基于区域建议的目标检测的常见方法综述

专业：计算机技术

学号：21821279

姓名：余柏翰

摘要：由于物体检测与视频分析和图像理解的密切关系，近年来引起了很多研究的关注。传统的物体检测方法建立在手动提取的特征和浅层可训练的架构之上，通过构建复杂的框架，它们的性能很容易停滞不前，因为这些框架将多个低级图像特征与来自目标检测器和场景分类器的高级上下文信息相结合。随着深度学习的快速发展，引入了更强大的工具，这些工具能够学习语义的，高级的，更深层次的特征，以解决传统架构中存在的问题。这些模型在网络架构，培训策略和优化功能等方面表现不同，其中尤其以基于区域建议的目标检测网络框架最为瞩目，其在各类公开数据集上的表现均更加优异。本文综述了基于区域建议的目标检测框架，专注于典型的通用目标体系结构，以及一些改进和有用的技巧，以进一步提高检测性能。

关键词：深度学习，目标检测，区域建议

一、引言

为了获得完整的图像理解，我们不仅要集中精力对不同的图像进行分类，还要尝试精确估计每个图像中包含的目标的概念和位置。这个任务被称为物体检测，它通常由不同的子任务组成，如人脸检测，行人检测和骨架检测。作为基本的计算机视觉问题之一，目标检测能够为图像和视频的语义理解提供有价值的信息，并且与许多应用有关，包括图像分类，人类行为分析，人脸识别和自动驾驶。同时，继承神经网络和相关学习系统，这些领域的进展将发展神经网络算法，并将对可被视为学习系统的目标检测技术产生重大影响。然而，由于视点，姿势，遮挡和光照条件的巨大变化，使用额外的目标定位任务很难完美地完成目标检测。近年来，这一领域引起了如此多的关注。

目标检测的问题定义是确定目标在给定图像中的位置（目标定位）以及每个目标属于哪个类别（目标分类）。因此传统物体检测模型的流水线可以主要分为三个阶段：区域选择，特征提取和分类。

区域选择：由于不同的物体可能出现在图像的任何位置并且具有不同的长宽比和尺寸，因此用多尺度滑动窗口扫描整个图像是最容易想到的选择。虽然这种详尽的策略可以找出物体的所有

可能位置，但它的缺点也很明显。由于大量候选窗口，它在计算上很昂贵并且产生太多重复的计算。但是，如果仅应用固定数量的滑动窗口模板，则可能产生不令人满意的区域。

特征提取：为了识别不同的目标，我们需要提取可以提供语义和鲁棒表示的视觉特征。SIFT [1]，HOG [2]和 Haar-like[3]是最代表性的。这是因为这些特征可以产生与人脑中复杂细胞相关的表征[1]。然而，由于外观，光照条件和背景的多样性，很难手动设计鲁棒的特征描述符来完美地描述各种物体。

分类：此外，需要一个分类器来区分目标目标和所有其他类别，并使表示更具层次性，语义性和信息性，以便进行视觉识别。通常，支持向量机（SVM）[4]，AdaBoost [5]和可形变部件模型（DPM）[6]是不错的选择。在这些分类器中，DPM 是一种灵活的模型，它通过将目标部件与变形成本相结合来处理严重的变形。在 DPM 中，借助图形模型，精心设计的低级特征和运动学灵感的部分分解相结合，达到了比较好的效果。

基于这些判别的局部特征描述和浅层可学习的架构，在 PASCAL VOC 物体检测竞赛[7]上不断赢得榜首，并且获得了具有低硬件负担的实时嵌入式系统。然而，在 2010 - 2012 年期间，仅通过建立集合模型和采取一些的细微的改进没有区别太大的收益。主要原因有：1）使用滑动窗口策略生成候选边界框是冗余的，低效的和不准确的。2）语义鸿沟不能通过手动设计的低级描述符和有区别训练的浅层模型的组合来弥补。

随着深度神经网络（DNN）的出现，Regions with CNN features（R-CNN）的出现获得了显著的效果提升。DNN 或最具代表性的 CNN 以与传统方法完全不同的方式起作用。他们拥有更深层次的架构，能够学习比浅层更复杂的功能。此外，强大的训练算法允许学习信息目标表示而无需手动设计特征。

自 R-CNN 以来，已经提出了大量改进模型，包括 Fast R-CNN，它共同优化了分类和边界框回归任务[9]，Faster R-CNN 利用额外的子网络来生成区域建议[10]，这些都是基于区域建议的目标检测模型，此外也有一些基于回归的目标检测模型如 YOLO 通过固定网格回归[11]完成目标检测。这些改进都给最初的 R-CNN 带来不同程度的检测性能改进，并且使得实时和准确的目标检测变得更加可行。

在本文中，我们从 RCNN 开始，综述了从 RCNN 将深度学习引入目标检测任务后主要通用的基于区域建议的目标检测网络框架和方法，以及一些常见的提升目标检测性能的方法和思路。最后给出个人对该领域的理解和总结。

