基于深度学习的目标检测算法综述

吴雨露 张德贤

（河南工业大学 信息科学与工程学院，河南 郑州 450001）

摘 要：计算机网络和人工智能快速发展的时代，人身安全、社会安全以及国家安全越来越受到大众的关注。目标检测在视频处理中发挥至关重要的作用。传统目标检测算法已难以满足目标检测中数据处理效率、性能、智能化等方面的要求。当前流行的深度学习广泛应用于人工智能和目标检测与跟踪。基于此，介绍 SPPNet、R-CNN 等一系列基于区域提案（Region Proposal）的目标检测方法和 YOLO、SSD 等基于回归的目标检测方法及优缺点，总结与展望目标检测的未来。

关键词：目标检测；R-CNN；YOLO；深度学习

中图分类号：TP18；TP391.41 **文献标识码：**A **文章编号：**1003-9767（2019）12-046-03

A Survey of Target Detection Algorithms Based on Deep Learning

Wu Yulu, Zhang Dexian

(School of Information Science and Engineering, Henan University of Technology, Zhengzhou Henan 450001, China)

Abstract: In the era of rapid development of computer network and artificial intelligence, people pay more and more attention to personal security, social security and national security. Target detection plays an important role in video processing. Traditional target detection algorithms have been difficult to meet the requirements of data processing efficiency, performance and intelligence in target detection. The current popular deep learning is widely used in artificial intelligence and target detection and tracking. Based on this, this paper introduces a series of target detection methods such as SPPNet, R-CNN based on region Proposal, YOLO, SSD based on regression and their advantages and disadvantages, and summarizes and looks forward to the future of target detection.

Key words: object detection; R-CNN; YOLO; deep learning

0 引言

目前，计算机视觉领域的目标检测技术逐渐成熟发展。2013 年 Girshick 等人提出 R-CNN 框架后，在此基础上的一系列框架被不断提出，在网络结构上可以分为 one stage 框架和 two stage 框架。two stage 包括 R-CNN 系列框架，one stage 包括 YOLO、SSD 等系列框架[1]。one\two stage 区别在于，前者是输入图像后在卷积神经网络中提取特征，预测物体分类和位置，后者是输入图像后在网络中生成多个候选框（Proposal Boxes），进行更精细的检测，找到待测物体的位置并分类。

1 基于 Region Proposal 的 two stage 检测算法

1·1 R-CNN

Girshick 等人首次提出的第一个基于深度学习的目标检测方法 R-CNN[2]，经过实验验证，（VOC07 数据集）mAP 由原来的 34·3% 提升到 66%。算法步骤：（1）采用选择性搜索（SS）算法，在一张图像中生成多个候选框（一般 2 000 个左右）；（2）每个候选框的特征图缩放到相同大小，使用卷积神经网络（CNN）提取特征；（3）将提取出的特征送入 SVM 分类器，判别是否属于一个特定的类；（4）使用回归器进一步调整候选框位置。

根据实验结果，R-CNN 能使 mAP 大幅度提升，但其他方面还有许多不足。例如，时间上，由于众多候选框，卷积神经网络需要逐个向前传播提取特征，导致运算量庞大、效率低；空间上，训练步骤费时且占用过多资源。由于输入的候选区域必须是固定大小，因此经过 crop 或 warp 操作后，裁剪区域可能不包含整个对象，而扭曲内容可能导致几何失真，影响检测精确度。对于这些问题，何恺明等人提出空间金字塔池化网络（SPP-Net）进行解决[3]。

1·2 SPP-Net

金字塔池化网络与 R-CNN 相比，优点包括三方面。第一，

作者简介：吴雨露（1994—），女，河南偃师人，硕士研究生。研究方向：人工智能信息处理。
张德贤（1961—），男，河南密县人，博士研究生，教授。研究方向：人工智能信息处理。

是基于多尺度的目标检测框架，多尺度在于和 R-CNN 相比，在最后一个卷积层和全连接层之间增加一个 SPP 层，SPP 使用多级空间区间，而滑动窗口池仅使用单个窗口。第二，更深层次网络中，由于输入尺度的灵活性，SPP 可集合不同尺度下提取的特征，避初始就 crop 和 warp。第三，不论输入图像大小，SPP 都能生成固定长度输出。

算法步骤：（1）采用选择性搜索（SS）算法在一张图像中生成多个候选框（一般 2 000 个左右）；（2）将图像输入到 CNN 中，对含有多个候选框的整张图片进行特征提取（不是 R-CNN 中的逐个提取而是一次性提取），得到特征图后，在特征映射的每个候选框进行金字塔空间池化，得到固定大小的特征向量（一般使用四级空间金字塔 1×1、2×2、3×3 和 6×6）；（3）和（4）与 R-CNN 步骤相同。

由实验可知，SPP-Net 的计算速度比 R-CNN 快 24 ～ 102 倍，且具有更好的精度。在此基础上，何等人提出了一种基于 CNN 分类精确度的重要策略——模型组合，用于进一步改进 SPP-Net。使用两个结构相同初始化不同的 SPP-Net，对候选窗口执行非最大抑制 [4]，使更强的一方抑制稍弱的一方。结果表明，此方法使 mAP 提高到 60·9%，模型互补比单个模型更有效。但是，上述方法都存在重复计算候选框消耗时间的问题。

