E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中图法分类号:TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(2020)04-0629-26

论文引用格式: Zhao Y Q, Rao Y, Dong S P and Zhang J Y. 2020. Survey on deep learning object detection. Journal of Image and Graphics, 25(04): 0629-0654(赵永强, 饶元, 董世鹏, 张君毅. 2020. 深度学习目标检测方法综述. 中国图象图形学报, 25(04): 0629-0654) [DOI: 10.11834/jig.190307]

深度学习目标检测方法综述

赵永强, 饶元, 董世鹏, 张君毅

1. 西安交通大学软件学院社会智能与复杂数据处理实验室, 西安　710049; 2. 西安交通大学深圳研究院, 深圳　518057

摘　要:　目标检测的任务是从图像中精确且高效地识别、定位出大量预定义类别的物体实例。随着深度学习的广泛应用,目标检测的精确度和效率都得到了较大提升,但基于深度学习的目标检测仍面临改进与优化主流目标检测算法的性能、提高小目标物体检测精度、实现多类别物体检测、轻量化检测模型等关键技术的挑战。针对上述挑战,本文在广泛文献调研的基础上,从双阶段、单阶段目标检测算法的改进与结合的角度分析了改进与优化主流目标检测算法的方法,从骨干网络、增加视觉感受野、特征融合、级联卷积神经网络和模型的训练方式的角度分析了提升小目标检测精度的方法,从训练方式和网络结构的角度分析了用于多类别物体检测的方法,从网络结构的角度分析了用于轻量化检测模型的方法。此外,对目标检测的通用数据集进行了详细介绍,从4个方面对该领域代表性算法的性能表现进行了对比分析,对目标检测中待解决的问题与未来研究方向做出预测和展望。目标检测研究是计算机视觉和模式识别中备受青睐的热点,仍然有更多高精度和高效的算法相继提出,未来将朝着更多的研究方向发展。

关键词:目标检测; 深度学习; 小目标; 多类别; 轻量化

Survey on deep learning object detection

Zhao Yongqiang, Rao Yuan, Dong Shipeng, Zhang Junyi

1. Laboratory of Social Intelligence and Complex Data Processing, School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. Shenzhen Research Institution, Xi'an Jiaotong University, Shenzhen 518057, China

Abstract: The task of object detection is to accurately and efficiently identify and locate a large number of predefined objects from images. It aims to locate interested objects from images, accurately determine the categories of each object, and provide the boundaries of each object. Since the proposal of Hinton on the use of deep neural network for automatic learning of high-level features in multimedia data, object detection based on deep learning has become an important research hotspot in computer vision. With the wide application of deep learning, the accuracy and efficiency of object detection are greatly improved. However, object detection based on deep learning still have four key technology challenges, namely, improving and optimizing the mainstream object detection algorithms, balancing the detection speed and accuracy, improving the small object detection accuracy, achieving multiclass object detection, and lightweighting the detection model. In view of the above challenges, this study analyzes and summarizes the existing research methods from different aspects. On the basis of extensive literature research, this work analyzed the methods of improving and optimizing the mainstream

收稿日期:2019-06-28; **修回日期**:2019-09-15; **预印本日期**:2019-09-22

基金项目: 国家自然科学基金项目(F020807); 科技部重点研发计划项目(2019YFB2102300); 教育部"云数融合"基金项目(2017B00030); 中央高校基本科研业务项目(ZDYF2017006)

Supported by: National Natural Science Foundation of China(F020807)

suitable for the detection needs of specific scenarios, how to achieve accurate object detection problems under the condition of lack of prior knowledge, how to obtain high-performance backbone network and information, how to add rich image semantic information, how to improve the interpretability of deep learning model, and how to automate the realization of the optimal network architecture.

Key words: object detection; deep learning; small object; multi-class; lightweighting

0 引 言

自从 Hinton 提出利用神经网络(LeCun 等,2015)对多媒体数据中的高层特征进行自动学习以来,基于深度学习的目标检测已成为计算机视觉领域中一个重要的研究热点(Liu 等,2019b),其旨在从图像中定位感兴趣的目标,准确判断每个目标的类别,并给出每个目标的边界框。为了获得更加丰富的目标表示特征,人们一方面构建 ImageNet(Nielsen,2018)、COCO(common objects in context)(Havard 等,2017)等大规模图像数据库,另一方面通过构建 VGG(visual geometry group)网络(Shelhamer 等,2017)、GoogLeNet(Ning 等,2017)及残差网络(residual network, ResNet)(He 等,2016)将卷积网络推向更深层次,大幅提高了网络性能,极大地推动了多媒体目标识别的准确度与执行效率,并在视频监控、智能交通、手术器械定位、机器人导航、车辆自动驾驶、机器人环境感知和基于内容的图像检索等领域得到了广泛应用(Liu 等,2018a)。

一般地,传统目标检测算法主要包括预处理、窗口滑动、特征提取、特征选择、特征分类和后处理等6个关键步骤。其中,窗口大小、滑动方式与策略对特征提取的质量影响较大,常采用部位形变模型(deformable part model, DPM)及其扩展模型(Divvala 等,2012)对滑动窗口进行判别,如方向梯度直方图(histogram of oriented gradient, HOG)(Wang 等,2009)、尺度不变特征变换(scale invariant feature transform, SIFT)(Juan 和 Gwun,2013)等,整个检测过程效率与精度都较低。Girshick 等人(2014)首次采用基于区域的卷积神经网络(region based convolutional neural network, R-CNN)将深度学习用于目标检测,在 PASCAL VOC 2007 数据集上的检测精度从29.2%提升到66.0%,极大地提高了目标检测的准确率。这种基于端到端的训练,将目标的特征提取、特征选择和特征分类融合在同一模型中,实现了性能与效率的整体优化。随着大量基于深度学习的目标检测算法的提出与应用,需要面对并解决的问题与挑战也随之出现,主要体现在以下4个核心方面:

1)如何实现基于主流目标检测算法的性能改进与优化。高准确率和高效率(Liu 等,2018b)是每个算法设计与应用的核心目标,但在实际情况下,具有高准确率的双阶段目标检测算法往往会消耗更多的计算资源,导致效率下降,而具有高效率的单阶段目标检测算法则相反,因此如何改进与优化主流的目标检测算法,实现精度与效率的最佳平衡是一个关键问题。

2)如何实现小目标物体的高精度检测。在实际需求中,小目标在图像检测任务中所占比例极大,而目标检测算法多是针对通用目标数据集进行设计,对小目标检测效果较差,因此如何提高小目标的检测精度是目标检测算法应用中的一个关键问题。

3)如何实现多类别物体检测。多类别物体检测是目标检测算法运用到实际需求中必须达到的一个要求,现有的目标检测算法在目标类别较少的 PASCAL VOC 数据集(Everingham 等,2015)上的检测精度可达到80%以上,但是在类别较多的 MS COCO 数据集上的检测精度却只有40%左右,同时随着类别的增加,模型运算量也会急剧增大,因此如何处理多类别物体的检测问题将是一个关键。

4)如何满足目标检测算法的轻量化需求。目标检测算法在移动端(Wang 等,2019a)、自动驾驶(Shan 等,2019)及嵌入式等领域应用的前提条件是实现轻量化处理,而深度神经网络的使用通常伴随着模型体积大、运行消耗资源多等问题,如何将基于深度学习的目标检测算法进行轻量化处理则是一个关键。

围绕上述问题与挑战,本文对基于深度学习的目标检测算法的研究进展与现状进行分析和综述,详细介绍目标检测的通用数据集和不同算法在主流数据集上的实验结果,并对目标检测领域未来可能的发展方向进行展望。

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 25, No. 4, Apr. 2020

1 基于主流目标检测算法的性能改进与优化

基于深度学习的主流目标检测算法根据有无候选框生成阶段分为双阶段目标检测算法和单阶段目标检测算法两类。双阶段目标检测算法先对图像提取候选框，然后基于候选区域做二次修正得到检测结果，检测精度较高，但检测速度较慢；单阶段目标检测算法直接对图像进行计算生成检测结果，检测速度快，但检测精度低。因此在目标检测中需要不断地改进与优化主流目标检测算法，以实现检测精度和检测速度的最佳平衡。根据研究思路不同，改进与优化主流目标检测算法的方法可以分为 3 类：基于双阶段目标检测算法的改进、基于单阶段目标检测算法的改进、基于双阶段和单阶段目标检测算法的结合。

