基于深度学习的通用目标检测研究综述

程 旭[1],宋 晨[1],史金钢[2],周 琳[3],张毅锋[3],郑钰辉[1]
(1. 南京信息工程大学计算机与软件学院,江苏南京 210044;
2. 西安交通大学软件学院,陕西西安 710049;
3. 东南大学信息科学与工程学院,江苏南京 210096)

摘 要: 目标检测是计算机视觉领域中最基础且最重要的任务之一,是行为识别与人机交互等高层视觉任务的基础.随着深度学习技术的发展,目标检测模型的准确率和效率得到了大幅提升.与传统的目标检测算法相比,深度学习利用强大的分层特征提取和学习能力使得目标检测算法性能取得了突破性进展.与此同时,大规模数据集的出现及显卡计算能力的极大提高也促成了这一领域的蓬勃发展.本文对基于深度学习的目标检测现有研究成果进行了详细综述.首先回顾传统目标检测算法及其存在的问题,其次总结深度学习下区域提案和单阶段基准检测模型.之后从特征图、上下文模型、边框优化、区域提案、类别不平衡处理、训练策略、弱监督学习和无监督学习这八个角度分类总结当前主流的目标检测模型,最后对目标检测算法中待解决的问题和未来研究方向做出展望.

关键词: 计算机视觉;深度学习;目标检测;卷积神经网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2021) 07-1428-11
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200570

A Survey of Generic Object Detection Methods Based on Deep Learning

CHENG Xu[1],SONG Chen[1],SHI Jin-gang[2],ZHOU Lin[3],ZHANG Yi-feng[3],ZHENG Yu-hui[1]
(1. *School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, Jiangsu* 210044, *China*;
2. *School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi* 710049, *China*;
3. *School of Information Science and Engineering, Southeast University, Nanjing, Jiangsu* 210096, *China*)

Abstract: Object detection is one of the most fundamental and important tasks in the field of computer vision, which is the basis of high-level vision tasks such as behavior recognition and human-computer interaction. With the development of deep learning technology, the accuracy and efficiency of object detectors have been greatly improved. Compared with traditional object detection algorithms, deep learning utilizes powerful hierarchical feature extraction and learning capabilities to make breakthroughs in the performance of object detectors. Meanwhile, the large-scale datasets and the tremendous improvement in computing power have also contributed to the vigorous development in this field. In this paper, the existing research of object detectors based on deep learning are reviewed in detail. First, we review the traditional object detection algorithms and its problems. Then, object detectors based on deep learning are introduced, and the region-based and single-stage benchmark detectors are summarized. After that, the current mainstream object detectors are concluded from eight perspectives of feature maps, context information, bounding box optimization, regional proposal, category imbalance processing, training strategy, weakly supervised learning and unsupervised learning. Finally, the problems to be solved in the object detectors are proposed and future research directions are prospected.

Key words: computer vision;deep learning;object detection;convolutional neural network

收稿日期:2020-06-15;修回日期:2021-01-15;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61802058, No. 61911530397, No. 61972206, No. 61672293);江苏省自然科学基金(No. BK20201267);中国博士后科学基金(No. 2019M651650)

1　引言

目标检测是计算机视觉领域中最基础且最具挑战性的任务之一,其包含物体分类和定位[1]. 与此同时,目标检测作为图像理解和计算机视觉的基石,它为实例分割、图像捕获、视频跟踪等任务提供了强有力的特征分类基础,因此探索高效实时的目标检测模型是近年来研究的热点.

传统的目标检测方法包括预处理、区域提案、特征提取、特征选择、特征分类和后处理六个阶段. 大多数检测模型关注于物体特征的提取和区域分类算法的选择,在PASCAL VOC数据集[2]上的检测准确率以较小步幅增长. Deformable Part-based Model(DPM)[3]算法三次在PASCAL VOC目标检测竞赛上获得冠军,是传统目标检测方法的巅峰之作. 然而在2008年至2012年期间,目标检测模型在PASCAL VOC数据集上的检测准确率逐渐达到瓶颈. 传统方法的弊端也展现出来,主要包括:(1)算法在区域提案生成阶段产生大量冗余的候选框且正负样本失衡;(2)特征提取器如HOG[4]、SIFT[5]等未能充分捕捉图像的高级语义特征和上下文内容;(3)传统检测算法分阶段进行,整体缺乏一种全局优化策略.

最近,深度学习经历了一段前所未有的发展热浪,AlexNet[6]在图像分类任务中的优异表现让人们重新燃起研究卷积神经网络的兴趣. 相比于传统算法,深度学习利用自动学习数据中的特征表达和学习能力加速了目标检测的发展,在检测速度和准确率方面均有显著提升. 正是由于目标检测技术的快速发展,如今其已广泛应用于智能视频监控、机器人视觉、基于内容的图像检索、自动驾驶[7,8]等领域.

本文首先介绍目标检测数据集及其评估指标,之后总结基于深度学习的目标检测基准模型,再从特征图、上下文模型、边框优化、区域提案、类别不平衡处理、训练策略、弱监督学习和无监督学习这八个方面归纳总结当前主流的目标检测模型,最后讨论目标检测技术的未来发展趋势与总结全文.