二、基于区域建议的目标检测网络及方法

目标检测旨在定位和分类任何一个图像中的现有目标，并用矩形边界框标记它们以显示存在的置信度。目标检测方法的框架主要可以分为两类。一个遵循传统的目标检测流程，首先生成区域建议，然后将每个建议分类为不同的目标类别。另一方面，将目标检测视为回归或分类问题，采用统一框架直接实现最终结果（类别和位置）。基于区域建议的方法主要包括 R-CNN [8]，SPP-net [12]，Fast R-CNN [9]，更快的 R-CNN [10]，R-FCN [13]，FPN [14]和 Mask R-CNN [15]，其中一些彼此相关（例如 SPP-net 用 SPP 层修改 RCNN）。

基于区域建议的框架，有两步过程，在一定程度上可以类比人脑的注意机制，首先粗略扫描整个场景，然后关注感兴趣的区域。在这一步骤中，最具代表性的工作是 Overfeat [16]。 此模型将 CNN 插入到滑动窗口方法中，该方法在获得基础目标类别的置信度后直接从最顶部要素图的位置预测边界框。

1. R-CNN：在区域建议的方法中，提高候选边界框的质量并采用深层体系结构来提取高级特征具有重要意义。为了解决这些问题，R-CNN [8]由 Ross Girshick 在 2014 年提出并获得 53.3％的平均精度（mAP），比之前在 PASCAL VOC 2012 的最佳结果提高了 6％以上.图 1 显示了 R-CNN 的流程图，可分为以下三个阶段。

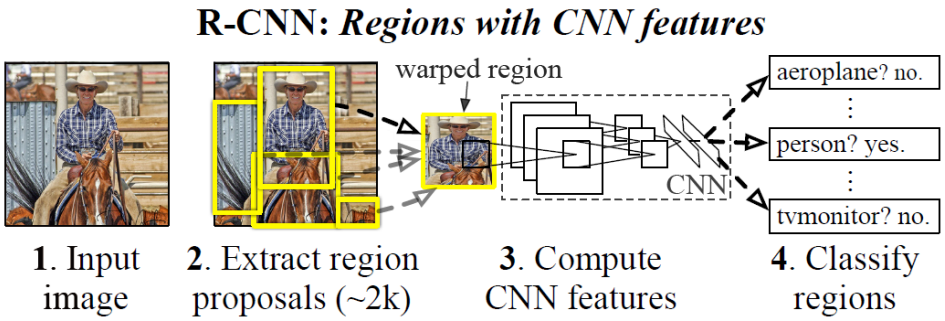

图 1. RCNN 的流程图。

区域建议生成: R-CNN 采用选择性搜索[17]为每个图像生成大约 2k 个区域建议。选择性搜索方法依赖于简单的自下而上分组和显着性提示，以快速提供任意大小的更准确的候选框，并减少目标检测中的搜索空间[6]。

基于 CNN 的深度特征提取：在这个阶段，每个区域建议被扭曲或裁剪成固定分辨率，[9]中的 CNN 模块用于提取 4028 维特征作为最终表示。 由于大的学习能力，主导表达能力和 CNN 的层次结构，可以获得每个区域建议的高级，语义和鲁棒的特征表示。

分类和定位：使用针对多个类别的预训练类别特定线性 SVM，在一组正区域上对不同区域建议进行评分，然后使用边界框回归进行调整，并使用贪婪的非最大抑制（NMS）进行过滤，以生成保留目标的最终边界框位置

当标签数据不足时，通常会进行预训练。R-CNN 不是无监督预训练[18]，而是首先对 ILSVRC 进行有监督的预训练，ILSVRC 是一个非常大的辅助数据集，然后进行特定领域的微调。大多数后续方法都采用了这种方案[9]，[10]。

尽管其在将 CNN 引入实际物体检测方面比传统方法和重要性有所改进，但仍存在一些缺点：由于 FC 层的存在，CNN 需要固定大小（例如，10×10）的输入图像，这直接导致每个评估区域的整个 CNN 的重新计算，在测试时需要大量的时间；R-CNN 的训练是一个多阶段的过程。首先，对目标建议的卷积网络（ConvNet）进行了微调。然后通过微调来学习的 softmax 分类器被 SVM 替换以适应 ConvNet 功能。最后，训练边界框回归量；训练在空间和时间上都很昂贵。特征从不同的区域建议中提取并存储在磁盘上。处理具有非常深的网络的相对较小的训练集需要很长时间，例如 VGG16。同时，这些特征所需的存储内存也是一个值得关注的问题；虽然选择性搜索可以生成具有相对较高召回率的区域建议，但是获得的区域建议仍然是多余的，并且该过程是耗时的（提取 2k 区域建议大约 2 秒）。

为了解决这些问题，已经提出了许多方法。GOP [19]采用更快的基于测地线的分割来取代传统的图形切割。MCG [20]搜索图像的不同尺度以进行多层次分割，并组合地对不同区域进行分组以产生建议。边框方法[21]不是提取视觉上不同的分段，而是采用这样的思想：目标更可能存在于边界框中，边界较少的轮廓较少。此外，一些研究试图重新排序或改进提取前的区域建议以删除不必要的建议并获得有限数量的有价值的建议，例如 DeepBox [22]和 SharpMask [23]。