1.1 基于双阶段目标检测算法的改进

针对传统目标检测算法的低精度问题，Girshick 等人(2015)提出了基于深度学习的目标检测算法 R-CNN。R-CNN 首先使用选择性搜索(selective search)算法(Uijlings 等，2013)从待检测图像中提取 2 000 个左右的候选框，这些候选框可能包含要检测的目标；然后将所有的侯选框缩放成固定大小的尺寸(227 × 227 像素)；接着利用深度卷积神经网络对候选区进行特征提取，得到固定长度的特征向量；最后把特征向量送入支持向量机(support vector machine，SVM)分类器进行分类得到类别信息，送入全连接网络进行回归得到对应位置坐标信息。R-CNN 算法使目标检测的精度得到了质的改变，是将深度学习应用到目标检测领域的里程碑之作，同时也奠定了基于深度学习的双阶段目标检测算法的基础。

R-CNN 算法虽然设计巧妙，但是由于模型的检测过程分为多个阶段导致检测效率的极大消耗，因此 He 等人(2015)与 Girshick 等人(2015)分别基于 R-CNN 提出了尺度金字塔池化网络(spatial pyramid pooling net，SPPNet)和 Fast R-CNN 算法，这两种方法不需要将所有的候选窗口送入深度卷积神经网络，只需将图像送入深度网络一次，再将所有的候选窗口在网络中某层上进行映射，提升了检测速度，在 PASCAL VOC 2007 数据集上的检测精度从 R-CNN 算法的 66% 提高到 70%。

Fast R-CNN 算法中提及的测试时间不包含选择性搜索时间，而在实际测试时很大一部分时间都会用于选择性搜索。针对这个问题，Ren 等人(2018)提出了 Faster R-CNN 算法，在 Fast R-CNN 的基础上增加了候选窗口网络(region proposal network，RPN)，候选窗口网络通过设置不同尺度的锚(anchor)来提取候选框，代替了选择性搜索等传统的候选框生成方法，实现了网络的端到端训练，提高了网络计算速度。Faster R-CNN 网络由卷积层(conv layers)、RPN 网络、ROI(regions of interest) pooling 层、分类和回归层等 4 部分组成，如图 1 所示。其中卷积层用于提取整幅输入图像(image)的特征，并实现特征图(feature maps)的输出；RPN 网络用于提取候选区域(proposals)，其输入为卷积层得到的特征图，输出为多个候选区域(proposals)；ROI pooling 层用于将不同大小的输入转换为固定长度的输出；分类和回归层用于判断候选区域所属的类别以及候选区域在图像中的精确位置。相比于 Fast R-CNN，Faster R-CNN 的所有任务都统一在单一的深度学习框架之下，计算速度大幅度提升，在 PASCAL VOC 2007 数据集上的检测精度提高到 73.2%。

随着深度学习的不断发展，基于 Faster R-CNN 框架的检测方法受到基础网络复杂度、候选框数量、分类与回归子网络复杂度等因素的影响，计算量不断增大。对于上述问题，直接优化前两点性价比不高，直接优化感兴趣区域判断子网络(ROI-wise sub-

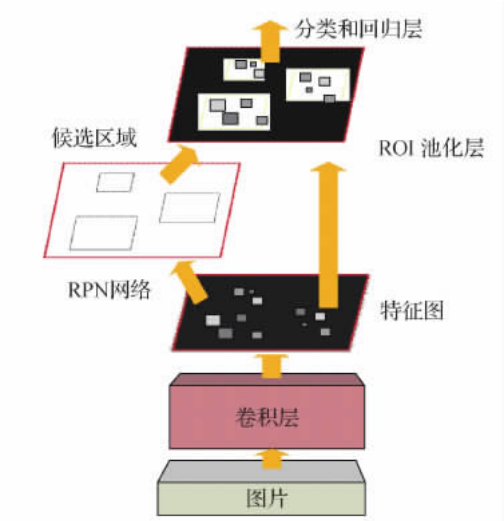


图 1 Faster R-CNN 结构图

Fig. 1 Architecture of Faster R-CNN

network)将子网络的深度减少又会造成基于分类任务的初始模型与目标检测的矛盾,因为分类要增加物体的平移不变性,目标检测则会减少物体的平移变化。微软亚洲研究院的Dai等人(2016)发现ROI池化后的网络层不再具有平移不变性,并且ROI池化后的层数会直接影响检测效率,因此提出了基于区域的全卷积网络(region based fully convolutional network, RFCN),通过位置敏感得分图(position-positive score maps)来解决这个矛盾。位置敏感得分图取消了感兴趣区域判断子网络,并利用位置敏感的ROI池化层,直接对池化后的结果进行判别,在PASCAL VOC 2007数据集上的检测精度从Faster R-CNN的73.2%提升到RFCN的80.5%。

Faster R-CNN在进行下采样与ROI pooling时都对特征图大小做了取整操作,这种做法对于分类任务基本没有影响,但对检测任务特别是对语义分割等像素级任务的精度影响较为严重,为此He等人(2017)遵循概念简单、效率和灵活性的原则提出了Mask R-CNN,对网络中特征图尺寸变化的环节都不使用取整操作,而是通过双线性差值填补非整数位置的像素。这使得下游特征图向上游特征图映射时没有位置误差,不仅提升了目标检测效果,还使得算法能满足语义分割任务的精度要求。

除上述方法外,还有从样本后处理的角度去改进与优化双阶段目标检测算法性能的方法,如非极大值抑制(non maximum suppression, NMS)算法、Soft-NMS(Bodla等,2017)、Softer-NMS(He等,2018)。NMS算法是绝大多数目标检测方法中必备的后处理步骤,首先将一定区域内的检测框按照得分高低进行排序,然后保留得分最高的框,同时删除与该框重叠面积大于一定比例(使用预先定义的阈值)的其他框,最后将此过程递归地应用于其余的框,从而保证为每个目标找到检测效果最好的一个框。这种方法存在两个问题:1)NMS的阈值不容易确定,设置小了会删除不该删除的框,设置大了会增加误检率;2)不满足要求的框会被直接删除,处理方式过于简单。针对这些问题,Bodla等人(2017)提出了Soft-NMS算法,其整体流程与NMS相似,不同之处在于不是直接删除所有重叠率(intersection over union,IOU)大于阈值的检测框,而是新设定一个置信度阈值,待处理框与最高得分框的IOU越大,该框置信度的得分就越低,最后得分大于置信度阈值的检测框可以保留,这样可以提高目标检测算法的召回率。为了进一步提高目标位置的预测精度,He等人(2018)提出了Softer-NMS算法,使用一种新的边框回归损失函数(Kullback-Leibler, KL)Loss,该损失函数可以同时学习到边框的形变量和位置变化量。同时Softer-NMS将KL Loss使用在基于权重平均的Soft-NMS算法上;最终,Softer-NMS算法在MS COCO数据集上,将基于VGG-16的Faster R-CNN的检测精度从23.6%提高到29.1%,将基于ResNet-50的Fast RCNN的检测精度从36.8%提高到37.8%。

1.2 基于单阶段目标检测算法的改进

针对双阶段目标检测算法的低效问题,YOLO(you only look once)v1(Redmon等,2016)舍去了算法中的候选框提取分支,直接将特征提取、候选框分类和回归在同一个无分支的深度卷积网络中实现,使得网络结构变得简单,检测速度从Faster R-CNN的7帧/s提升到了45帧/s,使得基于深度学习的目标检测算法在当时的计算能力下开始能够满足实时检测任务的需求。随着YOLOv1的出现,基于深度学习的目标检测算法开始有双、单阶段之分。YOLOv1的整体思路比较简洁,算法的网络结构图如图2所示。$s \times s$表示输入图像被划分成的网格个数,s代表图像的长或宽被平均分成s份。YOLOv1的核心在最后两层,卷积层之后接4 096维全连接层,然后又全连接到$7 \times 7 \times 30$维的张量上,整个过程不需先确定中间的候选区,单一网络即可完成类别的判定和位置的回归。

