基于卷积神经网络的目标检测综述

李 同，阮士峰，陈 卓，毛珍珍
（安阳工学院，河南 安阳 455000）

[摘要] 随着时代的发展，传统的目标检测效率和精度逐渐落后于人们的需求。传统的目标检测方法大多基于人工设计的特征，精度和鲁棒性较差。随着计算机性能瓶颈的突破，深度学习技术得到快速发展。通过对图像进行卷积处理，卷积神经网络能够提取优于人手工设计的图像特征。基于卷积神经网络的目标检测算法与传统算法相比，准确性和检验速度均有所提高，鲁棒性更强，在目标检测领域获得了广泛的关注。但不同的卷积神经网络算法结构存在着较大的差异，不同的结构在检验精度和速度上大不相同。从整体上看，以卷积神经网络为主的目标检测算法已成为主流。

[关键词] 目标检测；卷积神经网络；深度学习；计算机视觉
[中图分类号] TP389.1 **[文献标识码]** A **[文章编号]** 2096-1995(2020)27-0018-03

目标检测是一种应用特定计算机算法在图像中找到所需目标的技术[1]。结合目标的定位和分类，实现了更高的效率。随着时代的发展，计算机硬件已不再是限制发展的瓶颈，目标检测算法产生了巨大的突破，越来越多地应用于交通检测、智能支付、医疗影像等各个方面。

传统目标检测方法大多采用人工设计的特征进行检测，大致分为四个步骤：

使用不同大小的滑动窗口在给定图片上建立多个候选框；

采用不同的特征提取方式，将候选框转化为特征向量；

再采用基于上述特征的 SVM 等分类器对分类特征向量[2]；

利用非极大值抑制（NMS）消去冗余的候选框[3]。

传统的目标检测办法存在较多的缺陷，如滑动窗口在候选框的选择上采取枚举法，耗费大量的时间，候选框重复率高；人手工标记的特征鲁棒性较差，难以适应复杂的检测环境等。因此，随着深度学习的逐渐发展，卷积神经网络逐渐应用在目标检测领域，以卷积神经网络为基础的模型逐渐替换传统目标检测模型，进入大众的视野。

1 卷积神经网络在目标检测的发展

深度学习具有许多算法，卷积神经网络（Convolutional Neural Networks, CNN）是其代表算法之一[4][5]。它一个具有卷积计算和深度结构的前馈神经网络。

卷积神经网络可以追溯到 1962 年。在对猫大脑中视觉系统的研究中，Hubel 与 Wiesel 提出了感受野（Receptive fields）的概念[6]。这个结果对视觉系统中信息处理领域做出了巨大的贡献。福岛邦彦在 1980 年提出了具有卷积和池化层的神经网络结构[7]。在此基础上，LeNet-5 被提出，这是一个具有 2 个卷积层、2 个池化层和 3 个全连接层的网络，这是卷积神经网络的雏形[8]。直到 2012 年，ImageNet 图像识别大赛中，Hinton 组使用的 AlexNet[9] 成功颠覆了图像识别领域，使得基于卷积神经网络的目标检测模型被大家认可。

2 基于卷积神经网络的目标检测算法

卷积神经网络通过学习手动标记特征的数据集来获得目标的特征[10]。目前，基于卷积神经网络的算法大致可分为两种模式，即 two-stage 模式和 one-stage 模式，前者的检测过程分为两个步骤：首先由算法生成若干个候选框，再通过 CNN 对候选框进行分类；后者则是对目标的类别概率和位置坐标直接回归，相对来说精度有所损失，但速度较 two-stage 模式的算法更快。

2.1 基于 two-stage 的算法

2.1.1 R-CNN

2014 年 Ross B. Girshick 等人在基于 CNN 设计了 R-CNN 模型，模型采用了基于 AlexNet 的网络层结构[11]。在区域推荐（region proposal）上，R-CNN 算法通过选择性搜索（selective search）[12] 来确定候选框，该方法通过选择搜索在图像上确定不同形状和大小的约 2000 个候选框；为了使候选框能够输入到 CNN 中进行特征提取，统一将候选框压缩到 227 × 227 大小；然后运用 CNN 对候选框进行特征提取；最后使用多个支持向量机（SVM）分类器[13] 分类输出向量，采用边界回归生成目标区域。整个测试过程如图 1 所示。