1·3 Fast R-CNN

Girshick 等人提出 Fast R-CNN[5]，在 R-CNN 的基础上引入目标区域池化（ROI）。这实际上是一个单层金字塔池化层解决候选框重复计算的问题。Fast R-CNN 能根据输入的图像适当选取 Proposals（或多或少都会降低 mAP），训练过程中用多任务训练代替分阶段训练，将分类与回归任务同时进行，使用“蛮力学习”目标检测，即单尺度目标检测，扩充数据集（VOC10 和 VOC12），以提供更多训练数据提高训练精度，采用 Sofemax 分类器代替 SVM 分类器，用截断的 SVD 把全连接层压缩成两个没有线性关系的全连接层。

流程步骤：（1）输入一张图像，用滑动窗口查找图像中的对象，利用 SS 算法提取区域候选框大约 2 000 个；（2）利用 CNN 进行特征提取，投影到最后的特征层；（3）对特征层上的候选框进行单一金字塔池化操作，得到固定大小的特征表示；（4）通过两兄弟层，即两个全连接层（分类与回归），用 Sofemax 分类器分类，用边界回归模型进行微调。

在 VOC07 训练集上训练，Fast R-CNN 的 mAP 提高到 66·9%，之后用 VOC12 训练集扩充 VOC07 训练集，得到 VOC07+12 训练集，测试结果为 70%。由此可知，Fast R-CNN 提高了检测器的检测质量，但由于使用 SS 算法依旧花费太多时间，因此任等人在 2016 年提出了 Faster R-CNN 算法，提高了检测速度。

1·4 Faster R-CNN

Faster R-CNN 的主要方法是滑动窗口，核心是区域提案网络（Region Proposal Network，RPN）[6]，是全卷积神经网络。与 Fast R-CNN 不同，Faster R-CNN 通过四步训练法交替优化，学习 RPN 与基于区域对象检测 R-CNN 之间的共享 conv 层特征，最终形成两个网络共享的 conv 层。由于 Faster R-CNN 训练过程中各个任务互相配合、共享参数，因此大大缩短了检测时间。

具体步骤：（1）通过 CNN 对输入的任意大小的图像进行特征提取，输出一组矩形 Object Proposal；（2）在与 RPN 形成的共享 conv 层的最后一层的滑动网络滑动一遍，利用 K 个不同大小的锚（Anchor）定位滑动窗口（通常 K=9），从而预测 Region Proposal，生成特征区域候选框；（3）在 reg（盒回归层）和 cls（盒分类层）上进行回归和分类，每个锚都分配一个二进制类标签（an object or not）进行分类，训练一组 K 个边界回归模型，每个回归模型负责一个不同比例和纵横比的锚盒（Anchor Boxes），对每个候选框位置进行微调后再分类。

在 VGG16 网络中用 VOC07+12 训练集中的数据进行训练，测试结果 mAP 比 Fast R-CNN 高 3·2%，测试时间为 198 ms，减少了将近 10 倍。RPN 的提出，实现了高效准确的 Region Proposal 生成。通过与下游检测网络共享卷积特征，Region Proposal 几乎没有花费时间，使检测系统能够以 5 ～ 17 fps 的速度运行，提高了检测速度。此算法虽然检测速率有所提高，但是不具有实时性，实际运用希望检测算法能够实时在线检测目标，避免造成太大损失。

2 基于回归的 one stage 检测算法

以上方法都是先输入图像提取特征，然后用分类器和定位器识别特征空间中的对象。但是，这种检测方法不具有实时性，检测速度慢。以 YOLO 为代表的一步检测法，显著提高了检测速度，具有良好的实时检测效果。

2·1 YOLO

Redmon 等人提出的 YOLO 算法延续了 GoogleNet 模型的核心思想 [7]，将目标检测作为回归问题处理，直接在划分的网格上回归目标边界框和所属类别。其结构简单，算法为输入一张图像，缩放为统一大小，划分 S×S 个网格（S=7），若待检测对象中心落入网格中，则对其进行检测分类。

YOLO 与几种常见的检测框架（DPM、R-CNN、OverFeat 和其他快速检测器）相比，需要改进的地方有三方面。第一，一步检测。与以往优化检测管道相比，YOLO 完全抛弃了检测管道，其本身是一个完整的检测系统。YOLO 使用单个卷积神经网络同时执行特征提取、边界框预测、非最大抑制和上下文推理。第二，减少候选框。网络单元 proposal 设置了空间限制，每个图像提取的候选框只有 98 个，提高了检测效率。第三，YOLO 通用性好，不需要多个检测器，可以同时检测各种物体。