YOLOv1舍弃了候选框阶段,加快了检测速度,但是算法的定位和分类精度相对较低。Liu等人(2016)提出的SSD(single shot multiBox detector)在很大程度上平衡了单阶段目标检测算法的检测速度和检测精度。SSD采用VGG16作为模型基础,新增卷积层来获得更多的特征图用于检测。第1步与其他单阶段目标检测方法相同,利用卷积操作提取特征图,最后几层卷积对每一尺度上的特征图运用anchor方法进行候选框提取,依据anchor在不同尺度上得到的候选框,进行目标种类和位置的判断。与Faster R-CNN不同,SSD中的anchor分布在不同的特征图上,利用多层的特征达到多尺度要求。同时在SSD框架图中,若某一层特征图大小是8 × 8像素,就使用3 × 3像素的滑窗提取每个位置的特征,

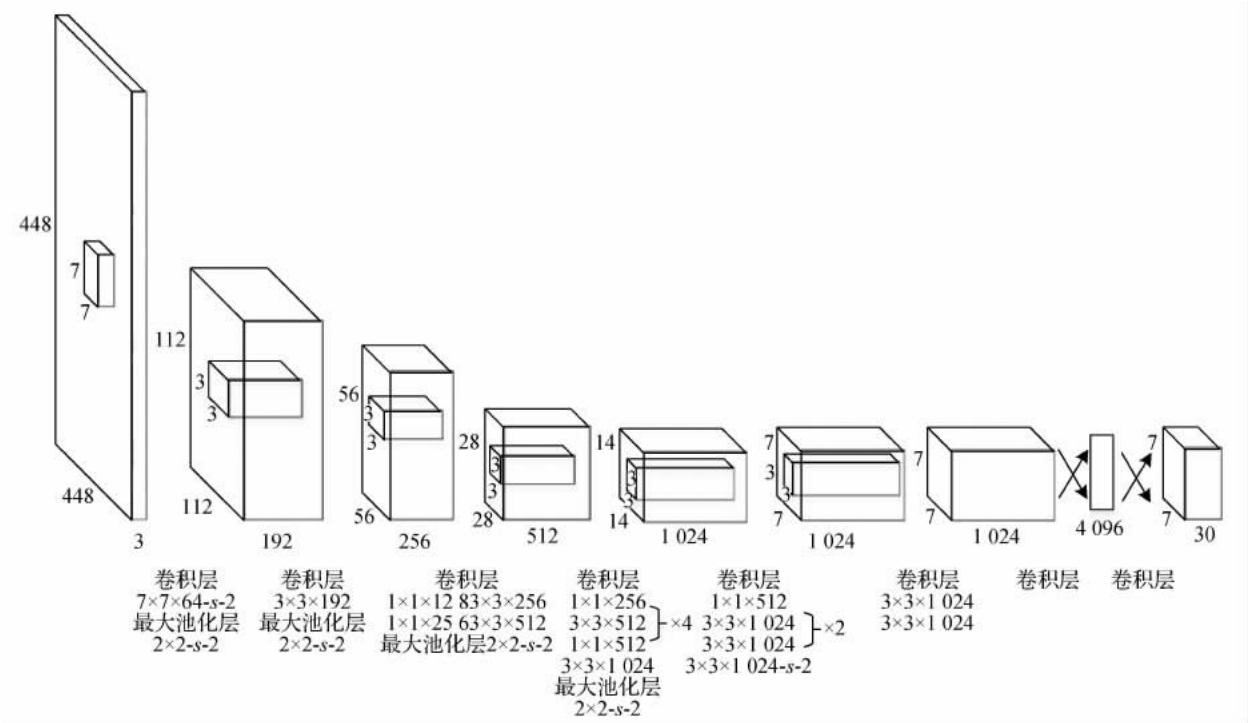


图 2　YOLOv1 结构图

Fig. 2　Architecture of YOLOv1

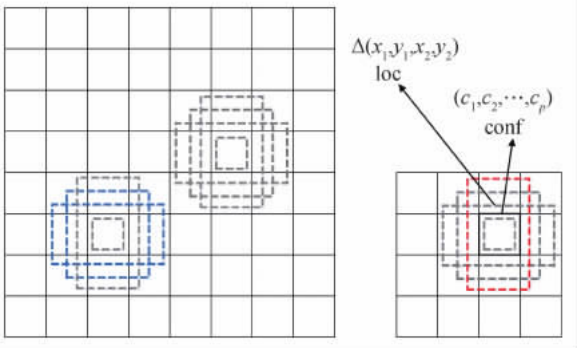


图 3　SSD 框架图

Fig. 3　SSD framework

这个特征回归得到目标的坐标信息(x_1,y_1,x_2,y_2)和类别信息(c_1,c_2,$\cdots$,c_p),如图 3 所示。最终在 PASCAL VOC 2007 数据集上 SSD 的检测精度从 YOLOv1 的 66.4% 提升到 75.1%,同时检测速度也提升到 58 帧/s。

YOLOv1 虽然检测速度较快,但是它在物体定位(localization) 方面不够准确,并且召回率(recall) 较低,因此它的检测精度较低。针对这一问题,Redmon 和 Farhadi(2017) 提出了 YOLOv2 的改进模型。YOLOv2 算法主要利用批归一化(batch normalization) 、高分辨率分类器(high resolution classifier) 、直接目标框位置检测(location prediction) 、多尺度训练(multi-scale) 等操作来提高模型的检测精度,最终将模型在 PASCAL VOC 2007 数据集上的检测精度从 YOLOv1 的 66.4% 提升到 78.6%。同时,YOLOv2 还专门训练了一个由 19 个卷积层(conv) 和 5 个最大池化(maxpooling) 层组成的 Darknet-19 网络作为模型的骨干网络,用于提取特征并减少模型的计算量。与 VGG 模型设计原则一致,Darknet-19 的卷积层主要使用 3 × 3 卷积,池化层使用 2 × 2 最大池化,最大池化之后,特征图的维度降低 2 倍,通道数增加 2 倍; 与 NIN(network in network) (Lin 等,2014) 类似,Darknet-19 最终采用全局平均池化(global average pooling) 做预测,并且在 3 × 3 卷积之间使用 1 × 1 卷积来压缩特征图通道数以降低模型计算量和参数; 同时,Darknet-19 每个卷积层后面使用归一化层加快收敛速度,降低模型过拟合概率,最终 Darknet-19 可以在不降低检测精度的前提下将 YOLOv2 的计算量降低 33%。

YOLOv3(Redmon 和 Farhadi,2018) 在 YOLOv2 的基础上,使用全新设计的 Darknet-53 残差网络并结合特征金字塔网络(feature pyramid networks, FPN) (Seferbekov 等,2018) 进行多尺度融合预测,其基本思想是先利用特征提取网络得到一定尺寸的特征图(如 13 × 13) ,然后将输入图像分成对应个

数(13 × 13)的网格单元,如果真实目标的中心坐标落在某一网格单元,则由该网格单元来预测该目标,因为每个网格单元都会预测固定数量的边界框(采用 YOLOv2 中的 K 均值聚类算法(K-means)获得 3 个初始尺寸不同的边界框),最终选择与真实值的 IOU 最大的边界框来预测该目标。YOLOv3 的 Darknet-53 相对于 YOLOv2 的 Darknet-19 改进了两个方面:1) YOLOv3 中做特征图尺寸变化的池化(pooling)层基本由卷积层来实现,减少了模型的运算量;2) 针对 YOLOv2 中直筒型网络结构层数太多所产生的梯度问题引入了 ResNet 网络中的残差结构(residual blocks),ResNet 的残差结构训练深层网络的难度较小,因此可以将网络做到 53 层来提升检测精度,这些改变使得 YOLOv3 用 1/3 的时间达到与 SSD 相当的精度。边界框的坐标预测方式沿用了 YOLOv2 的做法(如图 4 所示),其中 t_x,t_y,t_w,t_h 表示模型的预测输出,c_x 和 c_y 表示网格单元的坐标,p_w 和 p_h 表示预测前边界框的尺寸,b_x,b_y,b_w,b_h 就是预测得到的边界框的中心坐标和尺寸,坐标的损失采用平方误差损失。$\delta(t_x)$ 和 $\delta(t_y)$ 表示在 x 轴和 y 轴上预测得到的边界框的中心坐标相对于网络单元坐标的距离。