2　数据集和评估指标

数据、算力和算法是深度学习蓬勃发展的三大要素,其中数据是深度学习发展的基础,为目标检测技术发展提供了先决条件. 评估指标为检测算法的性能优劣提供了基准指标,为全面分析检测模型提供了先决条件.

2.1　目标检测数据集

目前主流的通用目标检测数据集有PASCAL VOC[2]、ImageNet[11]、MS COCO[12]、Open Images[13]和Objects365[14],相关数据集的信息在表1中列出.

表1　通用目标检测数据集

数据集	图像数量	图像尺寸	图像总类	提出年份	特点
PASCAL VOC(2012)[2]	11540	470×380	20	2012	包含日常生活中常见20种分类物体 图像接近于真实世界,拥有完整注释
ImageNet[9]	14000000+	500×400	21841	2009	充足的图像数量和丰富的物体种类 提供200种分类,共计50万张图像的训练集
MS COCO[10]	328,000+	640×480	80	2014	更加精细的图像注释和轮廓信息 提供多测试指标如 AP_S、AP_M和 AP_L
Objects365[12]	630000	/	365	2019	图像数量、种类、物体标注框多 规模大、质量高和泛化能力强
Open Images v6[11]	9000000+	/	19957	2020	掩码级的物体轮廓更加精细 新增局部叙事,人类动作等视觉关系标注

2.2　评价指标

当前用于评估检测模型的性能指标主要有帧率每秒(Frames Per Second, FPS)、准确率(accuracy)、精确率(precision)、召回率(recall)、平均精度(Average Precision, AP)、平均精度均值(mean Average Precision, mAP)等. FPS即每秒识别图像的数量,用于评估目标检测模型的检测速度;accuracy是正确预测类别的样本数占样本总数的比例;precision是预测正确的正样本数占所有预测为正样本个数的比例;recall是预测正确的正样本数占所有真实值为正样本个数的比例;PR曲线是对应precision和recall构成的曲线,AP是对不同召回率点上的精确率进行平均,在PR曲线图上表现为PR曲线下的面积;mAP是所有类别AP的平均.

3　基于深度学习的基准目标检测模型

基于深度学习的目标检测方法根据有无区域提案阶段划分为区域提案检测模型和单阶段检测模型,其最近发展历程在图1中画出.

3.1　区域提案目标检测基准模型

区域提案检测模型将目标检测任务分为区域提案生成、特征提取和分类预测三个阶段. 在区域提案生成阶段,检测模型利用搜索算法如选择性搜索(Selective

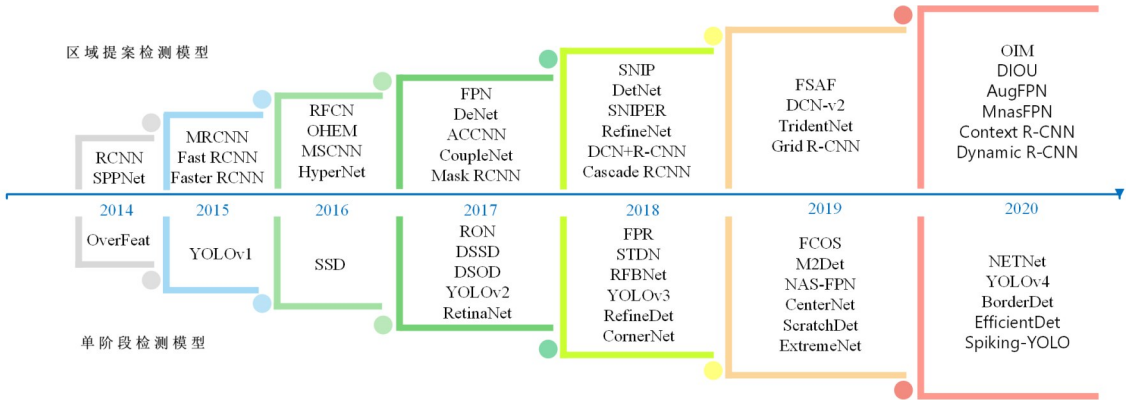

图1 基于深度学习的目标检测模型的发展历程

Search, SS)[13]、EdgeBoxes[14]、区域提案网络(Region Proposal Network, RPN)[15]等在图像中搜寻可能包含物体的区域. 在特征提取阶段,模型利用深度卷积网络提取区域提案中的目标特征. 在分类预测阶段,模型从预定义的类别标签对区域提案进行分类和边框信息预测.

2012年,卷积神经网络AlexNet[6]通过强大的分层特征学习能力在ImageNet分类任务中夺魁,其分类结果比传统算法要好得多. 之后Girshirk等[16]基于此提出通用的目标检测模型RCNN.

受空间金字塔匹配启发,He等[17]提出空间金字塔池化网络(Spatial Pyramid Pooling Network, SPPNet)解决RCNN模型中固定大小图像输入的问题. 尽管模型能解决固定大小图像输入的问题并大幅减少检测时间,但网络输出特征需要磁盘存储且网络不能通过反向传播更新卷积层参数. Fast RCNN[18],Faster RCNN[15]解决了该问题.

