空间金字塔池化的优点。

2.4.1 单一尺度训练

在我们之前的工作中，我们会首先从原始图片上截取大量固定尺寸的输入图像（224×224），这些截取得到的图像可以用于图像增强。对于一个给定了尺寸的输入图像，我们可以预先计算用于空间金字塔池化的空间区域大小。例如，在某网络结构最后一层卷积层输出的特征图像尺寸是a×a，某一层金字塔池化的空间区域大小是n×n，我们通过滑动窗口的形式来实现池化过程，窗口大小和步长都为 [a/n],方括号表示向下取整。通过最大池化得到了池化层后，我们将元素连接起来得到全连接层，图4展示了一个三级的金字塔池化方式。

我们使用单一尺寸来训练的主要目的是提高我们多尺寸训练的准确度，是我们进行多尺度训练的基础。

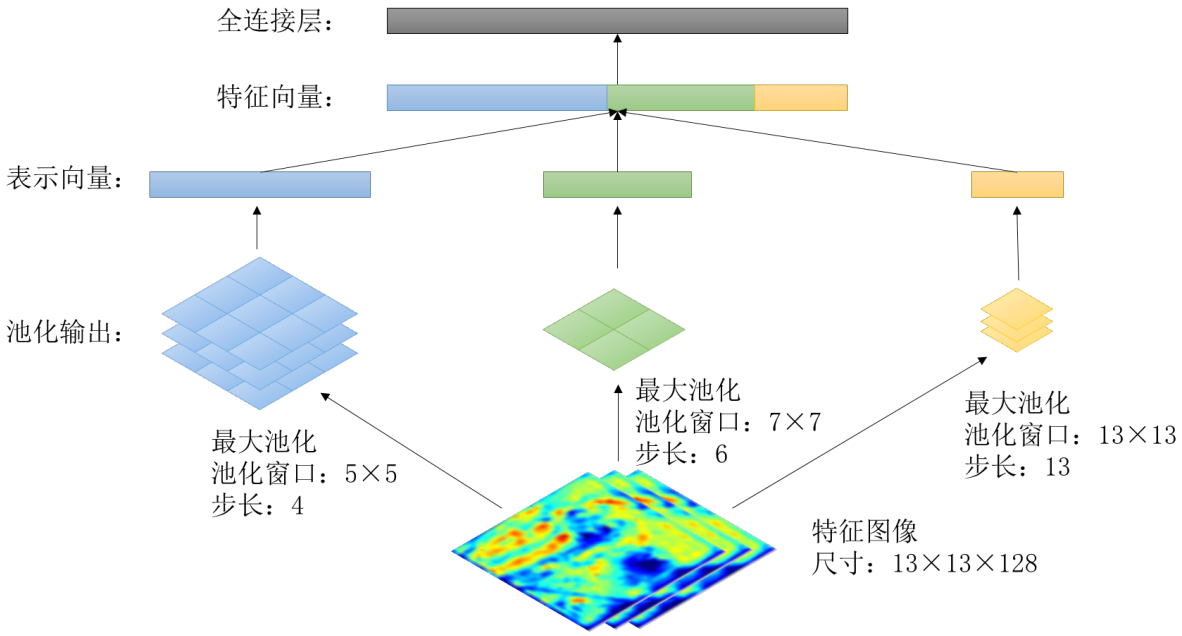

图 4: 一个三级金字塔池化的例子，这个金字塔池化层的输入是尺度大小为13×13的特征图像，所以，三个池化层输出的感受野大小为3×3、2×2、1×1

Fig 4: An example of three-level pyramid pooling is given. The input of the pyramid pooling layer is a feature image with a scale of 13 × 13. Therefore, the output receptive field size of the three pooling layers is 3 × 3, 2 × 2, 1 × 1.