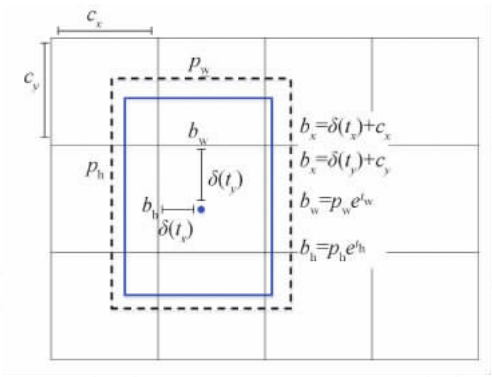


图 4 具有尺寸先验和位置预测的边界框

Fig. 4 Bounding box for size prior and position prediction

1.3 基于单阶段、双阶段目标检测算法的结合

双阶段目标检测算法可以利用其精度优势不断改进与创新,提高检测速度;单阶段目标检测算法则可以利用其速度优势,不断提升模型的检测精度,从而达到对检测速度和精度的需求。目前,双阶段目标检测算法和单阶段目标检测算法都已经具备了一定的理论基础,研究人员已开始关注将二者结合作为一种高效平衡检测精度和检测速度的方式。

RON(reverse connection with objectness prior networks)(Kong 等,2017)算法是在单阶段目标检测算法 SSD 和双阶段目标检测算法 Faster R-CNN 的基础上提出的高效通用的目标检测模型。与 SSD 相似,RON 使用 VGG-16 网络作为骨干网络,不同之处在于 RON 将 VGG-16 网络的第 14 与第 15 全连接层变成了核大小为 2 × 2、步长为 2 的卷积层以实现特征图降采样处理,同时采用反向连接方式将相邻位置的待检测特征图联系起来,实现多尺度目标定位并达到更好的效果;与 Faster R-CNN 相似,RON 提出了与 RPN 网络思想类似的目标先验(objectness prior)策略来引导目标对象搜索,同时利用多任务损失函数优化整个网络。在最终测试中,RON 达到了较高检测精度,同时在相同条件下检测速度比 Faster R-CNN 快了 3 倍。

RefineDet(Zhang 等,2018c)融合了 RPN 网络、FPN 算法和 SSD 算法,是单阶段和双阶段目标检测算法结合的典型代表,可以在保持 SSD 高效的前提下将在 PASCAL VOC 2007 数据集上的检测精度提高到 80.0%。RefineDet 的核心思想包括 3 个方面:1) 引入双阶段目标检测算法中对框的由粗到细的回归思想,即通过 RPN 网络得到粗粒度的框信息,然后通过常规的回归支路进行进一步回归,从而得到更加精确的框信息;2) 引入类似 FPN 网络的特征融合操作用于网络检测,可以有效提高小目标的检测效果;3) 检测网络以 SSD 为框架,保证了模型的检测速度。与核心思想相对应,RefineDet 网络结构主要由锚框优化模块(anchor refinement module,ARM)、转换连接模块(transfer connection block,TCB)和目标检测模块(object detection module,ODM)组成,如图 5 所示。

ARM 模块类似于 Faster R-CNN 算法中的 RPN 网络,主要用来得到候选框和去除一些负样本,与 RPN 网络的不同之处是算法的输入利用了多层特征,而 RPN 网络的输入是单层特征。ARM 基于4 层特征最后得到两条支路,一条是坐标回归支路,另一条是二分类支路。TCB 模块是进行特征的转换操作,即将 ARM 部分的输出特征图转换成 ODM 模块的输入,类似于 FPN 算法的特征融合。ODM 模块类似于 SSD,用来融合不同层的特征,然后进行分类和回归。改进的部分为输入 anchors 是 ARM 模块得到的预处理 anchor,类似 RPN 网络输出的候选框,同时与 FPN 算法类似,浅层特征图融合了高层特征图的信息,然后预测候选框是基于每层特征图进行,最后将各层结果再整合到一起,对小目标物体的检测效果更好。

RefineDet 的损失函数主要包含 ARM 和 ODM

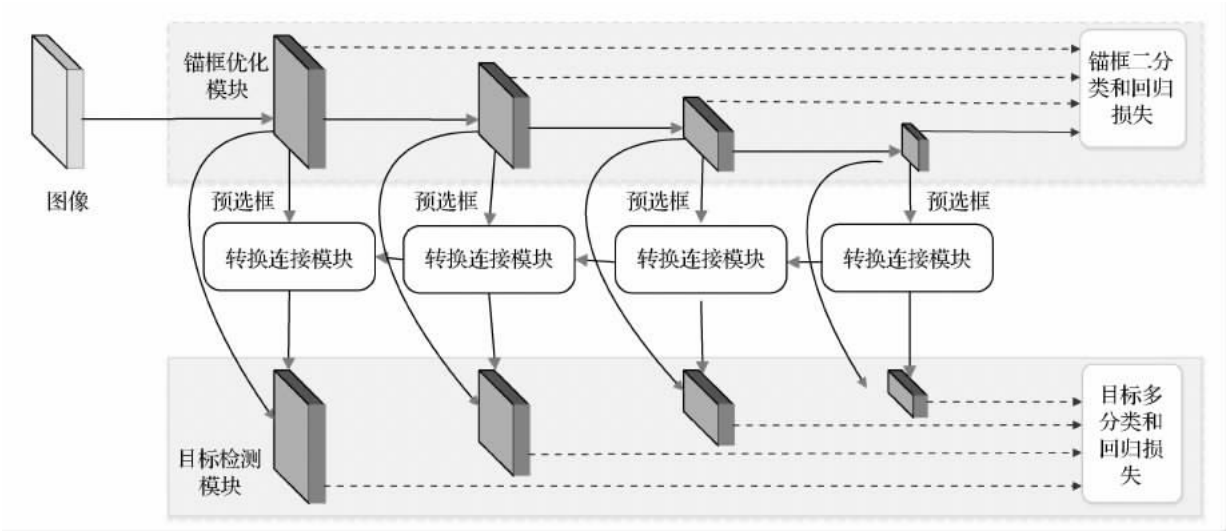


图 5　RefineDet 网络结构图

Fig. 5　RefineDet network structure

两方面，可表示为

$$L(\{p_i\},\{x_i\},\{c_i\},\{t_i\}) = \frac{1}{N_{\mathrm{ARM}}}\left(\sum_i L_{\mathrm{b}}(p_i, [l_i^* \geqslant 1]) + \sum_i [l_i^* \geqslant 1] L_{\mathrm{r}}(x_i, g_i^*)\right) + \frac{1}{N_{\mathrm{ODM}}}\left(\sum_i L_{\mathrm{m}}(c_i, l_i^*) + \sum_i [l_i^* \geqslant 1] L_{\mathrm{r}}(t_i, g_i^*)\right) \tag{1}$$

式中，N_{ARM} 代表 ARM 中正样本的数目，N_{ODM} 代表 ODM 中正样本的数目，p_i 代表置信度，x_i 代表 ARM 细化后预测的坐标，c_i 代表 ODM 中预测的物体类别，t_i 代表 ODM 中预测的框的坐标，l_i^* 代表目标的真实类别标签，g_i^* 代表目标的真实位置和大小。ARM 部分包含式中二分类(binary classification)损失 L_{b} 和回归损失 L_{r}，ODM 部分包含式中多类别分类(multi-class classification)损失 L_{m} 和回归损失 L_{r}，同时 ARM 与 ODM 两个部分的损失函数一起进行前向传递。

2　提升小目标物体检测精度

小目标通常存在两种定义，一种是相对尺寸的定义，即目标尺寸低于原图像尺寸的 10%，另外一种是绝对尺寸的定义，即目标尺寸小于 32 × 32 像素。在目标检测中，小目标物体检测精度低的原因主要是：1) 现有的目标检测算法大多是针对通用目标进行设计的，导致小目标的检测效果较差；2) 由于分辨率和信息的限制，小目标检测一直存在技术瓶颈。根据研究思路的不同，本文从骨干网络、增加视觉感受野、特征融合、级联卷积神经网络和模型训练方式 5 个角度对小目标检测算法的研究进展进行综述。