尽管Faster RCNN运用RPN能有效地缩短检测时间,但图像中上百个区域提案仍需输入到区域子网络来分类与坐标回归,这是模型的计算瓶颈. 若直接减少区域子网络的深度则会引起基于分类初始模型与检测模型的矛盾,因为分类会增加物体的平移不变性而检测则减少物体的平移不变性. 为解决上述问题,代表性的方法有RFCN[19],Mask RCNN[20]. 上述模型的相关信息在表2中列出.

表2 区域提案目标检测基准模型

检测模型	提出年份	网络架构	模型优点	模型缺点
RCNN[16]	2014	AlexNet	提出区域提案目标检测模型新范式	正负样本失衡、模型复杂度高 分阶段训练时间长且效率低
SPPNet[17]	2014	ZFNet	解决输入图像尺寸固定问题 提高基准检测模型准确率	中间特征需要磁盘存储 网络不能通过反向传播更新卷积层参数
Fast RCNN[18]	2015	VGG	提出ROI Pooling优化区域特征选择	模型训练时间长,正负样本失衡 不能满足实时检测需求
Faster RCNN[15]	2015	VGG	提出RPN替代SS算法 极大地缩短检测时间	锚框数量多且锚框正负样本失衡 训练分阶段进行且时间长
RFCN[19]	2016	ResNet	使用位置敏感得分图减少物体的平移不变性	训练时正负样本失衡 不能满足实时检测需求
Mask RCNN[20]	2016	ResNet	双线性插值填补非整数位置像素,使用额外分支输出候选区域的掩码预测,实现更准确检测	训练时正负样本失衡且训练时间长 不能满足实时检测需求

3.2 单阶段目标检测基准模型

虽然区域提案检测模型保持着高效的检测效率,但其耗费时间长难以得到实际应用. 相较于区域提案检测模型,单阶段检测模型联合区域提案和分类预测,输入整张图像到卷积神经网络中提取特征,最后直接输出目标类别和边框位置信息. 这类代表性的方法有:

YOLO[21], YOLOv2[22], YOLOv3[23], YOLOv4[24], SSD[25], CornerNet[27]等.

针对YOLO模型[21]中目标定位不准确的问题,Liu等人[25]提出更准确的单阶段实时检测模型SSD(Single Shot MultiBox Detector),其结合YOLO的回归思想与Faster RCNN的锚框机制. 之后DSSD(Deconvolutional Single Shot Detector)[26]用于小目标检测. 然而,锚框机制也存在明显的缺陷,如正负样本不平衡、引入过多的超参数而折戟检测速度与性能等. 基于此,研究者提出了无锚框单阶段检测模型[27,28],上述模型的相关信息在表3中列出.

表3　单阶段目标检测基准模型

检测模型	提出年份	网络架构	模型优点	模型缺点
YOLOv1[21]	2016	GoogLeNet	提出单阶段目标检测模型新范式	模型难以预测密集目标和小物体 检测准确率不高
YOLOv2[22]	2017	DarkNet	采取K-means聚类数据集边框信息 可预测1000类物体	模型复杂度高且训练步骤多 检测准确率不高
YOLOv3[23]	2018	DarkNet	独立的逻辑回归支持多标签预测	模型训练时间长,泛化性差
YOLOv4[24]	2020	CSPDarkNet	结合多种策略、方法与模型 检测速度快且准确率高	预测框误检率高
SSD[25]	2016	VGG	分层特征图预测不同尺度物体	模型对小物体检测准确率不高 锚框多且正负样本失衡
CornerNet[27]	2018	Hourglass	转换边界框检测为角点检测 模型训练代价小	预测框误检率高 检测准确率不高

4　基于深度学习的目标检测衍生算法

当前主流的基于深度学习的目标检测方法可分为检测部件、数据增强、优化方法和学习策略四个方面. 其中检测部件包含基准模型和基准网络;数据增强包含几何变换、光学变换等;优化方法包含特征图、上下文模型、边框优化、区域提案方法、类别不平衡和训练策略六个方面,学习策略涵盖监督学习、弱监督学习和无监督学习. 本节从优化方法和学习策略这两个大的方面归纳总结了深度学习下基准目标检测模型的衍生方法. 基于深度学习的目标检测部件及其代表性的检测方法如图2所示.