2.4.2 多尺度训练

我们的多尺度金字塔池化网络是为了能够用在多种尺度的图片上，为了解决训练过程中图像尺度大小变化的问题，我们首先要提前固定图像的尺寸大小，这次我们使用180×180的尺寸而非224×224。我们先截取到224×224尺寸的图片，然后将图片的大小压缩到180×180，前后两个图片的差别在于分辨率也就是尺寸而非内容。为了让网络能够接收180×180的输入尺寸，我们实现了另一个固定大小输入的网络（180×180），这个网络在最后一层卷积层输出的特征图像大小为10×10，我们使用类似于上一节的办法，池化区域的大小和步长依然是[a/n]，这样一来这个网络的全连接层大小和224×224图片大小的全连接层长度相等，这样，不同尺度的图片就可以有相同长度的特征向量，就实现了多尺度训练。

为了减少运算损耗，我们在一个网络上训练完成对同一尺度图片集的一次遍历，然后我们切换到另一个尺度的图片集完成一次遍历，这个过程中训练参数保持不变，然后进行如此反复的学习过程。在训练过程中，我们发现，这种训练方法的收敛速度和传统方式的差不多。

多尺度训练的目的是模拟变化的输入尺寸，利用现有的经过优化的固定尺寸的网络来训练另一个尺度的图片。包括上述的两个尺度的图像训练，我们同时使用了尺度大小随机处于[180，224]图片来进行训练，每个尺度各完成一次遍历迭代。我们在下文的实验部分会具体介绍。

上述的单尺度和多尺度的训练方法仅用于训练网络。所以在测试阶段，我们可以将SPP-net用于各种尺度的图片了。

2.5 用于目标检测算法的SPP-net

与图像分类算法不同，目标检测算法不仅需要得出图像的类别，同时也需要计算出物体在图像中的位置，我们以R-CNN网络的算法为基础，加入空间金字塔池化层来改善网络。对于每张图像，我们使用“快速”模式的选择性搜索算法[24]产生大约2000个生成区域。之后，我们将图像调整人小使得min(长，宽)=s，并提取整张图的特征图像。我们使用 (Zeiler and Fergus)[25]网络训练SPP-net。对于每个生成区域，我们使用4层空间金字塔(1×1,2×2,3×3,6×6，总共50个区域)来池化特征。每个区域会产生一个12800 (256×50)维的表示向量。这些表示被用于全连接层的输入。然后，我们在这些表示特征的基础上，对于每个类别训练二元的支持向量机分类器。该网络结构图如图5所示。