2.1　骨干网络

主流的基于深度学习的目标检测算法大都使用基于 ImageNet 数据集训练的 VGG-16(Simonyan 和 Zisserman，2014)、GoogLeNet、ResNet-50 及 ResNet-101 等网络作为骨干网络，这些骨干网络的作用是提取图像中的特征信息用于后续的融合、分类、回归等操作，但是对图像中的小目标无法进行高精度检测。针对这一问题，研究者从底层出发，在原有骨干网络的基础上不断提出新的骨干网络，如 DetNet 和 DenseNet 等，以适应小目标物体的检测需求。

DetNet 是 Li 等人(2018)基于 ResNet-50 提出的骨干网络，DetNet 在保留 ResNet-50 的前 3 个阶段的基础上，修改了第 4 阶段和第 5 阶段，增加了第 6 阶段，结构图如图 6 所示。A 表示具有特征金字塔的特征提取网络，B 表示图像分类网络，14 × 14 GAP 表示特征图尺寸大小为 14 × 14，GAP 为单位。ResNet-50 中阶段 5 的特征图是原图的 1/32，而 DetNet 中阶段 4、5、6 的特征图都是原图的 1/16，这样做可以保证特征图的尺度足够大，从而保证小物体不会消失以及特征图保存更清晰的边缘信息，同时为了减少模型的计算量和内存，DetNet 并没有像传统的骨干网络

一样,每个阶段通道数目都增加一倍,而是在第 5 阶段和第 6 阶段保持了相同的通道数,这种网络结构可以在提高检测效率的同时将在 MS COCO 数据集上的检测精度从 ResNet-50 的 35.8% 提高到 40.2%。

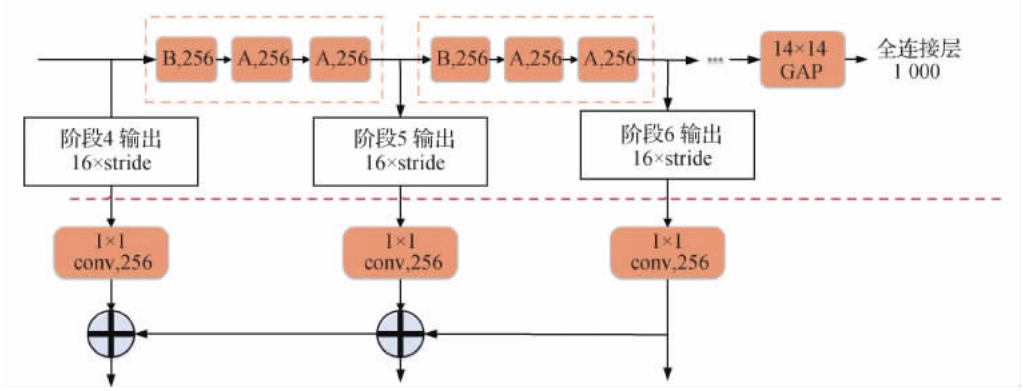


图 6 DetNet 网络结构图

Fig. 6 DetNet network structure

Huang 等人(2017)提出的 DenseNet 骨干网络摆脱了传统的以加深网络层数和加宽网络结构来提升网络性能的思维定式,从特征的角度出发,通过密集块(dense block)的使用和密集块之间的连接,大幅减少了网络的参数量,并在一定程度上缓解了梯度消失(gradient vanishing)和模型退化(model degradation)问题,使网络更易于训练并且可以提取到更有效的特征,进而提高模型的检测精度。在单个密集块中,为了保证层与层之间最大程度的信息传递,DenseNet 直接将所有的层连接起来,即将前面每一层得到的特征图进行融合,然后输入到后面的层中。同时,由于在 DenseNet 中需要对不同层的特征图进行融合操作,需要不同层的特征图保持相同的尺寸大小,限制了网络中下采样(down sampling)的实现。为解决这一问题,DenseNet 网络被分成多个密集块,且在同一个密集块中要求特征图尺寸保持相同大小,在不同密集块之间设置转换层(transition layers)实现下采样。

在 DenseNet 网络的基础上,Zhou 等人(2018b)提出了 STDN(scale transferrable object detection)算法。STDN 采用 DenseNet 网络作为特征提取网络,提高网络的特征提取能力,同时引入尺寸转换层(scale-transfer layer),将已有的特征图转换成大尺寸的特征图,与反卷积(deconvolution)、上采样(upsample)等尺寸转换方法不同,尺寸转换层可以在几乎不增加参数量和计算量的前提下生成大尺寸的特征图,最后利用多层特征做预测,并对预测结果融合得到最终的检测结果。STDN 算法通过对 DenseNet 网络的使用和相关改进,可以在保证检测速度的前提下,将在 PASCAL VOC 2007 数据集上的检测精度从以 ResNet-101 为骨干网络的 DSSD(deconvolutional single shot detector)算法(Fu 等,2017)的 78.6% 提高到 79.3%。

2.2 增加视觉感受野

深度学习中感受野(receptive fields)是指卷积神经网络每一层输出特征图上的像素点在原始输入图像上映射的区域大小。在深度学习的其他领域已经采用通过增加感受野的方式来提高性能,比如在人体姿态估计中利用大的感受野来学习长距离的空间位置关系,建立内隐空间模型(implicit spatial model)。在目标检测领域针对小目标的信息限制,同样可以通过增加视觉感受野来保证特征图的尺寸,从而提高网络的特征提取能力得到更多的目标信息,进而提高小目标物体检测精度。

通过增加感受野提高目标检测网络的特征提取能力,首次在感受野块(receptive field block,RFB)网络(Liu 等,2018b)中使用,主要是在 SSD 架构中引入感受野块,如图 7 所示。图中 pRF in hV4 表示不同感受野的特征图在偏心率为 hV4 的条件下组合在一起形成了感受野的空间阵列。感受野块通过模拟人类视觉的感受野加强网络的特征提取能力,因此可以在兼顾速度的同时将在 PASCAL VOC 2007 数据集上的检测精度提高到 80.5%。在结构上,RFB 借鉴了 Inception(Szegedy 等,2016)的思想,主要是在 Inception 的基础上加入了加宽卷积层(又称膨胀卷积或空洞卷积),从而有效增大了感受野。RFB 结构主要有两个特点:1)不同尺寸卷积核的卷积层构成的多分支结构,这部分改进参考了 Inception 思想,图 7 中用不同大小的圆形表示不同尺寸卷积核的卷积层;2)引入了加宽(dilated)卷积层,加宽卷积层曾应用在分割算法 DeepLab(Chen 等,2018)中,主要作用也是增加感受野,与 Deformable CNN(Dai 等,2017)类似,RFB 结构也用不同比例表

示加宽卷积层的参数。在 RFB 结构中,最后将不同尺寸和比例的卷积层输出进行结合,达到融合不同特征的目的。RFB 结构中用 3 种不同大小和颜色的输出叠加来展示,图 7 最后一列将融合后的特征与人类视觉感受野做对比,结果非常接近。通过引入 RFB 模块,目标检测算法将在 PASCAL VOC 2007 数据集上的检测精度从 SSD 的 75.1% 提高到 RFB Net 的 80.5%,特别是对小目标的检测效果提升显著。

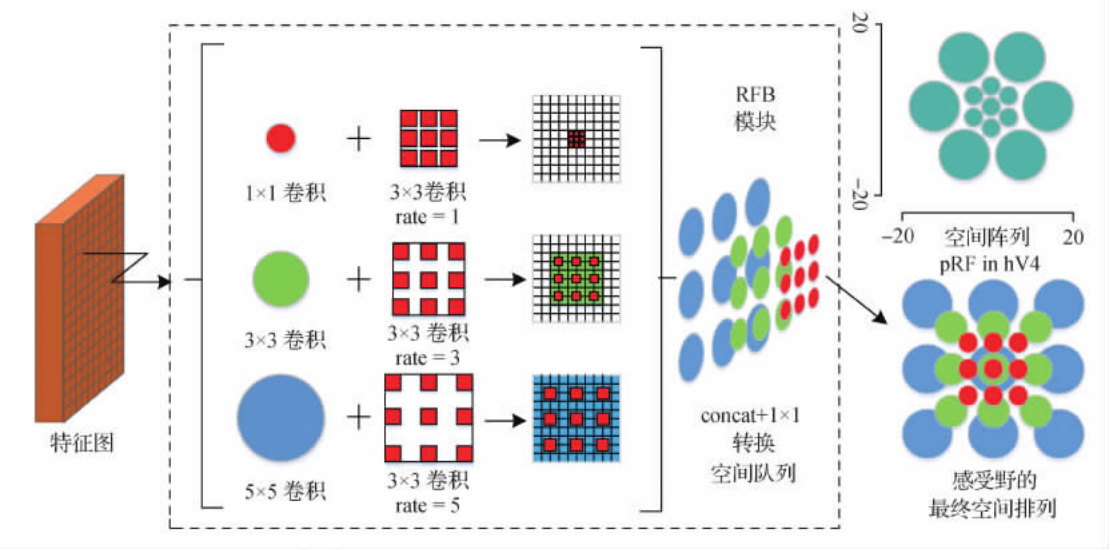


图 7 RFB 效果示意图

Fig. 7 RFB effect diagram

在 RFB 的基础上增加视觉感受野还有另外一种结构形式: RFB-s,其主要有两处改进: 1) 用 3 × 3 卷积层代替 5 × 5 卷积层; 2) 用 1 × 3 和 3 × 1 卷积层代替 3 × 3 卷积层,主要目的是减少计算量,类似 Inception 后期版本对 Inception 结构的改进,如图 8 所示。