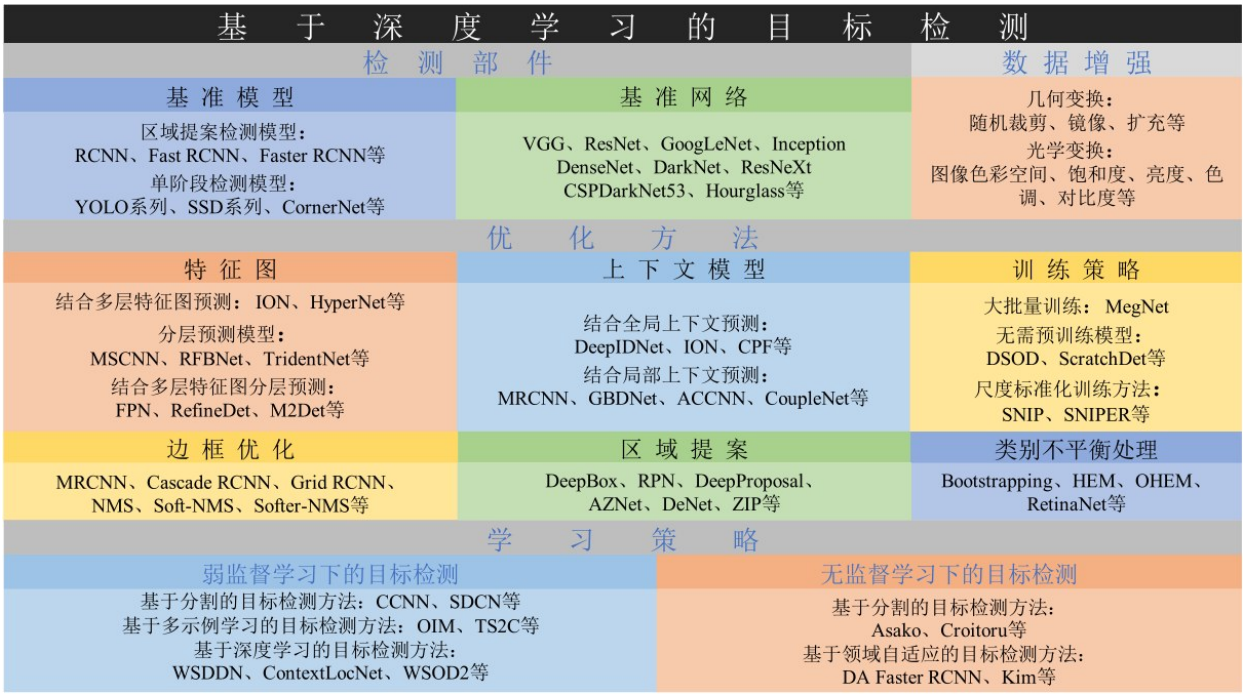

图2　基于深度学习的目标检测部件及其代表性的检测方法

4.1 融合特征图的目标检测模型

特征图是图像经过卷积池化层输出的结果,大多数基准检测模型只在顶层特征图做预测,这在很大程度上限制了模型的性能. 为了充分提取特征信息,现有检测模型从三个角度融合浅中深层特征,分别是:结合多层特征图单层预测模型(ION[29]、HyperNet[30])、分层预测模型(MSCNN[31]、SSD[25]、RFBNet[32]、TridentNet[33])和结合多层特征图多层预测模型(FPN[34]、DSSD[26]、STDN[35]、DetNet[36]、M2Det[37]、FCOS[38]、EfficientDet[39]). 相关模型信息在表4中列出.

表4 融合特征图的目标检测模型

检测模型	提出年份	卷积架构	基准架构	模型亮点
结合多层特征图单层预测的检测模型				
ION[29]	2016	VGG16	Fast RCNN	跳跃连接和循环神经网络分别用于提取多层特征
HyperNet[30]	2016	VGG16	Faster RCNN	融合从粗糙到精细的多个特征图
分层预测的检测模型				
MSCNN[31]	2016	VGG16	Faster RCNN	区域提案和分类同时在多层进行
SSD[25]	2016	VGG16	/	分层特征图预测不同尺度物体
RFBNet[32]	2018	VGG16	SSD	采取类似Inception模块的多分支卷积块
TridentNet[33]	2019	ResNet101	Faster RCNN	提出三分支权重共享且多扩张参数的卷积层
结合多层特征图分层预测的检测模型				
FPN[34]	2017	ResNet101	Faster RCNN	结合浅中深层特征图用于预测
DSSD[26]	2017	ResNet101	SSD	提出反卷积层和跳跃连接传递更多语义信息
STDN[35]	2018	DenseNet169	SSD	提出尺度迁移模块用于获得不同分辨率特征
DetNet[36]	2018	DetNet59	Faster RCNN	引入扩张卷积获得不同分辨率的特征图
M2Det[37]	2019	ResNet101	SSD	提出多分支模块用于更精确地分层预测
FCOS[38]	2019	ResNet101	RetinaNet	无锚框、无区域提案、对像素点预测
EfficientDet[39]	2020	EfficientNet	/	提出加权双向特征金字塔网络学习特征

4.2 结合上下文信息的目标检测模型

在物体遮挡、背景信息杂乱或图像质量不佳的情况下,根据图像的上下文信息能更有效更精确地检测. 现有的目标检测模型主要考虑将上下文信息分为全局上下文信息(DeepIDNet[40]、ION[29]、CPF[41])和局部上下文信息(MR-CNN[42]、GBDNet[43]、ACCNN[44]、CoupleNet[45]). 相关模型的信息在表5中列出.