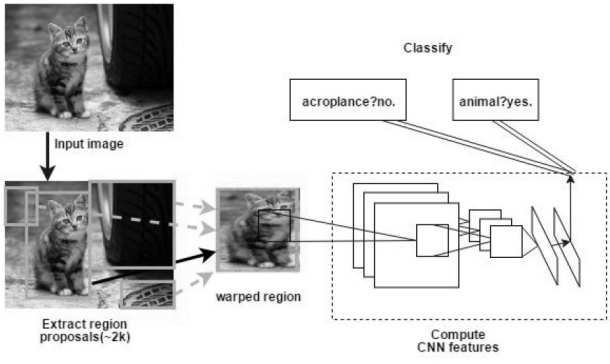

图 1 R-CNN 的测试步骤

R-CNN 的应用使得在数据集 PASCAL VOC2012 的结果达到了 53.3%[11]，较之前最佳结果提升了三成。R-CNN 的出现已经成功地将 CNN 应用到目标检测领域，但 R-CNN 也存在着一定问题：

在提取候选框时效率低，消耗存储空间较多；

在输入网络前需要对候选框进行归一化，有可能导致输入 CNN 的信息缺失；

所有候选框都要进入 CNN 计算，并且有大量计算重叠。

2.1.2 Fast R-CNN

由于卷积神经网络的全连接层对输入图像的尺寸有着一定的要求，所以需要对 R-CNN 中的候选框进行归一化，无论是采用剪切还是变形，都很难保留图像的完整信息，对于高像素图像，则很容易造成构成缺失和模糊等问题。正是因为这个原因，何凯明等人于 2015 年提出了 SPPNet[14]。SPPNet 解决了 R-CNN 需要重复提取候选区域并使候选框变形的问题，但

依然需要分多步骤训练，占用大量磁盘空间。Fast R-CNN 在吸收 SPPNet 优势的同时改进了 R-CNN[15]。Fast R-CNN 仍然使用选择性搜索来确定候选框，但 Fast R-CNN 将整张图片输入到 CNN，在卷积特征层上使用用感兴趣区域（Region of interest pooling，ROI pooling）操作，并从特征图中提取一个特定长度的特征向量；然后将特征向量输入到全连接层，用 softmax 对其进行分类；最后对属于同一特征的候选框进行分类并回归其位置。

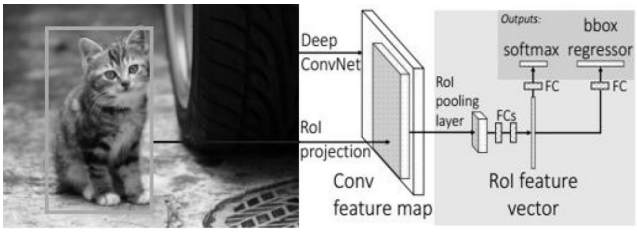

图 2 Fast R-CNN

Fast R-CNN 创新性地对将整张图片进行特征提取，提升了算法运行的速度，同时使用 ROI 池化操作使每一个候选框生成固定大小的特征图，免去了归一化的步骤，也降低了对磁盘的占用量。但 Fast R-CNN 依旧没有解决使用选择性搜索方法占用时间的问题。缓慢的候选框提取速度与后续的计算速度不对等，造成计算资源浪费，难以满足实时应用需求，并没有实现真正的端到端。

2.1.3 Faster R-CNN

从 R-CNN 模型开始，通过引入卷积神经网络，新的算法获得了高于传统算法的检测速度，但由于传统区域候选算法的限制，很难提高检测速度。为了突破候选区域算法的时间瓶颈，任少卿等人在 2016 年提出了 Faster R-CNN[16]。Faster R-CNN 使用 RPN 而不是选择性搜索，大大减少了提取候选框的时间。该算法可大致理解为将 RPN 和 Fast R-CNN 相结合，首先提取整张图片的特征；再将特征结果输入到 RPN；然后使用 ROI 池化层固定候选框的大小；最后对属于某一特征的候选框回归和调整。

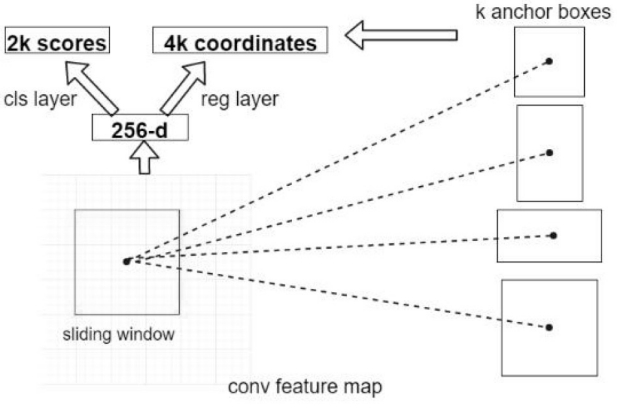

图 3 RPN 结构

Faster R-CNN 的结构实现了端到端，减少了确定候选框的时间，可以满足实时需求。从 R-CNN 到 Faster R-CNN，其检测速度和检测精度都不断在提高，此类算法至今依然是目标检测算法的重要分支。