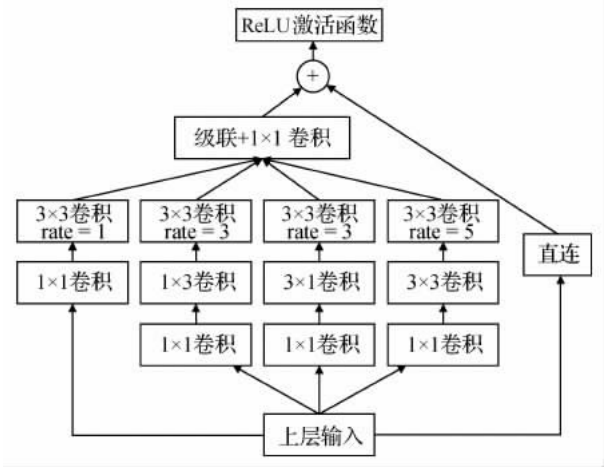


图 8 RFB-s 结构示意图

Fig. 8 The architectures of RFB-s

TridentNet 算法(Li 等,2019)同样从感受野入手,以 ResNet-101 网络作为骨干网络,通过引入空洞卷积增加网络的感受野,实现不同尺度目标的检测,最终提高小目标物体的检测精度。TridentNet 算法通过多分支结构、权重共享(weight sharing among branches)和指定尺度过滤训练(scale-aware training scheme) 3 个方面的改进,使得算法中的多分支结构与 RFB Net 中的多分支结构类似,即将原本特征提取网络的单支路卷积层替换成 3 支路卷积层,同时每个支路卷积层又由不同加宽参数的空洞卷积层组成。权重共享是指 3 个支路的卷积层参数是共享的,权重共享一方面可以减少网络前向计算的时间,另一方面可以使网络学到的参数具有更好的泛化能力,此外还能够帮助实现快速预测(inference)。TridentNet 中使用一种快速预测做法,即选择一个分支的输出作为最终结果,假如没有权重共享,那么单分支的结果很难近似多分支结果。指定尺度过滤训练是指 3 个支路分别检测不同尺度的目标。由于不同尺度的目标对应的网络最佳感受野不同,因此可以为这 3 条支路分配不同尺度的目标,这种方法可以减少每条支路训练的目标尺寸差异,提高检测精度。

2.3 特征融合

随着计算机性能的大幅提升以及深度学习的快

速发展,特征融合的优势越来越明显,特征融合算法主要分为 3 类: 基于贝叶斯决策理论的算法、基于稀疏表示理论的算法和基于深度学习理论的算法。在目标检测中,主要利用基于深度学习理论的算法,即将多个神经网络得到的多类别特征进行融合得到融合的特征。通过特征融合可以得到更多的小目标信息,从而提高相应的检测精度。

在特征金字塔网络(feature pyramid networks, FPN)之前,大多数的目标检测方法与分类方法一样,都使用单层的特征来进行处理,并没有将高层的语义信息添加到低层特征图,但在目标检测中,高层的语义信息特别重要,FPN 采用多尺度特征融合的方式,对高低层特征图进行融合,在 MS COCO 测试集上将小目标物体的检测精度提高到 35.8%,同时 FPN 也被 YOLOv3 等单阶段目标检测方法使用。

整个 FPN 网络同样可以嵌入到 RPN 中,生成不同尺度的特征并融合作为 RPN 网络的输入,以提高双阶段目标检测算法的精度,具体方式如图 9 所示。每一个金字塔层级应用单尺度的 anchor,{$P2$, $P3$, $P4$, $P5$, $P6$} 分别对应的 anchor 尺度为{32^2, 64^2, 128^2, 256^2, 512^2},同时每层都有 3 个长宽对比度,即: 1∶2、1∶1、2∶1,这样整个特征金字塔有15 种 anchor,因此可以融合更多的特征信息,提高算法的检测精度。

由于目标检测算法通常只包含图像的特征信息,并没有图像的语义信息,Zhang 等人(2018b)基于 SSD 框架融合图像的低层与高层特征语义信息提出了 DES(detection with enriched semantics),通过提高低层特征图的语义信息的分割模型(segmentation module)与提高高层特征图的语义信息的总体激活模型(global activation module)解决了 SSD 中对于小目标物体的检测效果较低的问题,将在 PASCAL VOC 2007 数据集上的检测精度提升到 84.3%。DES 的具体网络结构主要分为 3 个方面: 1)与 SSD 架构基本相同的检测分支; 2)分割模型,以 conv4_3 和边界框层的真实值作为输入来增强低层特征图的语义信息; 3)总体激活模型利用注意力机制(attention mechanism)(Qi 等,2017)的思想来提高高层特征图的语义信息,网络结构如图 10 所示。