表5 上下文模型和边框优化模型

检测模型	提出年份	卷积架构	基准架构	模型亮点
结合全局上下文的检测模型				
DeepIDNet[40]	2015	ZFNet	RCNN	在放大区域中提取特征作为附加信息
CPF[41]	2016	VGG16	Faster RCNN	语义分割用于上下文推理和迭代反馈
结合局部上下文的检测模型				
MR-CNN[42]	2015	VGG16	SPPNet	多区域特征与语义特征辅助检测
GBDNet[43]	2016	ResNet269	Fast RCNN	双向门控卷积神经网络用于传递信息
ACCNN[44]	2017	VGG16	Fast RCNN	提出注意力上下文子网和多尺度本地子网
CoupleNet[45]	2017	ResNet101	RFCN	提取区域提案周围多尺度上下文区域特征
边框优化模型				
MRCNN[46]	2016	VGG16	Fast RCNN	利用迭代边框优化的方式选取预测框
CascadeRCNN[47]	2018	ResNet101	FPN	使用一系列递增IOU阈值级联训练
Grid RCNN[48]	2018	ResNet101	FPN	采取网格引导本地化精确目标检测机制
Soft NMS[50]	2017	ResNet101	RFCN	新设一个置信度阈值用于处理候选预测框
Softer NMS[51]	2018	ResNet50	FPN	提出新的边框回归损失函数KL Loss

4.3 优化边框定位的目标检测模型

当前检测模型在小目标检测表现不佳的主要原因是定位错误偏多,包含定位偏差大和重复预测,因此部分研究着眼于优化边框定位来提升检测性能,代表性的模型有MRCNN[46]、Cascade RCNN[47]、Grid RCNN[48]等. 此外,一些算法使用后处理步骤来优化预测框位置,如NMS[49]、Soft-NMS[50]、Softer-NMS[51]等.

4.4 高效区域提案的目标检测模型

区域提案是图像中可能包含物体的区域,它是两阶段检测模型中性能保障的关键. 早期的检测模型DPM使用滑动窗口方法,存在计算复杂度高和定位性能差的问题. RCNN使用SS算法提取候选区域并利用卷积神经网络提取图像特征,其检测效率和性能上均有大幅提高. EdgeBox利用图像中低维线索如颜色、纹理、边缘、梯度等对其分类,表现出良好的检测性能. Kuo等人[52]在EdgeBox基础上提出DeepBox检测模型,运行速度更快且提案窗口召回率更高. Ren等人[15]提出使用RPN生成候选区域的Faster RCNN检测模型,在特征图上以每个像素点为中心生成三个尺度和三个长宽比总共九个锚框. Ghodrati等人[53]提出DeepProposal检测模型,使用多个级联的卷积特征来生成对象提案再构建逆级联选择图像中可能存在的对象位置.

4.5 处理类别不平衡的目标检测模型

类别不平衡的主要矛盾是负样本数远多于正样本数,导致训练的深度模型效率低. 传统检测算法常用Bootstrapping[54]方法处理此问题,之后RCNN模型使用困难样本挖掘(Hard Example Mining, HEM)方法来处理. Shrivastava等人[55]在HEM基础上提出在线困难样本挖掘方法(Online Hard Example Mining, OHEM),其根据区域提案损失有选择性地反向传播负样本区域更新梯度. 最近,Lin等人[56]提出使用Focal Loss的单阶段检测模型RetinaNet,使模型更关注于那些少量的困难样本. 表6总结了类别不平衡处理模型和训练策略方法.

表6　类别不平衡处理模型和训练策略方法

检测模型	提出年份	卷积架构	基准架构	模型亮点
OHEM[55]	2016	VGG16	Fast RCNN	有选择地反向传播困难样本区域梯度
RetinaNet[56]	2017	ResNet101	FPN	Focal Loss取代常用标准的交叉熵损失 在单阶段检测模型上有效处理类别不平衡
MegNet[57]	2018	ResNet50	Faster RCNN	大批量样本训练方法; 跨GPU批量归一化
LargeDet[58]	2020	ResNet50	FPN	提出一种周期性动量衰减层级自适应动量优化器 采取同步批处理标准化快速收敛模型
SNIP[59]	2018	DCN	RFCN	选择性传播不同尺寸物体的梯度作为损失
SNIPER[60]	2018	ResNet101 DCN	Faster RCNN	多尺度训练策略 采取Negative chip sampling策略
DSOD[61]	2017	DenseNet	SSD	无需预训练,只在检测数据集上训练模型
ScratchDet[62]	2019	ResNet34	SSD	BatchNorm、Root-ResNet变种网络

4.6 训练策略

大多数目标检测模型采取小批量样本进行训练,然而小批量样本训练存在梯度不稳定、训练时间长等问题. 研究者们提出一些高效的方法解决上述问题,典型的方法有:MegNet[57], LargeDet[58], SNIP[59], SNIPER[60], DSOD[61], ScratchDet[62]等. 相关训练策略的信息在表6中列出.

4.7 基于弱监督学习的目标检测方法

数据标注的昂贵性和人工标注的主观性已成为一个棘手的问题. 基于弱监督学习的目标检测方法主要划分为三类:基于分割的目标检测方法[63~65]、基于多示例学习的目标检测方法[66~68]和基于深度学习的目标检测方法[69~71]. 这些模型的相关信息在表7中列出.

4.8 基于无监督的目标检测方法

尽管基于弱监督学习的目标检测方法仅需要图像级别信息即可训练,表现出了良好的性能. 然而,在现实应用中图像往往没有标注信息. 目前,基于无监督学习的目标检测方法大致可分为两类:基于分割的目标检测方法[72,73]和基于领域自适应的目标检测方法[74~76]. 模型的相关信息在表8中列出.