在 PASCAL VOC07 和 VOC12 数据集上比较实验结果，

YOLO 的 mAP 达到 63·4%，检测速度 45 fps，Fast YOLO 的 mAP 为 52·7%，检测速度 155 fps。与其他实时检测器相比，Fast YOLO 速度最快，YOLO 的 mAP 最高。由于其是训练整个图像，因此检测精度有所下降。YOLO 存在的局限性还有两方面。第一，对小目标检测率低。因为对边界框预测施加了强大的空间约束，所以限制模型预测邻近对象的数量，例如成群出现的鸟类，这些小物体 YOLO 很难检测。第二，不正确的定位。训练一个近似于检测性能的损失函数时，损失函数对大边界框和小边界框的错误处理相同，但大盒中的小错误没太大影响，小盒中的小错误对 IOU 有很大影响。

针对 YOLO 难定位的缺点，Redmon 等人将定位误差小、背景检测误差大的 Fast R-CNN 与定位误差大、背景检测误差小的 YOLO 结合训练，测试结果与 Fast R-CNN 相比，mAP 提升了 3·2%。VOC12 的测试结果与 YOLO 相比，mAP 提升了 12·8%。Redmon 等人针对检测精度下降问题，在 2016 年提出了 YOLO9000 算法 [8]，在保证原本 YOLO 算法速度的基础上改进提高检测精确度。YOLOV2 算法改进网络结构，采用降维采样的方法进行动态调整，以预测不同大小的图片，使检测精度与速度达到平衡。

2·2 SSD

Liu 等人提出的 SSD 是单层深度神经网络，能够应用于多类别对象检测。其核心是使用小卷积滤波器，预测特征图中固定的一组默认边界框的类别 scores 和框偏移量 [9]。SSD 模型为了提高检测速度与精度，在基础网络中添加了辅助结构，包括多尺度特征图、卷积预测因子、默认框与宽高比。其中，默认框（Default Boxes）类似于 Faster R-CNN 中的锚盒（Anchor Boxes），可以在不同比例的多个特征图上的每个特征位置使用不同宽高比的默认框。与基于滑动窗口和基于 Region Proposal 分类的目标检测方法相比，SSD 没有使用 Proposal 步骤，而是使用默认框。因此，这种一步检测方法更加灵活。在 Pascal VOC07 数据集上训练 SSD 与其他模型的比较结果如表 1 所示。

表 1 Results on Pascal VOC07 test

Method	mAP	FPS	Batch size
Faster R-CNN(VGG16)	73·2	7	1
Fast YOLO	52·7	155	1
YOLO(VGG16)	66·4	21	1
SSD300	74·3	46	1
SSD512	76·8	19	1
SSD300	74·3	59	8
SSD512	76·8	22	8

由表 1 可以看出，SSD 结合了 YOLO 的检测速度和 Faster R-CNN 的检测精度。SSD512 在检测精度上优于其他方法，达到 76·8%，且检测速度基本达到实时检测的要求。虽然 Fast YOLO 检测速度最快，但是其检测精度最低。因为 SSD 没有像 Faster R-CNN 中的特征重采样步骤，所以小目标分类相对困难。数据训练中，增强小目标精度数据可以提高 SSD 的性能。将图像随机放在原始图像大小为 16 倍的画布上，其中填充平均值，进行随机裁剪操作，可以扩展数据，获得更多训练图像，达到数据增强的目的。与 YOLO 相比，SSD 提高了小目标检测精度。

3 结 语

当前基于深度学习的目标检测算法与传统检测算法相比，性能有很大改善。虽然检测速度与精度大幅度提高，但是仍然存在一些未解决的问题，比如数据集不完整、针对小目标检测的数据较少、检测性能不高，未能实现不降低检测精度前提下的实时在线检测。目前，社会国家等各方面的发展中，目标检测是一大课题，实时在线检测是重中之重，值得人们继续研究发展。

参考文献

[1] 郑伟成 , 李学伟 , 刘宏哲 · 基于深度学习的目标检测算法综述 [C]·// 中国计算机用户协会网络应用分会 2018 年第二十二届网络新技术与应用年会 ,2018·

[2] 周晓彦 , 王珂 , 李凌燕 · 基于深度学习的目标检测算法综述 [J]· 电子测量技术 ,2017,40(11):89-93·

[3]He K,Zhang X,Ren S,et al·Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]· IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1·

[4]Bailo O,Rameau F,Joo K,et al·Efficient Aadaptive Nnon-maximal Suppression Algorithms for Homogeneous Spatial Keypoint Distribution[J]·Pattern Recognition Letters,2018,106:53-60·

[5]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 Intelligen ce,2015,39(6):1137-1149·

[6]Redmon J,Divvala S,Girshick R,et al·You Only Look Once:Unified, Real-Time Object Detection[C]·// Proc of IEEE Conference on Computer Vision and Pattern Recognition,2016·

[7]Redmon J,Farhadi A·YOLO9000:Better,Faster,Stronger [C]·// Proc of IEEE Conference on Computer Vision and Pattern Recognition,2017·

[8]Liu W,Anguelov D,Erhan D,et al·SSD:Single Shot Multibox Detector[C]·// Proc of European Conference on Computer Vision·,2016·

[9]Everingham M,Winn J·The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit[J]· International Journal of Computer Vision,2006,111(1):98-100·