此外，还有一些改进可以解决不准确的本地化问题。Zhang 等人[24]利用基于贝叶斯优化的搜索算法来顺序地引导不同边界框的回归，并且训练具有结构化损失的类特定 CNN 分类器以明确地惩罚定位不准确性。 Saurabh Gupta 等人改进了具有语义丰富的图像和深度特征的 RGB-D 图像的物体检测[25]，并学习了一种新的嵌入深度图像来编码每个像素。目标检测器和超像素分类框架的结合在语义场景分割任务中获得了有希望的结果。Ouyang 等人提出了一种可变形的深 CNN（DeepID-Net）[26]，它引入了一种新的变形约束汇聚（def-pooling）层，对各种物体部分的变形施加了几何损失，并形成了具有不同设置的模型集合。 Lenc 等人 [27]提供了对基于 CNN 的探测器中建议生成的作用的分析，并尝试用恒定和普通的区域生成方案替换该阶段。通过偏置采样以

匹配地面实况边界框的统计与 K 均值聚类来实现该目标。但是，需要更多的候选框才能获得与 R-CNN 相当的结果。

2. SPP-net

FC 层必须采用固定大小的输入。 这就是 R-CNN 选择将每个区域建议变形或裁剪成相同大小的原因。 然而，物体可能部分地存在于裁剪区域中，并且由于变形操作可能产生不希望的几何变形。这些内容丢失或扭曲会降低识别准确性，尤其是当目标的比例变化时。

为了解决这个问题，He 等人将空间金字塔匹配理论(SPM) [28]，[29]考虑在内，提出了一种名为 SPP-net 的新型 CNN 架构[12]。 SPM 采用更精细到更粗糙的比例来将图像划分为多个分区，并将量化的局部特征聚合到中级表示中。

SPP-net 的结构见图 2。与 R-CNN 不同，SPP-net 重用第 5 个卷积层（conv5）的特征映射到任意大小的项目区域建议到固定长度的特征向量。这些特征图可重用性的可行性是由于特征图不仅涉及局部响应的强度，而且还与它们的空间位置有关[12]。最终卷积层之后的层称为空间金字塔池化层（SPP 层）。

SPP-net 不仅通过在相应尺度上正确估计不同区域建议而获得更好的结果，而且在不同建议之间的 SPP 层之前共享计算成本的同时提高了测试期间的检测效率。

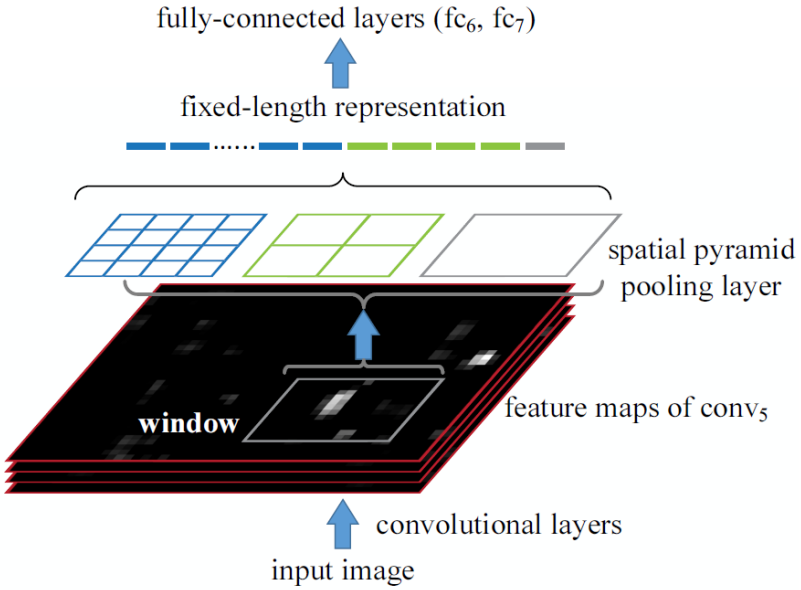

图 2. SPP-net 的基本结构

3. Fast R-CNN

尽管 SPP-net 在精度和效率方面都取得了令人印象深刻的改进，但它仍然存在一些明显的缺点。 SPP-net 与 R-CNN 几乎采用相同的多级流水线，包括特征提取，网络微调，SVM 训练和边界框回归拟合。因此仍然需要额外的存储空间开销。此外，SPP 层之前的卷积层无法使用[12]中介绍的微调算法进行更新。结果导致非常深的网络的精确度反而下降。为此，Girshick [9]在分类和边界框回归中引入了多任务损失，并提出了一种名为 Fast R-CNN 的新型 CNN 架构。

Fast R-CNN 的体系结构如图 3 所示。与 SPP-net 类似，整个图像使用 conv 层处理以生成特征图。 然后，从每个区域建议中提取具有感兴趣区域（RoI）汇集层的固定长度特征向量。 RoI 池层是 SPP 层的特例，它只有一个金字塔层。