5 目标检测技术的研究趋势

近年来随着深度学习技术的快速发展,通用目标检测技术发展迅速且取得重大突破,但目前检测模型效率和速度与人性化的表现之间仍存在巨大差距. 已有的研究方法表明:基于深度学习的通用目标检测技术待解决的问题和未来研究趋势主要包括:

(1)如何高效权衡检测速度与精度. 区域提案和单阶段检测模型都表现出各自强力的优势,但鲜有检测方法能共具两者优点. 对于区域提案检测模型,计

表7 基于弱监督学习的通用目标检测方法

弱监督学习下的目标检测方法	提出年份	模型亮点
弱监督学习下基于分割的目标检测方法		
Liu[63]	2011	提出利用条件随机场表征图像视觉特征
CCNN[64]	2015	转化任务为线性条件约束下的训练模型最优化
SDCN[65]	2019	采取协作循环的方式指导分割模块与检测模块
弱监督学习下基于多示例学习的目标检测方法		
Arun[66]	2019	提出利用相异系数概率学习图像位置特征信息
OIM[67]	2020	提出在空间和外观图中引入信息传播检测对象实例
Ren[68]	2020	引入示例空间多样化约束计算预测与真实标签差异
弱监督学习下基于深度学习的目标检测方法		
WSDDN[69]	2016	利用深度神经网络极强的非线性映射能力描述特征
ContextLocNet[70]	2016	引入图像目标的上下文语义信息实现目标精确定位
WSOD2[71]	2019	联合考虑低维特征与分类置信分预测目标提案

表8 基于无监督学习的通用目标检测方法

无监督学习下的目标检测方法	提出年份	模型亮点
无监督学习下基于分割的目标检测方法		
Asako[72]	2018	提出利用超像素分割的方法预测图像像素点类别
Croitoru[73]	2019	提出一种双路径下的无监督对象学习方法
无监督学习下基于领域自适应的目标检测方法		
DA Faster RCNN[74]	2018	提出基于H-散度理论的域自适应组件学习域间差异
Kim[75]	2019	提出弱自训练和对抗背景正则化方法用于减弱偏移
Hsu[76]	2020	提出利用中间域弥补源域与目标域间数据分布差异

算瓶颈主要在候选区域生成阶段,未来研究会聚焦于设计召回率高且区域提案少的基准网络. 对于单阶段检测模型,性能瓶颈在于模型提取较少的图像特征,高效特征提取网络的设计是值得研究的一个方向.

(2)如何在先验知识缺失的条件下实现目标的精确检测. 当前大多数目标检测方法只能检测识别数据集上的常规物体,对现实环境如遮挡、模糊、视点和光照变化、变形等下的物体检测识别鲁棒性不高,因此如何利用迁移学习和强化学习方法来拓展识别种类和增强检测算法是未来的研究热点之一.

(3)如何设计高效的特征提取网络. 目前大多数目标检测模型的特征提取网络使用性能卓越的分类网络完成. 然而,分类和检测任务之间的差异性导致学习过程会产生偏差,因此研究适用于目标检测领域的专用特征提取网络显得很有必要. 此外,特征提取网络的参数规模需要耗费大量硬件资源来训练,如何对网络进行压缩和加速以满足实时目标检测需求也是值得考虑的一个方向.

(4)如何获得更加丰富的图像语义信息. 更多的图像语义信息对应图像中更多的物体特征,更丰富的图像语义信息对应图像中更多的物体联系,这两点常对应于特征图和上下文模型. 最近已有研究着眼于图像语义理解. 图像的语义理解可帮助模型更深层次地学习图像信息,其结果可作为辅助信息实现更精确的目标定位.

(5)如何自动生成和设计最优的网络架构. 目前神经架构搜索已在目标检测和识别任务中表现出很强的能力,因此如何利用神经网络架构自动搜索与生成技术来提升目标检测领域模型的检测性能将是未来有希望的研究方向.

6 总结

本文回顾了传统的目标检测算法并指出其存在的问题,引入了目标检测数据集和评估指标;然后分类回顾并总结区域提案和单阶段基准检测模型,并指出他们的优缺点;从特征图、上下文模型、边框优化、区域提案方法、类别不平衡处理和训练策略这六个角度总结两类目标检测基准模型的衍生方法;最后,基于已有的检测方法和最近的研究思路,从5个角度总结了通用目

标检测模型的未来发展趋势.

参考文献

[1] Fischler M A, et al. The representation and matching of pictorial structures[J]. IEEE Transactions on Computers, 1973, 100(1): 67 – 92.

[2] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303 – 338.

[3] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627 – 1645.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2005. 886 – 893.

[5] Lowe D G. Object recognition from local scale-invariant features[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 1999. 1150 – 1157.

[6] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[A]. L Bottou. Advances in Neural Information Processing Systems[C]. CA: Morgan Kaufmann, 2012. 1097 – 1105.

[7] Chen C, Seff A, Kornhauser A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2015. 2722 – 2730.

[8] Chen X, Ma H, Wan J, et al. Multi-view 3D object detection network for autonomous driving[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2017. 1907 – 1915.