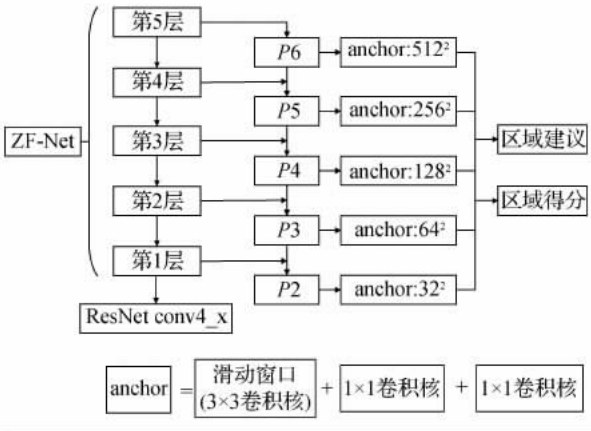


图 9 融合 FPN 的 RPN 网络

Fig. 9 RPN network with FPN

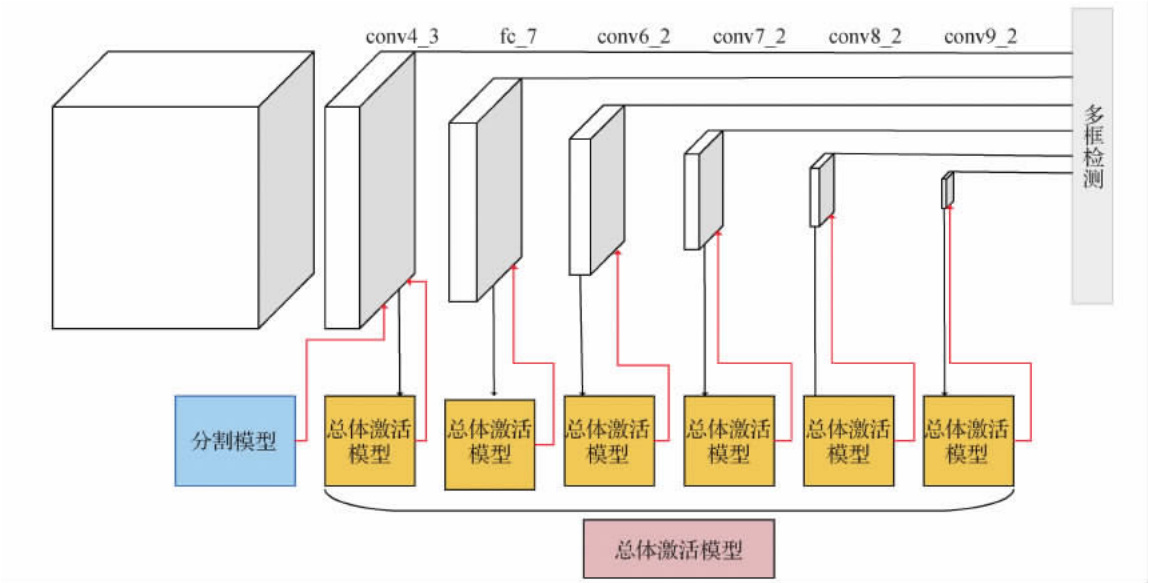


图 10 DES 网络结构图

Fig. 10 DES network structure

FPN 算法可以将高层的语义信息与低层的图像信息相结合,提升网络的特征提取能力,提高检测精度,但 FPN 网络架构是人工设计的,融合效果并非最好,因此 Ghiasi 等人(2019)利用神经架构搜索(neural architecture search,NAS)(Zoph 和 Le,2017)技术生成了一个新的特征金字塔网络 NAS-FPN,生成过程是利用强化学习训练控制器在一个包含所有跨尺度连接的可扩展搜索空间中选择最优的特征金字塔架构,控制器利用子模型在搜索空间中的准确度作为奖励信号来更新其参数,通过不断迭代,得到最优的特征金字塔模型架构。最终,NAS-FPN 与 RetinaNet(Lin 等,2017)框架中的若干骨干模型相结合,实现了当前目标检测模型的较高准确率,在 MS COCO 数据集上的检测精度达到了 48.3%。

2.4 级联卷积神经网络

越复杂的深度卷积神经网络从大量的数据中自动学习高层特征的能力越抽象,越容易找到高位特征之间的关联以及小目标的相关信息,因此通过级联卷积神经网络提高检测模型的复杂度,从而提高网络的特征提取能力得到更多的小目标信息,最终提高小目标的检测精度。

在 Fast R-CNN 的基础上,Faster R-CNN 采用两阶段级联 RPN 网络进行优化,第 1 阶段是 RPN 子网络(H0),第 2 阶段是以 Fast R-CNN 为基础的感兴趣区域检测子网络(H1),如图 11(a)所示,级联网络使 Faster R-CNN 在 PASCAL VOC 2007 数据集上的检测精度从 70% 提升到 73.2%。

Intertive BBox(intertive BBox at inference)在检测阶段同样采用了级联的网络架构,完成了迭代式的边界框回归,将前一个检测模型回归得到的边界框坐标初始化为下一个检测模型的边界框,迭代 3 次(H1),如图 11(b)所示,提升了小目标检测的精度。

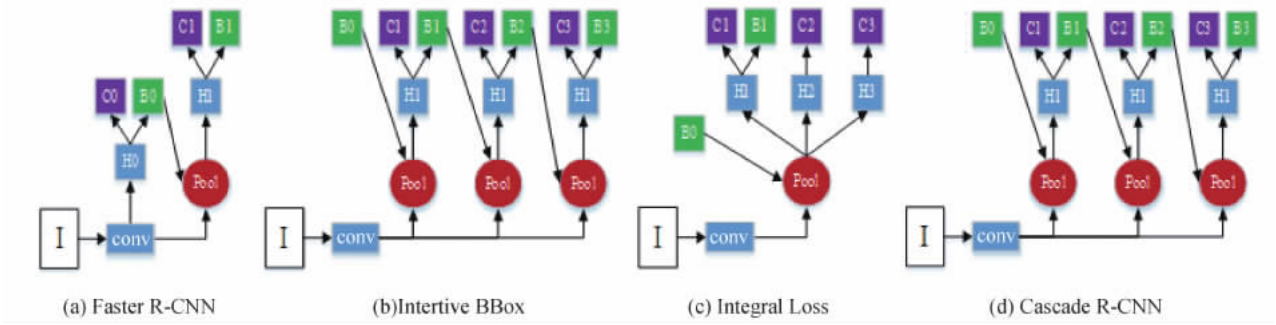
针对迭代式边界框回归的单一 IOU 阈值问题,Integral Loss(Zagoruyko 等,2016)(图 11(c))用 3 条支路(H1,H2,H3)并行的策略,对输出边界框的标签界定采取不同的 IOU 阈值来提高小目标物体的检测精度。

基于 Faster R-CNN、Intertive BBox 和 Integral Loss 架构,Cai 和 Vasconcelos(2018)提出了改进算法 Cascade R-CNN。与 Intertive BBox 不同,检测模型是基于前一阶段的输出进行训练,而不是基于最初始的数据。与 Integral Loss 的差别在于 Cascade R-CNN 中每个阶段的输入边界框是前一阶段边界框的输出,而不仅仅是基于不同重叠度阈值训练得到。同时针对单一的 IOU 阈值难以实现大范围跨度的 IOU 检测的问题,Cascade R-CNN 结构包含了不同 IOU 阈值的回归网络,如图 11(d)所示,同时提高了 IOU 阈值下的正样本比例,改善了训练中正负样本不平衡的问题。Cascade R-CNN 对特征图进行多次检测,通过级联不同阈值的检测网络得到不同的检测结果,最终在 MS COCO 测试集上的检测精度提升到 42.8%。Cascade R-CNN 的结构在损失函数中也有着良好的体现,损失函数为

$$L(x^t,g) = (h_t(x^t),y^t) + \lambda\,[y^t \geqslant 1]\,L_{\text{loc}}(f_t(x^t,b^t),g) \tag{2}$$

式中,x^t 代表对应的输入,g 是 x^t 对应的真实值,λ 是折中系数,y^t 是 x^t 的标签,$b^t = f_{t-1}(x_{t-1},b_{t-1})$。

深度神经网络的特征抽取导致特征图的尺寸不断缩减,解决这个问题的一个突破口是如何获得具有高清表示信息的特征来提升关键点检测的精度。例如,反卷积和空洞卷积都是在下采样的基础上进



(a) Faster R-CNN (b)Intertive BBox (c) Integral Loss (d) Cascade R-CNN

图 11 Faster R-CNN、Interative BBox、Integral Loss 和 Cascade R-CNN 算法框架体系结构

Fig. 11 Algorithm framework architectures of Faster R-CNN, Interative BBox, Integral Loss and Cascade R-CNN ((a) Faster R-CNN; (b) Intertive BBox; (c) Integral Loss; (d) Cascade R-CNN)

行高清信息的恢复,精度提升不理想。针对此问题,中国科技大学和微软亚洲研究院的研究人员联合提出了 HRNet 算法(Sun 等,2019a),该算法的整体思路是将采用高分辨率的子网络作为第 1 阶段,然后逐渐并行添加较低分辨率的子网络,得到多个阶段的子网络的输出,其中第 1 阶段的每一个高分辨率特征图表示都可以一次又一次地从其他并行表示分支接收信息,从而得到信息更丰富的高分辨率表示。最终第 1 阶段子网络的输出是通过多次加入低分辨率融合而形成的一种高清表示,关键点检测更精确,空间分辨率精度更高。在 MS COCO 数据集上进行的关键点检测、姿态估计、多人姿态估计这 3 项任务中,HRNet 方法都超越了已有的算法。

2.5 模型训练方式

在图像的内容检测中,小目标检测占有极大比例,而现有的目标检测算法多是针对通用目标数据集进行设计的,训练集与测试集的巨大差异直接影响了小目标物体的检测精度。因此一些研究通过更改模型的训练方式提高小目标物体的检测精度。

YOLOv2 考虑到在 ImageNet 数据集上预训练模型时输入图像大小是 224 × 224 像素,与检测网络用的图像 416 × 416 像素存在较大差别,所以将预训练模型在 416 × 416 像素的 ImageNet 数据集上继续预训练,然后再用检测模型提取特征,通过这种训练方式实现了预训练模型和检测模型的良好过渡,将在 PASCAL VOC 2007 数据集上的目标检测精度提升了 12.2%。