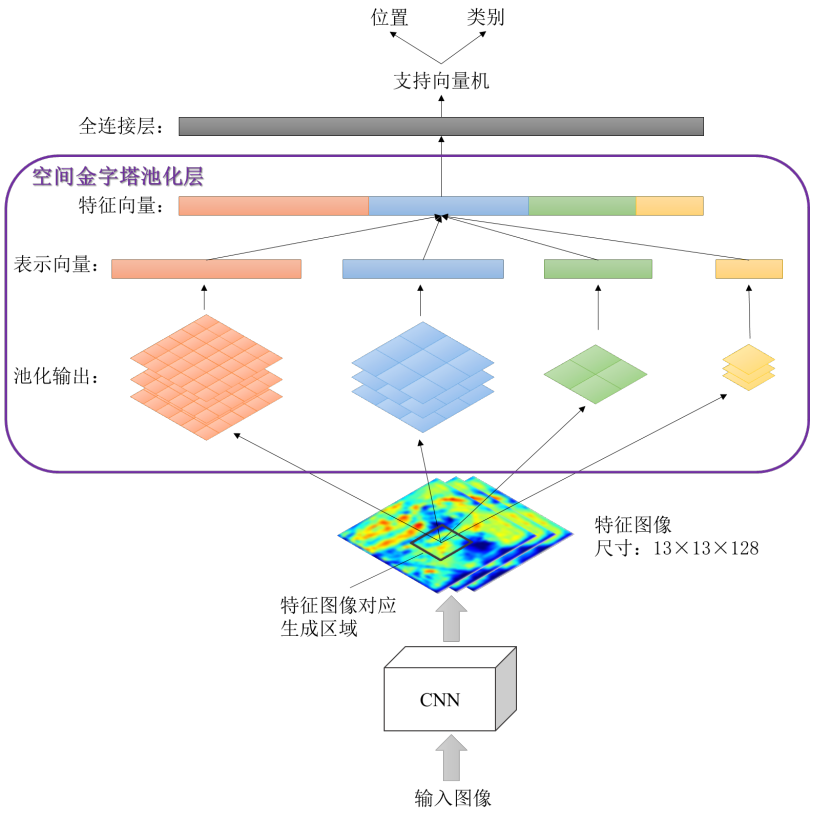

图5：以空间金字塔池化层改进的R-CNN目标检测算法架构，输入可为任意尺度的图像

Fig5：Based on the improved r-cnn target detection algorithm architecture of spatial pyramid pooling layer, images of any scale can be input.

3 实验系统及测量结果

3.1 Pascal VOC上的实验结果

我们在PASCAL VOC 2007测试集上对我们的方法进行了验证。表4显示了基于不同层特征的检测精度。对于我们的方法，我们分别使用单一尺度（s=688）以及5个尺度的特征。这里R-CNN的结果是[26]中报告的使用5个卷积层AlexNet[27]基础网络的版本。只使用$pool_5$层特征，我们的结果（44.9%）与R-CNN的（44.2%）精度相差无几。但是，如果使用没有经过微调的fc_6层特征，我们的方法在精度上会略差，一种可能的解释是，在预训练中全连接层是基于图像区域训练的，但是在检测中，它们的输入是特征图像的区域。特征图像区域在靠近边界的地方可能会有强响应，而基于图像区域的特征一般不会。这种用法上的不同带来的精度损失可以通过微调网络来弥补。使用微调之后的全连接$ftfc_{67}$，我们的结果与微调网

络之后R-CNN的精度相似或更好一些。在使用框回归之后，我们5个尺度的结果（59.2%）比R-CNN的（58.5%）好（0.7%），而我们单尺度的版本（58.0%）略差（0.5%）。

在表3中，我们进一步比较了在使用同一个预训练网络（ZF）时，R-CNN及SPP-net的精度。在这种情况下，我们的方法R-CNN有着相同的性能。R-CNN的性能得益于这个预训练网络，有一定的提升。表5比较了每类的检测精度。

表5同时也增加了其他检测方法的结果。选择性搜索算法（SS）[24]在尺度不变特征变换特征图像上使用空间金字塔匹配。可形变部件模型（DPM）[28]和Regionlet[29]使用方向梯度直方图（HOG）特征[30]。

Regionlet同结合包括conv5特征在内的多种特征将检测精度提升到46.1%[31]。DetectorNet[32]训练了一个深度网络用来输出像素级的物体掩码（mask）。这种方法与我们的方法一样，只需要对于整张图像使用一次深度卷积网络，但是他们的检测精度比较低，只有30.5%。

表 3. PASCAL VOC 2007上的检测结果（mAP）。使用与SPP(ZF)相同的基础网络结构

Table 3. Test results (map) on Pascal VOC 2007. Use the same infrastructure as spp (ZF).

	SPP(1-sc)	SPP(5-sc)	R-CNN
$ftfc_7$	54.5	55.2	55.1
$ftfc_7bb$	58.0	59.2	59.2
卷积时间（GPU)	0.053s	0.0293s	14.37s
全连接时间（GPU）	0.089s	0.089s	0.089s
总时间（GPU)	0.142s	0.382s	14.46s
加速比（vs.RCNN）	102x	38x	-

表 4. PASCAL VOC 2007上的检测结果（mAP）。”ft”和”bb”表示网络微调和框回归

Table 4. Test results (map) on Pascal VOC 2007. ”ft” and ”bb” indicate network tuning and box regression.