[9] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2009. 248 – 255.

[10] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2014. 740 – 755.

[11] Kuznetsova A, Rom H, Alldrin N, et al. The Open Images Dataset v4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale[EB/OL]. arXiv preprint arXiv:1811.00982, 2018.

[12] Shao S, Li Z, Zhang T, et al. Objects365: A large-scale, high-quality dataset for object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2019. 8430 – 8439.

[13] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154 – 171.

[14] Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2014. 391 – 405.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[A]. Advances in Neural Information Processing Systems[C]. CA: Morgan Kaufmann, 2015. 91 – 99.

[16] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2014. 580 – 587.

[17] 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, 38(9): 1904 – 1916.

[18] Girshick R. Fast R-CNN[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2015. 1440 – 1448.

[19] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks[A]. U. Luxburg. Advances in Neural Information Processing Systems[C]. CA: Morgan Kaufmann, 2016. 379 – 387.

[20] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2017. 2961 – 2969.

[21] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18 – 28.

[22] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2016. 779 – 788.

[23] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2017. 7263 – 7271.

[24] Redmon J, Farhadi A. Yolov3: An Incremental Improvement[EB/OL]. arXiv preprint arXiv:1804.02767,

2018.

[25] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[EB/OL]. arXiv preprint arXiv:2004.10934, 2020.

[26] Wang C Y, Mark Liao H Y, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of cnn[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2020. 390 – 391.

[27] Misra D. Mish: A Self Regularized Non-monotonic Neural Activation Function[EB/OL]. arXiv preprint arXiv:1908.08681, 2019.

[28] Ghiasi G, Lin T Y, Le Q V. Dropblock: A regularization method for convolutional networks[A]. H. Wallach. Advances in Neural Information Processing Systems[C]. CA: Morgan Kaufmann, 2018. 10727 – 10737.

[29] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2016. 21 – 37.

[30] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional Single Shot Detector[EB/OL]. arXiv preprint arXiv:1701.06659, 2017.

[31] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2018. 734 – 750.

[32] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2019. 6569 – 6578.

[33] Bell S, Lawrence Zitnick C, Bala K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2016. 2874 – 2883.

[34] Kong T, Yao A, Chen Y, et al. Hypernet: Towards accurate region proposal generation and joint object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2016. 845 – 853.

[35] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2016. 354 – 370.

[36] Liu S, Huang D. Receptive field block net for accurate and fast object detection[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2018. 385 – 400.

[37] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2019. 6054 – 6063.

[38] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2017. 2117 – 2125.

[39] Zhou P, Ni B, Geng C, et al. Scale-transferrable object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2018. 528 – 537.

[40] Li Z, Peng C, Yu G, et al. DetNet: Design backbone for object detection[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2018. 334 – 350.

[41] Zhao Q, Sheng T, Wang Y, et al. M2Det: A single-shot object detector based on multi-level feature pyramid network[A]. Yang Q. American Association for Artificial Intelligence[C]. New York: IEEE, 2019. 9259 – 9266.

[42] Tian Z, Shen C, Chen H, et al. FCOS: Fully convolutional one-stage object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2019. 9627 – 9636.

[43] Tan M, Pang R, Le Q V. Efficientdet: scalable and efficient object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2020. 10781 – 10790.

[44] Ouyang W, Wang X, Zeng X, et al. DeepID-Net: Deformable deep convolutional neural networks for object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2015. 2403 – 2412.

[45] Shrivastava A, Gupta A. Contextual priming and feedback for faster R-CNN[A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2016. 330 – 348.

[46] Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware CNN model[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2015. 1134 – 1142.

[47] Zeng X, Ouyang W, Yan J, et al. Crafting GBD-net for object detection[J]. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 2017, 40(9): 2109 – 2123.

[48] Li J, Wei Y, Liang X, et al. Attentive contexts for object detection[J]. IEEE Transactions on Multimedia, 2016, 19(5): 944 – 954.

[49] Zhu Y, Zhao C, Wang J, et al. Couplenet: Coupling global structure with local parts for object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2017. 4126 – 4134.

[50] Zagoruyko S, Lerer A, Lin T Y, et al. A Multipath Network for Object Detection [EB/OL]. arXiv preprint arXiv:1604.02135, 2016.

[51] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2018. 6154 – 6162.

[52] Lu X, Li B, Yue Y, et al. Grid R-CNN[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2019. 7363 – 7372.

[53] Rosenfeld A, Thurston M. Edge and curve detection for visual scene analysis[J]. IEEE Transactions on computers, 1971, 100(5): 562 – 569.

[54] Bodla N, Singh B, Chellappa R, et al. Soft-NMS—improving object detection with one line of code[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2017. 5561 – 5569.

[55] He Y, Zhu C, Wang J, et al. Bounding box regression with uncertainty for accurate object detection[A]: Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2019. 2888 – 2897.

[56] Kuo W, Hariharan B, Malik J. Deepbox: Learning objectness with convolutional networks[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2015. 2479 – 2487.

[57] Ghodrati A, Diba A, Pedersoli M, et al. Deep proposal: Hunting objects by cascading deep convolutional layers [A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2015. 2578 – 2586.