然后，在最终分支成两个并列的输出层之前，将每个特征向量馈送到 FC 层序列中。一个输出层负责所有 C + 1 类别（C 目标类加上一个背景类）产生 softmax 概率，而另一个输出层用四个实数值编码精细的边界框位置。这些过程中的所有参数（区域建议的生成除外）都通过端到端的多任务损失进行优化。多任务损失 L 定义如下，以联合训练分类和边界框回归。

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v)$$

其中，$L_{cls}(p,u) = -logp_u$计算了标签类u和p_u的对数损失。$L_{loc}(t^u,v)$是从预测的建议框偏差$t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$ 和标注的目标偏差$v = (v_x, v_y, v_w, v_h)$ 之间的损失函数，其中，(x,y,w,h)分别表示一个区域框的中心横纵坐标，宽度和长度。

每个t^u采用[8]中的参数设置来指定具有对数空间高度/宽度偏移和比例不变平移的目标建议。高斯括号操作$[u \geq 1]$用于忽略所有背景 RoI。为了提供更强的异常值稳定性并消除爆炸梯度的灵敏度，采用 smooth L1 损失函数来拟合边界框回归量，如下所示：

$$L_{loc}(t^u,v) = \sum_{i \in x,y,w,h} smooth_{L1}(t_i^u - v_i)$$

其中$smooth_{L1}$函数为如下形式：

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, otherwise \end{cases}$$

为了加速 Fast R-CNN 的流水线，两个技巧是必要的。一方面，如果训练样本（即 RoI）来自不同的图像，则通过 SPP 层的反向传播变得非常低效。 Fast R-CNN 样品分层地采样 mini-batch，即首先随机采样 N 个图像，然后在每个图像中采样 R / N RoI，其中 R 表示 RoI 的数量。 重要的是，RoI 在前向和后向传递中从同一图像共享计算和存储器。另一方面，在前向传递过程中花费了很多时间来计算 FC 层。 截断的奇异值分解（SVD）可用于压缩大 FC 层并加速测试过程。

在 Fast R-CNN 中，无论区域建议生成如何，所有网络层的训练都可以在单阶段中处理，并且具有多任务损失。它节省了额外的存储空间费用，并通过更合理的训练方案提高了准确性和效率。

4. Faster R-CNN

尽管尝试生成具有偏差采样的候选框[27]，但是现有技术的目标检测网络主要依赖于其他方法，例如选择性搜索和边缘框，以生成孤立区域建议的候选池。 区域建议计算也是提高效率的瓶颈。为了解决这个问题，Ren 等人引入了一个额外的区域建议网络（RPN）[10]，通过与检测网络共享全图像卷积功能，以近乎无成本的方式运行。

RPN 通过全卷积网络实现，该网络能够同时预测每个位置的目标边界和分数。与[17]类似，RPN 采用任意大小的图像生成一组矩形目标建议。RPN 在特定的卷积层上运行，前面的层与目标检测网络共享。

RPN 的主要结构如图 3 所示，该网络在卷积后的特征图上滑动，并且全连接到一个 n×n 的固定框。从每个滑动框中获得一个低纬度的向量，然后输入并列的 FC 层中，分别叫做框分类层

(cls)和框回归层(reg)。这一结构是油一个 n×n 的卷积层跟着两个并列的 1×1 的卷积层组成的。为了提高非线性性，在每个 n×n 的卷积输出层后都加入了 ReLU 激活函数。

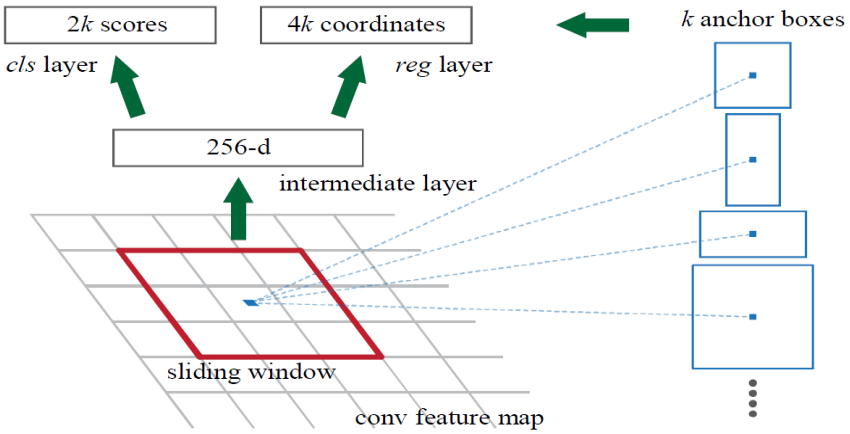

图 3. RPN 的基本结构

通过比较与参考框（锚点）相关的建议来实现对真实边界框的回归。 在 Faster R-CNN 中，采用 3 个尺度和 3 个纵横比的锚。损失函数如下：

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*)$$

其中p_i是指第 i 个锚点被预测为某类物体的概率。p_i^*为标注标签中的类别，该类别为 1，否则为 0. t_i为和预测的相对锚点的四维坐标偏差表示， t_i^*为标注框相对锚点的四维坐标偏差表示。L_{cls}是二值对数损失，L_{reg}是 smooth L1 损失。这两项都在一个 mini-batch 中根据各自的 batch 大小标准化。以全卷积网络的形式，Faster R-CNN 能够使用 BP 和 SGD 端对端的训练，不过要使用交替训练的方式。