	SPP(1-sc)	SPP(5-sc)	R-CNN
$pool_5$ 43.0	42.9	44.2	
fc_6 42.5	44.8	46.2	
$ftfc_6$ 52.3	53.7	53.1	
$ftfc_7$	54.5	55.2	54.2
$ftfc_7bb$	58.0	59.2	58.5
卷积时间（GPU)	0.053s	0.0293s	8.96s
全连接时间（GPU）	0.089s	0.089s	0.07s
总时间（GPU)	0.142s	0.382s	9.03s
加速比（vs.RCNN）	64x	24x	-

表 5. PASCAL VOC 2007上的检测结果（mAP）。使用与SPP(ZF)相同的基础网络结构

Table 5. Test results (map) on Pascal VOC 2007. Use the same infrastructure as spp (ZF).

方法	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	personplant	sheep	sofa	train	tv
DPM [23]	33.7	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5
SS [32]	33.8	43.5	46.5	10.4	12.0	9.3	49.4	53.7	39.4	12.5	36.9	42.2	26.4	47.0	52.4	23.5	12.1	29.9	36.3	42.2	48.8
Regionlet [74]	41.7	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	49.1	26.6	57.0	54.5	43.4	16.4	36.6	37.7	59.4	52.3
DetNet [53]	30.5	29.2	35.2	19.4	16.7	3.7	53.2	50.2	27.2	10.2	34.8	30.2	28.2	46.6	41.7	26.2	10.3	32.8	26.8	39.8	47.0
RCNN $ftfc_7$ (A5)	54.2	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7
RCNN $ftfc_7$ (ZF5)	55.1	64.8	68.4	47.0	39.5	30.9	59.8	70.5	65.3	33.5	62.5	50.3	59.5	61.6	67.9	54.1	33.4	57.3	52.9	60.2	62.9
SPP $ftfc_7$ (ZF5)	55.2	65.5	65.9	51.7	38.4	32.7	62.6	68.6	69.7	33.1	66.6	53.1	58.2	63.6	68.8	50.4	27.4	53.7	48.2	61.7	64.7
RCNN bb (A5)	58.5	68.1	72.8	56.8	**43.0**	36.8	**66.3**	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8
RCNN bb (ZF5)	**59.2**	68.4	**74.0**	54.0	40.9	35.2	64.1	**74.4**	69.8	**35.5**	66.9	53.8	**64.2**	69.9	69.6	**58.9**	**36.8**	**63.4**	**56.0**	62.8	64.9
SPP bb (ZF5)	**59.2**	**68.6**	69.7	**57.1**	41.2	**40.5**	**66.3**	71.3	**72.5**	34.4	**67.3**	**61.7**	63.1	**71.0**	**69.8**	57.6	29.7	59.0	50.2	**65.2**	**68.0**

3.2 多模型融合的提升

多模型融合在基于卷积神经网络的图像分类任务中是提升精度的重要策略[27]。这里我们为物体检测任务提出了一种简单的模型融合方法。

我们使用相同的网络结果但是不同的随机初始化在ImageNet分类任务上预训练了另外一个网络。然后在这个新训练的网络上，我们重复以上检测任务的训练过程。表6. （SPP-net（2））显示了这个网络的检测结果。它的mAP与第一个网络相当（59.1% vs. 59.2%），并且在11个子类上比第一个网络的效果好。

表 6. PASCAL VOC 2007上两个模型融合的检测结果（mAP）

Table 6. Detection results of fusion of two models in Pascal VOC 2007 (map).

方法	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	personplant	sheep	sofa	train	tv
SPP-net (1)	59.2	**68.6**	69.7	57.1	41.2	40.5	66.3	71.3	72.5	34.4	**67.3**	61.7	63.1	71.0	69.8	57.6	29.7	59.0	50.2	65.2	68.0
SPP-net (2)	59.1	65.7	71.4	57.4	**42.4**	39.9	67.0	71.4	70.6	32.4	66.7	61.7	64.8	71.7	70.4	56.5	30.8	59.9	53.2	63.9	64.6
combination	**60.9**	68.5	**71.7**	**58.7**	41.9	**42.5**	**67.7**	**72.1**	**73.8**	**34.7**	67.0	**63.4**	**66.0**	**72.5**	**71.3**	**58.9**	**32.8**	**60.9**	**56.1**	**67.9**	**68.8**