2.2 基于 one-stage 的算法

2.2.1 YLOL v1

与基于候选区域的 two-stage 目标检测模型相比，基于回归方法的 one-stage 目标检测模型免去了候选框提取的步骤，而是直接在特征图是采取了回归方法。

YOLO 系列算法是一种典型的 one-stage 算法[17]，它的核心思想是将完整图片输入网络，在输出层对边界框（bounding box，BBox）所属的种类和 BBox 的坐标进行回归。YOLO v1 的算法具有一个独立的 CNN 模型，从而实现端到端的检测，其实现方法为先将图片分为 5×5 个网格单元，再将整张图片输入 CNN，最后通过整理 CNN 预测结果获得检测的目标。由于 YOLO v1 算法具有统一的框架结构，与 R-CNN 系列算法相比，其速度更快，训练过程更加简洁。

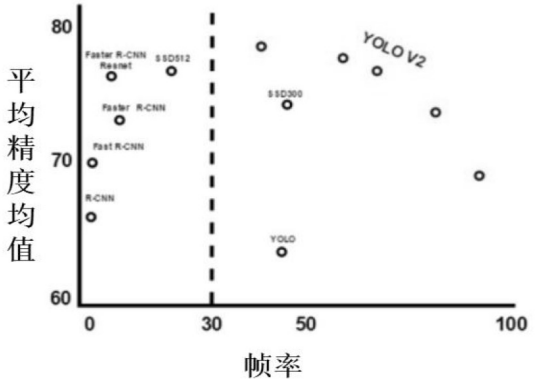

图 4 Yolo 在 PASCAL VOC 2007 上的对比[18]

YOLO v1 的效率虽然很高，检测速度很快，但 YOLO v1 的网格单元只预测两个从属于一个类别的边界框，对于尺度较小的物体，YOLO 的表现往往不尽人意，同时，YOLO v1 对物体的形变方面的泛化比率较低，很难定位比例不同寻常的物体，这也是 YOLO 的一个缺陷。

2.2.2 YOLO 改进算法

基于 YOLO v1 存在的问题，已经产生了许多后续的改进算法，如 YOLO 9000、YOLO v2、YOLO v3。

YOLO-9000 大幅提高了算法可以识别目标的类别数量。YOLO v2 则针对 YOLO v1 的部分内容进行了改进。由于 YOLO v1 在预训练阶段和检测阶段分别采用了 224×224 和 448×448 的分辨率，导致模型需要适应分辨率的改变。为了解决这个问题，YOLO v2 将预训练分为了两步。先用 224×224 的输入训练 160 个循环，再改用 448×448 的输入训练 10 个循环。经过微调后，新的模型可以在 448×448 的图像上顺利过渡。YOLO v2 在 VOC2007 上 mAP 为 78.6%，高于 Faster R-CNN[18]。2018 年，Redmon 等人提出的 YOLO v3[19] 则在泛化能力上和较小目标的检测上进一步提高，检测速度和精度均有一定的保障。

YOLO 系列算法是目前一种先进的目标检测算法。因为整个检测框架是一个整体，所以可以端到端地对算法的性能进行优化。同时，因为其优异的性能而被广泛使用，不断产生新的衍生算法。

2.3 two-stage 和 one-stage 的对比

由于结构的差异，two-stage 和 one-stage 算法在检测速度上和精度上也存在着较大的差异。Microsoft COCO 数据集是目前计算机视觉领域最权威的数据集之一。作为大型数据集，它