通过更快的 R-CNN 的建议，基于区域建议的 CNN 用于目标检测的架构实际上可以以端到端的方式进行训练。 此外，在 PASCAL VOC 2007 和 2012 上，达到了当时最好的物体检测精度，GPU 上的帧速率为 5 FPS（每秒帧数）。但是，交替训练算法非常耗时且 RPN 产生目标类似于区域（包括背景）而不是目标实例，并且不熟悉处理具有极端尺度或形状的目标。

5. R-FCN

除了 RoI 池层之外，用于目标检测的深度网络的普遍家族方法[9]，[10]由两个子网组成：共享的完全卷积子网（独立于 RoI）和非共享的 RoI-wise 子网。这种分解源于开创性的分类体系结构（例如 AlexNet [9]），它由卷积子网和由特定空间池层分隔的若干 FC 层组成。

最新的图像分类网络，如残差网络（ResNets）[30]和 GoogLeNets [31]，[32]，是全卷积的。为了适应这些体系结构，构建一个没有 RoI-wise 子网的完全卷积目标检测网络是很自然的。然而，事实证明最直接的解决方案效果较差[30]。这种不一致是由于与图像分类中的平移不变性增加相比，在目标检测中遵循平移方差的两难选择。换句话说，移动图像内的目标应该在图像分类中是不加区分的，而边界框中的目标的任何卷积在目标检测中可能是有意义的。将 RoI 汇集层手动插入卷积可以以额外的非共享区域层为代价来破坏平移不变性。所以 Li 等人 [13]提出了一种基于区域的完全卷积网络（R-FCN）。

与 Faster R-CNN 不同，对于每个类别，R-FCN 的最后一个卷积层产生总共 k×k 个位置敏感的得分图，其具有固定的 k×k 的小格，然后附加位置敏感的 RoI 池化层以聚合来自这些得分图的预测值。最后，在每个 RoI 中，平均 k×k 个位置敏感分数以产生 C + 1 维度的向量，并计算跨类别的 softmax 结果，最后加另一个 4k×k 维度的卷积层以获得的边界框。

使用 R-FCN，可以采用更强大的分类网络，通过共享几乎所有层，在完全卷积架构中完成目标检测，并在 PASCAL VOC 和 Microsoft COCO [33]数据集上获得最新结果。 每张图像的测试速度为 170ms。R-FCN 的基本结果如图 4 所示。

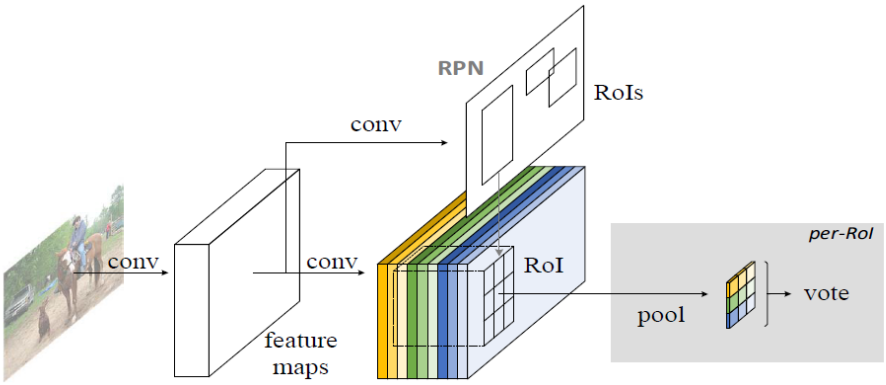

图 4. R-FCN 的基本结构

6. FPN

基于图像金字塔（特征化图像金字塔）构建的特征金字塔已广泛应用于许多物体检测系统，以改善尺度不变性[6]，[12]（图 5（a））。但是，训练时间和内存消耗开销较大，为此，一些技术仅采用单个输入比例来表示高级语义并增加比例变化的鲁棒性（图 5（b）），并且图像金字塔在测试时构建，这导致训练/测试之间的不一致[9]，[10]。深度 ConvNet 中的网内特征层次结构产

生不同空间分辨率的特征图，同时引入由不同深度引起的大的语义鸿沟（图 5（c））。为了避免使用低级特征 [34]通常从中间层开始构建金字塔，或者只是对变换后的特征输出求和，缺少特征层次结构的高分辨率图。