有了这两个网络之后，我们首先分别使用它们对测试图片上的生成框进行打分。然后我们对于两个模型检测结果的并集（包含分数）使用非极大值抑制。某个模型输出的高置信度的输出会抑制其他模型低置信度的结果。在模型融合之后，mAP提升到了60.9%（表6）。对于20个子类别，融合模型在其中17上比单个模型的精度要高。这说明了这两个模型是一定程度上互补的。

进一步，我们发现这种互补性主要来源于卷积层。我们尝试融合两个使用相同卷积层，以及在微调时分别随机初始化的网络，并没有观察到提升。

4 讨论部分

4.1 特点

在本文中，我们打破了卷积神经网络只能处理固定大小输入图像的限制。我们提出的空间金字塔池化层使得使用卷积神经网络处理任意大小的图片成为可能。在此基础上，我们提出的算法使得多图像区域得以共享卷积层特征，同时对于一张图像而言，卷积层只需要被计算一次。

从物体检测角度看，我们的方法大大降低了算法的计算代价，提升了检测的效率。从长远的角度看，这种神经网络中多区域特征共享的机制会为图像相关算法的发展提供更多的可能性。

4.2 不足

卷积层特征图像的计算成为了神经网络检测系统的重要时间瓶颈。对于物体检测问题而言，一般我们需要对较大尺度的图像进行处理，以检测图像中较小的物体，这样所带来的卷积层计算量会成倍增加，计算代价的问题尤其严重。如何从卷积神经网络的根本入手为检测任务设计独特的网络，以减少卷积层计算量，是个值得研究的问题。

5 结　论

本文提出了一种利用空间金字塔池化层来训练深度网络的解决方案，新的网络我们称之为SPP-net，可以处理任意输入尺寸的图片。主要的创新手段为，基于原有的网络，对卷积层输出的特征图像进行空间金字塔池化，以生成固定长度的表示向量，解决网络需要输入固定大小图片的问题。在针对物体检测问题上，我们基于R-CNN算法，使SPP-net只需要进行一次卷积运算，避免了重复计算卷积特征。实验结果表明，我们提出的方法是R-CNN的24-102倍，并有较高的准确率。在ImageNet和Pascal VOC 2007数据集上，SPP-net在分类或物体检测任务中显示出了突出的准确性，并极大地加快了基于深度卷积神经网络的检测。从长远的角度看，这种神经网络中金字塔池化方法会为图像相关算法的发展提供更多的可能性。

致谢

感谢中国科学技术大学博士任少卿的博士毕业论文为我们提供了内容。

参考文献

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Comput., vol. 1, no. 4,pp. 541 – 551, 1989.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248 – 255.

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv:1312.6229, 2013.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580 – 587.

[5] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdevr, "Panda: Pose aligned networks for deep attribute modeling," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1637 – 1644.

[6] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in Proc. 10th IEEE Int. Conf. Comput. Vis., 2005, pp. 1458 – 1465.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2006, pp. 2169 – 2178.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1106 – 1114.

[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," arXiv:1311.2901, 2013.

[10] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in Proc. British Mach. Vis. Conf., 2011, pp. 1 – 12.

[11] A. Coates and A. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 921 – 928.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91 – 110, 2004.

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in Proc. 9th IEEE Int. Conf. Comput. Vis., 2003, p. 1470.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2006, pp. 2169 – 2178

[15] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in Proc. 10th Eur. Conf. Comput. Vis., 2008, pp. 696 – 709.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1794 – 1801.

[17] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Localityconstrained linear coding for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, p. 3306.

[18] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in Proc. 10th IEEE Int. Conf. Comput. Vis., 2005, pp. 1458 – 1465.

[19] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv:1312.4400, 2013.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," arXiv:1409.4842, 2014.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1717 – 1724.

[23] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1106 – 1114.

[24] Uijlings JR, Sande K E, Gevers T, et al. Selective search for object recognition [J]. International journal of computer vision,2013,104(2):15417l.

[25] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks IC]. Proceedings of Computer visionECCV 2014. Springer.2014:818833.