具有精确的物体分割和重要的日常检测价值[20]。

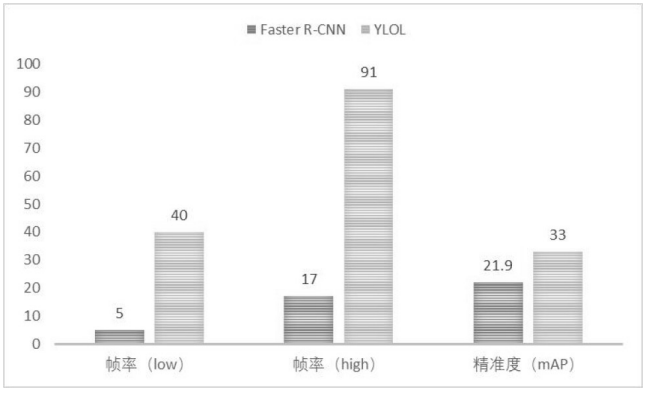

图 5 Microsoft COCO 上的检测结果

Faster R-CNN 和 YOLO v3 分别作为两种结构的代表算法，在 Microsoft COCO 数据集上的准确度和检验速度如图 5 所示。得益于其一体化的网络结构，YOLO 算法具有显著高于 Faster R-CNN 的速度；在准确度上，YOLO 略高于 Faster R-CNN。但在尺度较小的目标检测中，如图 6，在对不同大小的目标进行检测的过程中，由于 YOLO 每个网格单元只能回归同一种类的两个边界框，导致其对小目标的检测检测精度远低于 Faster R-CNN 算法。

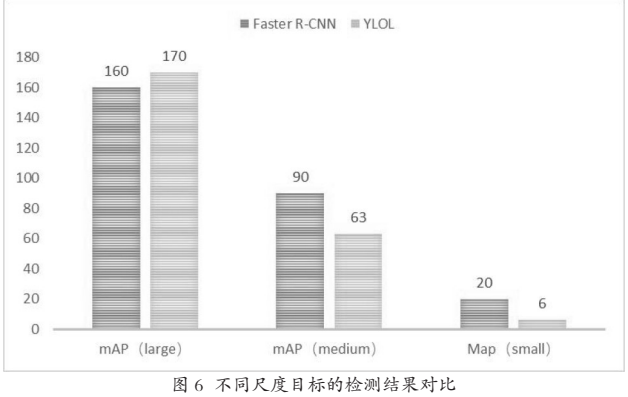

图 6 不同尺度目标的检测结果对比

3 结语

近年来，目标检测领域高速发展，算法改进方面的研究也在稳步进行，新的成果层出不穷。但整体来说，单一的目标检测算法并不能应用于所有检测环境，各种算法都具有不同的发挥空间。通过将不同的算法进行结合，能获得单一算法所不具备的效率。当前卷积神经网络已成为主流，但是传统算法依然具有一定的价值，能对新的目标检测算法进行启发。

【参考文献】

[1] 田合雷，丁胜，于长伟，等 . 基于目标检测及跟踪的视频摘要技术研究 [J]. 计算机科学，2016，43(11)：297-299.
[2]Ying Z, Li B, Lu H, et al. Sample-Specific SVM Learning for Person Re-identification[C]//IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas, USA: IEEEComputerSociety,2016: 1278-187.
[3] 王静，王海亮，向茂生，等 . 基于非极大值抑制的圆目标亚像素中心定位 [J]. 仪器仪表学报，2012(07)：22-30.
[4]Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G. and Cai, J., 2015. Recent advances in convolutional neural networks. Ar Xiv preprint arXiv:1512.07108.
[5]Goodfellow, I., Bengio, Y., Courville, A..Deep learning(Vol. 1).Cambridge：MIT press，2016：326-366
[6]Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex. J. Physiol. 160, 106-154(1962).
[7]Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernetics 36, 193-202(1980).
[8]LECUN, Y., BOTTOU, L., BENGIO, Y., & HAFFNER, P..(1998).Gradient-Based Learning Applied to Document Recognition. 86(11), 2278-2324.
[9]Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
[10]Cireşan, Dan, Meier U , Schmidhuber J . Multi-column Deep Neural Networks for Image Classification[J]. Eprint Arxiv, 2012.
[11]Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Columbus, USA: IEEE Computer Society, 2014: 580-587.
[12]Uijlings J R R , K. E. A. van de Sande…. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.
[13]L é on Bottou, Lin C J . Support vector machine solvers[J]. Large Scale Kernel Machines, 2013:1-27.
[14]K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014.
[15]Girshick R . Fast R-CNN[J]. Computer Science, 2015.
[16]Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015,39(6):1137-1149.
[17]Redmon J, Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Computer Society,2016:429-442.
[18]Redmon J, Farhadi A.YOLO9000: Better, Faster, Stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE ComputerSociety, 2017: 6517-6525.
[19]Redmon J, Farhadi A.YOLOv3: An Incremental Improvement[J].2018.
[20]Belongie S. "Microsoft COCO: Common Objects in Context",[J]. 2014.