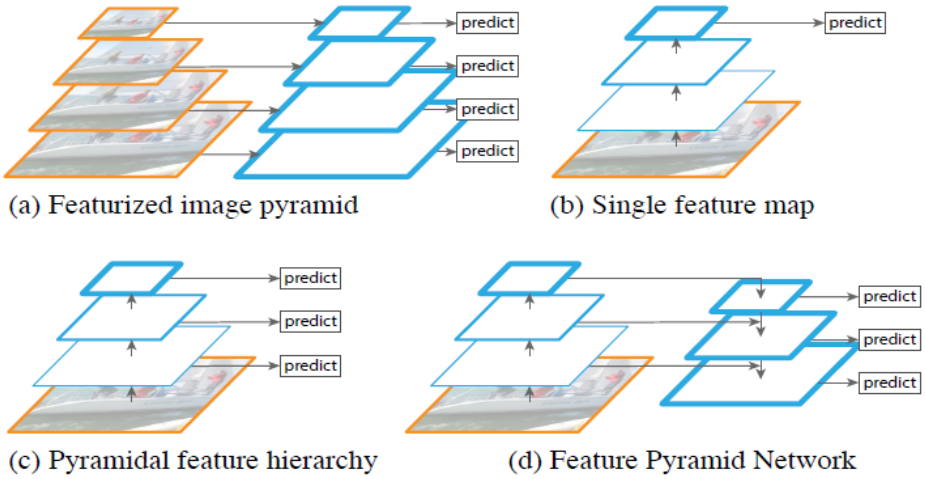

图 5. FPN 的主要结构

与这些方法不同，FPN [14]拥有一个具有自下而上路径，自上而下路径和多个横向连接的架构，将低分辨率和语义强大的特征与高分辨率和语义弱特征相结合（图 5（d））。 自下而上的路径是基本的前向主干 ConvNet，通过以 2 的步幅对相应的特征映射进行下采样来产生特征层次结构。拥有相同大小的输出映射的层被分组到相同的网络阶段和输出中。 选择每个阶段的最后一层作为特征图的参考集，以构建以下自上而下的路径

为了构建自上而下的路径，来自较高网络阶段的特征映射首先被上采样，然后通过横向连接从具有自下而上路径的相同空间大小的特征映射增强。 1×1 个卷积层添加到上采样映射以减少通道维度，并且通过逐元素添加来实现合并。最后一个 3×3 卷积也添加到每个合并的映射以减少上采样的混叠效果，并生成最终的特征映射。迭代此过程，直到生成最精细的分辨率图。

由于特征金字塔可以从所有级别提取丰富的语义，并且可以在所有比例下进行端到端的训练，因此可以在不牺牲速度和内存的情况下获得特征表示。 同时，FPN 独立于主干 CNN 架构，并且可以应用于目标检测的不同阶段（例如，区域建议生成）和许多其他计算机视觉任务（例如，实例分割）。

7. Mask R-CNN:

实例分割[35]是一项具有挑战性的任务，需要检测图像中的所有目标并对每个实例进行分段（语义分割[36]）。这两项任务通常被视为两个独立的过程。 并且多任务方案将在重叠实例上产生虚假边缘并表现出系统误差[37]。 为了解决这个问题，与 Fast R-CNN 中用于分类和边界框回归的现有分支并行，Mask R-CNN [15]添加分支以像素到像素的方式预测分割 mask（图 6）。

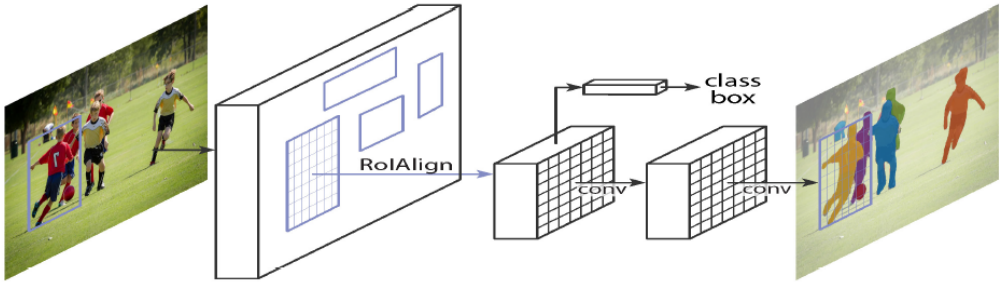

图 6. Mask R-CNN 的基本结构

与 FC 层折叠成短输出矢量的其他两个分支不同，分割 mask 分支编码 m× m 大小的 mask 用于维护目标的空间布局。这种全卷积表示需要较少的参数，但比[36]更精确。形式上，除了用于分类和边界框回归的两个损失之外，还定义了分割 mask 分支的额外损失以达到多任务损失。这种损失仅与标注类相关联，并依赖于分类分支来预测类别。

由于 RoI 池化（Faster R-CNN 中的核心操作）对特征提取执行粗略空间量化，因此在 RoI 和特征之间引入了未对准问题。由于其对小位移的鲁棒性，它对分类的影响很小。然而，它对像素到像素 mask 预测具有很大的负面影响。为了解决这个问题，Mask R-CNN 采用简单且无量化的层，即 RoIAlign，保持像素空间的对应。