[58] Sung K K. Learning and Example Selection for Object and Pattern Detection[D]. Massachusetts, USA: MIT AI Lab, 1995.

[59] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2016. 761 – 769.

[60] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2017. 2980 – 2988.

[61] Peng C, Xiao T, Li Z, et al. MegDet: A large mini-batch object detector[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2018. 6181 – 6189.

[62] Wang T, Zhu Y, Zhao C, et al. Large batch optimization for object detection: training coco in 12 minutes [A]. Vittorio Ferrari. European Conference on Computer Vision[C]. Berlin: Springer, 2020. 481 – 496.

[63] Singh B, Davis L S. An analysis of scale invariance in object detection snip[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2018. 3578 – 3587.

[64] Singh B, Najibi M, Davis L S. Sniper: Efficient multi-scale training[A]. H Wallach. Advances in Neural Information Processing Systems [C]. CA: Morgan Kaufmann, 2018. 9310 – 9320.

[65] Shen Z, Liu Z, Li J, et al. Dsod: Learning deeply supervised object detectors from scratch[A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2017. 1919 – 1927.

[66] Zhu R, Zhang S, Wang X, et al. ScratchDet: Training single-shot object detectors from scratch [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2019. 2268 – 2277.

[67] Liu T, Yuan Z, Sun J, et al. Learning to detect a salient object[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(2): 353 – 367.

[68] Li J, Li X, Yang B, et al. Segmentation-based image copy-move forgery detection scheme[J]. IEEE Transactions on Information Forensics and Security, 2014, 10 (3): 507 – 518.

[69] Li X, Kan M, Shan S, et al. Weakly supervised object detection with segmentation collaboration [A]. Jim Little. IEEE International Conference on Computer Vision[C]. New York: IEEE, 2019. 9735 – 9744.

[70] Arun A, Jawahar C V, Kumar M P. Dissimilarity coefficient based weakly supervised object detection[A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition[C]. New York: IEEE, 2019. 9432 – 9441.

[71] Lin C, Wang S, Xu D, et al. Object instance mining for weakly supervised object detection [A]. Yang Q.

American Association for Artificial Intelligence [C]. New York: IEEE, 2020. 11482 – 11489.
[72] Ren Z, Yu Z, Yang X, et al. Instance-aware, context-focused, and memory-efficient weakly supervised object detection [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2020. 10598 – 10607.
[73] Bilen H, Vedaldi A. Weakly supervised deep detection networks [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2016. 2846 – 2854.
[74] Kantorov V, Oquab M, Cho M, et al. ContextLocNet: Context-aware deep network models for weakly supervised localization [A]. Vittorio Ferrari. European Conference on Computer Vision [C]. Berlin: Springer, 2016. 350 – 365.
[75] Zeng Z, Liu B, Fu J, et al. WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection [A]. Jim Little. IEEE International Conference on Computer Vision [C]. New York: IEEE, 2019. 8292 – 8300.
[76] Kanezaki A. Unsupervised image segmentation by back-propagation [A]. IEEE International Conference on Acoustics, Speech and Signal Processing [C]. New York: IEEE, 2018. 1543 – 1547.
[77] Croitoru I, Bogolin S V, Leordeanu M. Unsupervised learning of foreground object segmentation [J]. International Journal of Computer Vision, 2019, 127 (9): 1279 – 1302.
[78] Chen Y, Li W, Sakaridis C, et al. Domain adaptive faster R-CNN for object detection in the wild [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2018. 3339 – 3348.
[79] Zhu X, Pang J, Yang C, et al. Adapting object detectors via selective cross-domain alignment [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2019. 687 – 696.
[80] Hsu H K, Yao C H, Tsai Y H, et al. Progressive domain adaptation for object detection [A]. Satya Nadella. IEEE Conference on Computer Vision and Pattern Recognition [C]. New York: IEEE, 2020. 749 – 757.

作者简介

程 旭(通信作者) 男,1983年出生,山西祁县人,南京信息工程大学计算机与软件学院副教授、硕士生导师,研究方向为计算机视觉、图像理解.
E-mail:xcheng@nuist.edu.cn

周 琳 女,1976年出生,江苏镇江人,东南大学信息科学与工程学院副教授、硕士生导师,主要研究方向为深度学习、语音信号处理及其应用.
E-mail:Linzhou@seu.edu.cn

宋 晨 男,1997年出生,江苏盐城人,南京信息工程大学计算机与软件学院软件工程专业硕士研究生,主要研究方向为深度学习、目标检测及其应用.
E-mail:20191221015@nuist.edu.cn

张毅锋 男,1963年出生,安徽芜湖人,东南大学信息科学与工程学院副教授、硕士生导师,研究方向为计算机视觉、图像理解.
E-mail:yfz@seu.edu.cn

史金钢 男,1988年出生,黑龙江哈尔滨人,西安交通大学软件学院副教授、博士生导师,研究方向为计算机视觉、深度学习.
E-mail:jingang.shi@hotmail.com

郑钰辉 男,1982年出生,山西芮城人,南京信息工程大学计算机与软件学院教授、博士生导师,研究方向为计算机视觉、模式识别.
E-mail:zhengyh@vip.126.com