[26] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. Proceedings of Proceedings of the IEEE conference on computer vision and patern recognition.2014.580587.

[27] Krizhevsky A, Sutskever l, Hinton G E. Imagenet classification with deep convolutional neural networks [C]. Proceedings of Advances in neural information processing systems,2012.10971105.

[28] Felzenszwalb PF. Girshick RB. McAllester D, et al. Object detection with discriminatively-trained partbased models [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on,2010,32(9):16271645.

[29] Wang X, Yang M. Zhu S, et al. Regionlets for generic object detection [C]. Proceedings of Proceedings of the IEEE International Conference on Computer Vision,2013,1724.

[30] Dalal N. Triggs B. Histograms of oriented gradients for human detection IC]. Proceedings of Computer Vision and Paterm Recognition,2005. CVPR 2005. IEEE Computer Society Conference on, volume 1. IEEE,2005.886893.

[31] Zou WY, Wang X, Sun M. et al. Generic object detection with dense neural patterns and regionlcts [J] arXiv preprint arXiv:1404.4316,2014.

[32] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection IC] Proceedings of Advances in Neural Information Processing Systems,2013.25532561.

	姓名	性别	班级	学号	序号	贡献
组长	陈若愚	男	1701班	20178210	170103	1.组内工作任务的分配，组员间的协调 2.前期的论文选题与创新点选择 3.撰写部分method，并补充method部分原论文未展示或表示不形象的图，主要是图1，图4，图5 4.全局修改，理清文章脉络与结构。其中摘要部分大修，引言部分中修与补充，方法部分补充与纠正学术错误，讨论部分大修，结论部分大修
组员	赵占宇	男	1701班	20178523	170127	1.撰写部分引言部分 2.论文的latex排版 3.摘要等的英文翻译 4.全文语言的润色
	黄锟	男	1701班	20178391	170105	1.撰写引言第一稿 2.后期润色引言 3.复制整理参考文献
	刘宜奇	男	1704班	20179327	170419	1.撰写结果部分
	刘少文	男	1701班	20178706	170112	1.撰写讨论部分
	皮峻银	男	1701班	20179147	170129	1.前期与组长一起讨论选题选创新点 2.撰写部分method 3.后期整理method部分参考文献
	郭子沐	男	1701班	20178831	170134	1.撰写结论部分
	王旭薇	女	1701班	20177304	170133	1.撰写摘要部分

Spatial Pyramid Pooling in Deep Convolutional Networks for Efficient Object Detection *

Chen Ruoyu[1)2)†] Huang Kun[1)] Pi Junyin[1)] Zhao Zhanyu[1)] Wang Xuwei[1)] Liu Yiqi[1)] Liu Shaowen[1)] Guo Zimu[1)]

1) (*School of control engineering, Northeast University at Qinhuangdao, 066004, China*)

2) (*Department of mechanical engineering, City University of Hong Kong, Hong Kong, 999077, China*)

Abstract

Object detection, which aims to locate and recognize objects in images, is one of the core problems of computer vision. The existing deep convolution neural network (CNN) needs a fixed size input image. The target detection needs to convolute the generated region of the image many times, which has a large complexity of object detection. In order to eliminate these limitations, a spatial pyramid pooling strategy for deep convolution neural network is proposed. The new composite network is called SPP-net. In the ImageNet classification datasets, it is verified that SPP-net can improve the accuracy of convolutional neural network. Based on the R-CNN object detection algorithm, combined with spatial pyramid pooling, an efficient object detection method is proposed, which can input any size image. Using SPP-net for object detection, we can only calculate the feature map once for the whole image, avoiding the repeated calculation of convolution features. Compared with the R-CNN method, SPP-net improves the computing speed by 24-102 times, and achieves higher accuracy on Pascal VOC 2007 datasets.

Keywords: object detection, convolution neural network, spatial pyramid pooling, high efficiency

PACS: 07.05.Mh, 07.05.Pj

* Project supported by the key scientific research projects of colleges and universities in Hebei Province (Grant No. ZD2019305), the National Natural Science Foundation of China (Grant Nos. 61873307)