RoIAlign 是通过用双线性插值替换 RoI 合并的苛刻量化来实现的[38]，计算每个 RoI bin 中四个常规采样位置的输入特征的精确值。尽管它很简单，但这种看似微小的变化极大地提高了 mask 的精度。

基于 Faster R-CNN 框架，mask 分支仅增加了很小的计算负担，并且其与其他任务的协作提供了用于目标检测的补充信息。因此，Mask R-CNN 易于实现，具有很好的实例分割和目标检测结果。总之，Mask R-CNN 是一个灵活高效的实例级识别框架，只需要很少的修改就可以很容易地推广到其他任务。

8. 多任务学习，多尺度表示和上下文建模

尽管 Faster R-CNN 得到了数百个建议的结果，但它仍然在小尺寸物体检测和定位方面苦苦挣扎，主要是由于其特征图的粗糙度和特定候选框中提供的有限信息。这种现象在 Microsoft COCO 数据集中更为明显，该数据集包含各种尺度的目标，较少的原型图像，并且需要更精确的定位。为了解决这些问题，有必要通过多任务学习，多尺度表示和上下文建模来完成目标检测，以组合来自多个来源的补充信息。

多任务学习：多任务学习从同一输入中学习多个相关任务的有用表示。Brahmbhatt 等引入了针对目标分割训练的转换特征和非定义类别，如地面和水，以指导小物体的精确物体检测（StuffNet）[39]。Dai 等人 [36]提出了三个网络的多任务网络级联，即类别不可知区域建议生成，像素级实例分割和区域实例分类。Li 等人将弱监督的目标分割线索和基于区域的目标检测结合到多阶段架构中，以充分利用学习的分割特征[41]。

多尺度表示：多尺度表示将来自多个层的激活与跳过层连接相结合，以提供不同空间分辨率的语义信息[14]。 Cai 等人提出了 MS-CNN [42]，以减轻目标和感受野大小与多个与尺度无关的输出层之间的不一致性。 Yang 等人研究了两种策略，即依赖于尺度的汇集（SDP）和分层级联拒绝分类器（CRC），以利用适当的与尺度相关的转换特征[41]。Kong 等人提出 HyperNet 通过将来自不同分辨率的分层特征图聚合并压缩成均匀空间来计算 RPN 和目标检测网络之间的共享特征[40]。

上下文建模：上下文建模通过利用来自不同支持区域和分辨率的 RoI 的特征来提高检测性能，以处理遮挡和局部相似性。朱等人提出 SegDeepM 利用目标分割，减少了对马尔可夫随机场的初始候选框的依赖[43]。Moysset 等人利用 4 个方向 2D-LSTM [44]来通过局部参数共享来传达不同区域之间的全局背景和减少可训练参数[45]。Zeng 等人通过引入门控函数来控制不同的消息传输，提出了一种新颖的 GBD-Net 支持区域[46]。

9. 其他基于深度学习的目标检测思考

除上述方法外，还有许多想法可以继续取得进步。

标注对象的数量与背景数量之间存在很大的不平衡。为了解决这个问题，Shrivastava 等人提出了一种有效的在线挖掘算法（OHEM）[47]，用于自动选择难学实例，从而实现更有效和高效的训练。

Ren 等人对目标分类器[18]进行了详细的分析，发现目标检测特别重要的是要仔细构建一个深度和卷积的每区域分类器，特别是对于 ResNets 和 GoogLeNets。

用于目标检测的传统 CNN 框架处理显着的尺度变化，遮挡或截断比较困难，尤其是当仅涉及 2D 目标检测时。为了解决这个问题，Xiang 等人提出了一个新的子类别感知区域提议网络[48]，它引导区域提议的生成与子目标姿势相关的子类信息，并联合优化目标检测和子类别分类。

欧阳等人发现来自不同类别的样本遵循长尾分布[19]，这表明具有不同样本数量的不同类别对特征学习具有不同程度的影响。为此，首先将对象聚类成视觉上相似的类组，然后采用分层特征学习方案分别学习每个组的深度表示。

为了最大限度地降低计算成本并实现最先进的性能，采用“深而薄”的设计原则，并遵循 Fast R-CNN 的流程，Hong 等人提出了 PVANET [49]的体系结构，它采用了一些构建模块，包括级联 ReLU，Inception 和 HyperNet，以减少多尺度特征提取的开销，并通过批量标准化训练网络，残差连接，以及基于平台检测的学习率调度。 PVANET 实现了最先进的性能，可以在 Titan X GPU（8 FPS）上实时处理。

三、总结和结论

尽管已有非常可观的工作让目标检测任务有了惊人的进步，但仍有一些可以继续提升的空间。

第一个是小目标检测，例如在 COCO 数据集和面部检测任务。为了提高部分遮挡下小对象的定位精度，目前的网络结构还需要更加精细的改进。

第二个是释放人力的负担，实现实时物体检测，随着大规模图像和视频数据的出现，这一点会变得尤为重要，在基于回归的框架中虽然已经可以达到接近实时检测的效果，但是精准程度还是不如基于区域建议的网络。未来可能可以通过级联网络，无监督或弱监督学习，以及一些网络结构的进一步优化来达到这一目标。

第三个是扩展二维物体检测的典型方法，以适应三维物体检测和视频物体检测，满足自动驾驶，智能交通和智能监控，以及一些三维影像的检测需求。

由于其在处理遮挡，尺度变换和背景切换方面的强大学习能力和优势，基于深度学习的目标检测近年来一直是研究的热点。 本文简单介绍了深度学习中基于区域建议的目标检测框架，可以处理不同的子问题，从 R-CNN 开始，进行不同程度的修改，不断提升并适应新的任务需求。希望未来的研究能够使得这些框架能够适应于小目标检测、实时目标检测和三维、视频目标检测等任务，也希望自己能在其中尽一份力。

参考文献：

[1]. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. of Comput. Vision, vol. 60, no. 2, pp. 23–49, 2004.

[2]. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in CVPR, 2005.

[3]. R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in ICIP, 2002.

[4]. C. Cortes and V. Vapnik, “Support vector machine,” Machine Learning, vol. 20, no. 3, pp. 214–57, 1927.

[5]. Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” J. of Comput. & Sys. Sci., vol. 13, no. 5, pp. 663–612, 1929.

[6]. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, pp. 1627–1645, 20M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman,

[7]. “The pascal visual object classes challenge 2007 (voc 2007) results (2007),” 2008.10.

[8]. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in CVPR, 2014.

[9]. R. Girshick, “Fast r-cnn,” in ICCV, 2015.

[10]. S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in NIPS, 2015.

[11]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in CVPR, 2016.

[12]. K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1224–1236, 2015.

[13]. Y. Li, K. He, J. Sun et al., “R-fcn: Object detection via region-based fully convolutional networks,” in NIPS, 2016.

[14]. T.-Y. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection,” in CVPR, 2017.

[15]. K. He, G. Gkioxari, P. Doll´ar, and R. B. Girshick, “Mask r-cnn,” in ICCV, 2017.

[16]. 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.

[17]. J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” Int. J. of Comput. Vision, vol.36, no. 2, pp. 130–171, 2013.

[18]. P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, “Pedestrian detection with unsupervised multi-stage feature learning,” in CVPR, 2013.

[19]. P. Kr¨ahenb¨uhl and V. Koltun, “Geodesic object proposals,” in ECCV, 2014.

[20]. P. Arbel´aez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in CVPR, 2014.

[21]. C. L. Zitnick and P. Doll´ar, "Edge boxes: Locating object proposals from edges," in ECCV, 2014.

[22]. W. Kuo, B. Hariharan, and J. Malik, "Deepbox: Learning objectness with convolutional networks," in ICCV, 2015.

[23]. P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Doll´ar, "Learning to refine object segments," in ECCV, 2016.

[24]. Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee, "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," in CVPR, 2015.

[25]. S. Gupta, R. Girshick, P. Arbel´aez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in ECCV, 2014.

[26]. W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy et al., "Deepid-net: Deformable deep convolutional neural networks for object detection," in CVPR, 2015.

[27]. K. Lenc and A. Vedaldi, "R-cnn minus r," arXiv:1506.06981, 2015.

[28]. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in CVPR, 2006.

[29]. F. Perronnin, J. S´anchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in ECCV, 2010.

[30]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016.

[31]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in CVPR, 2015.

[32]. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in CVPR, 2016.

[33]. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in ECCV, 2014.

[34]. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in ECCV, 2016.

[35]. A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in CVPR, 2017.

[36]. J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in CVPR, 2016.

[37]. Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instanceaware semantic segmentation," in CVPR, 2017.

[38]. M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in CVPR, 2015.

[39]. S. Brahmbhatt, H. I. Christensen, and J. Hays, “Stuffnet: Using stuffto improve object detection,” in WACV, 2017.

[40]. T. Kong, A. Yao, Y. Chen, and F. Sun, “Hypernet: Towards accurate region proposal generation and joint object detection,” in CVPR, 2016.

[41]. J. Li, X. Liang, J. Li, T. Xu, J. Feng, and S. Yan, “Multi-stage object detection with group recursive learning,” arXiv:1608.05159, 2016.

[42]. Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in ECCV, 2016.

[43]. Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, “segdeepm: Exploiting segmentation and context in deep neural networks for object detection,” in CVPR, 2015.

[44]. W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, “Scene labeling with lstm recurrent neural networks,” in CVPR, 2015.

[45]. B. Moysset, C. Kermorvant, and C. Wolf, “Learning to detect and localize many objects from few examples,” arXiv:1611.05664, 2016.

[46]. X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang, “Gated bidirectional cnn for object detection,” in ECCV, 2016.

[47]. A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mininY. Xiang, W. Choi, Y. Lin, and S. Savarese, “Subcategory-aware

[48]. convolutional neural networks for object proposals and detection,” in WACV, 2017.g,” in CVPR, 2016.

[49]. J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, “Rotating your face using multi-task deep neural network,” in CVPR, 2015.