ImageNet 和 MS COCO 数据集在物体的尺寸分布上同样存在较大差异:1) ImageNet 数据集中的目标尺寸占图像尺寸的百分比(中值 50%)为55.6%,而 MS COCO 数据集中仅有 10.6%;2) MS COCO 数据集中的小目标比 ImageNet 数据集多;3) MS COCO 数据集中物体尺寸变化范围较大。以上 3 点共同导致了目标检测算法在 MS COCO 数据集上的低准确率。Singh 等人(2018)基于对数据集的分析,提出一种新的模型训练方式:图像金字塔的尺寸归一化(scale normalization for image pyramids,SNIP),主要包含两个改进点:1) 借鉴多尺寸训练思想,引入图像金字塔来处理数据集中不同尺寸的数据;2) 为了减少训练集和测试集的差异,在梯度回传时只将与预训练模型所基于的训练数据尺寸相对应的 ROI 的梯度进行回传。从而将在 MS COCO 数据集上的检测精度提高到 48.3%。SNIP 中无论是训练检测器还是 RPN 网络,都是基于所有标定的真实值来定义 proposal 和 anchor 的标签,某个 ROI 在某次训练中是否回传梯度与预训练模型的数据尺寸相关,即当某个 ROI 的面积在指定范围内时,该 ROI 就是有价值的,会在此次训练中回传梯度,否则就是无效的,RCN 为分类器,用于判断检测出的候选框是否满足要求。具体信息如图 12 所示。

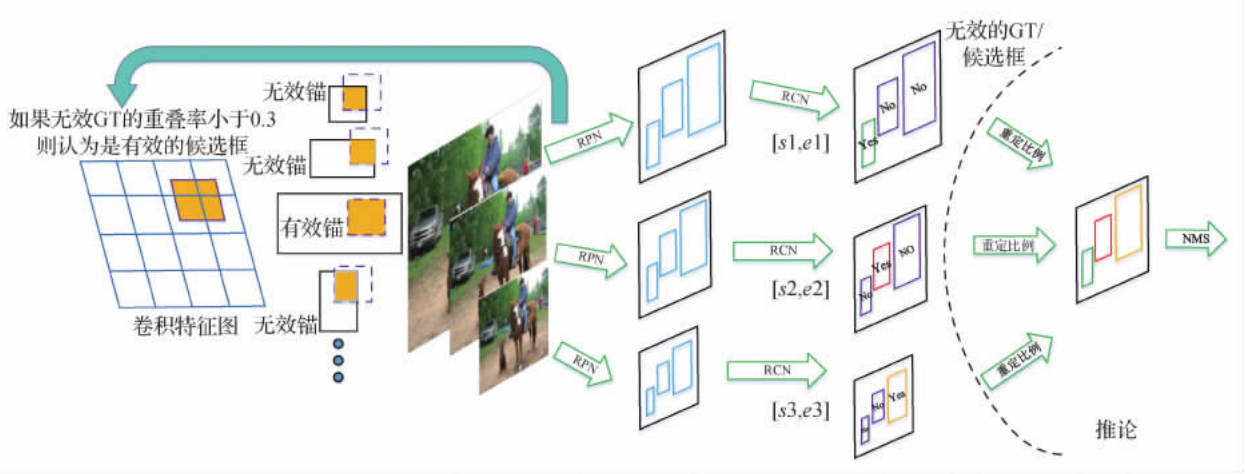


图 12 SNIP 网络结构示意图

Fig. 12 Schematic diagram of SNIP network structure

生成式对抗网络(generative adversarial networks,GAN)(Goodfellow 等,2014)是一种深度学习模型,是近年来复杂分布上无监督学习最具前景的方法之一。GAN 通过生成模型(generative model)和判别模型(discriminative model)的互相博弈学习产生最佳输出。在训练过程中,生成网络的目标就是尽量生成真实的图像去欺骗判别网络,而判别网络的目标就是尽量将生成网络生成的图像与真实的

图像分别开来，这样生成网络和判别网络就构成了一个动态的博弈过程，进而得到更好的输出。GAN的应用需要有良好的训练方式，否则可能会因为神经网络模型的自由性而导致输出不理想。Li等人(2017)提出的感知生成式对抗网络(perceptual GAN)为GAN带来了一种很好的训练方式。perceptual GAN将所有的物体分为大物体和小物体两个子集，通过挖掘不同尺度物体间的结构关联，提高小物体的特征表示，使之与大物体类似，从而提高小物体检测率。perceptual GAN包含两个子网络：生成网络(小物体)和判别网络(大物体)。生成网络是一个深度残差特征生成模型，通过引入低层细粒度的特征将原始较差的特征转换为高分辨的特征；判别网络包含两个分支：对抗分支和感知分支，对抗分支用来分辨小物体生成的高分辨率特征与真实的大物体特征，感知分支则通过感知损失提升检测率。perceptual GAN的目标函数为

$$\min_G \max_D L(D,G) = E_{x\sim p_{\text{data}}(x)} \log D(x) + E_{z\sim p_z(z)} [\log(1 - D(G(z)))] \tag{3}$$

式中，G表示生成网络，D表示判别网络，E表示数学期望，x，z分别表示大物体和小物体；整个模型是在训练G的过程中，最大化D出错的概率；另外，生成网络G将小物体的标识转换为超分辨的形式，使之类似于大目标的特征，由于信息的缺乏，这个过程比较困难。同时引入附加信息

$$\min_G \max_D L(D,G) = E_{F_l\sim p_{\text{data}}(F_l)} \log D(F_l) + E_{F_s\sim p_{F_s}(F_s|f)} [\log(1 - D(\underbrace{F_s + G(F_s|f)}_{\text{residual learning}}))] \tag{4}$$

来学习大物体和小物体间的残差表示。其中，F_s表示小物体的特征，F_l表示大物体的特征，G表示生成器函数，D表示分类器。最终，该算法在PASCAL VOC 2007数据集上的检测精度从Faster R-CNN的73.2%提高到84%。

3 多类别物体检测

图像中经常存在多个不同类别的物体，而实现对多类别物体的检测也面临两个问题：1)目标检测数据集类别数目较少，模型训练效果较差；2)随着物体类别的增加，计算量随之增大，导致检测速度和检测精度降低。针对这些问题，研究者从训练方式和网络的角度提出了多种优化多类别物体检测效果的算法。

3.1 训练方式的优化

经典算法在包含20个类别的对象目标检测场景中具有较高的检测准确度，但是现实场景中，对象种类成千上万，基于大尺寸场景的经典检测算法表现并不突出，针对这个问题，Hoffman等人(2014)提出了一个基于迁移学习训练方式的自适应多类别检测算法LSDA(large scale detection through adaptation)，针对包含上千类目标的多分类数据集，可以训练出很好的分类网络，而检测数据集只有几十类的目标，检测这几十类目标准确率很高，但是应用于其他类别的对象则表现不佳，因此，LSDA首先利用分类数据集初始化神经网络，然后再利用检测数据集进行微调，最后将检测数据集训练的参数迁移到没有检测标签的类别上来实现提高目标检测类别数目。

在YOLOv2(Redmon等，2016)的基础上，Redmon和Farhadi(2017)针对上述问题提出了YOLO9000的联合训练策略，即将检测和分类数据集联合，共同训练目标检测模型，实现了多类别物体检测。联合训练策略通过使用标记为检测的图像来学习边界框的坐标预测和目标类别的特定信息，将检测和分类数据集混合用于模型训练。当网络看到标记为检测的图像时，能够基于完整的YOLOv2损失函数进行反向传播；当看到一个分类的图像时，只能够从该架构的分类特定部分反向传播损失。由于检测数据集只有通用的目标和标签，分类数据集具有更广、更深的标签范围，为了使用一种连贯的方式来合并这些标签，Redmon和Farhadi(2018)提出了层次分类(hierarchical classification)方法，借鉴ImageNet中的词典(WordNet)思想来构建分层树，提出词树(WordTree)来简化组合假定不互斥的数据集。WordTree是一个视觉概念的分层模型，为了使用WordTree进行分类，Redmon和Farhadi等人(2018)预测每个节点的条件概率，以此得到同义词集合中每个同义词的下义词概率。例如，当样本标签为叶子节点时，父节点也激活成为正样本；当样本标签为非叶子节点时，将对非叶子节点和父节点进行反向传播，如图13和图14所示。

在目标检测过程中，未标记的对象实例将被视为背景，给检测器的生成带来不正确的训练信号，但

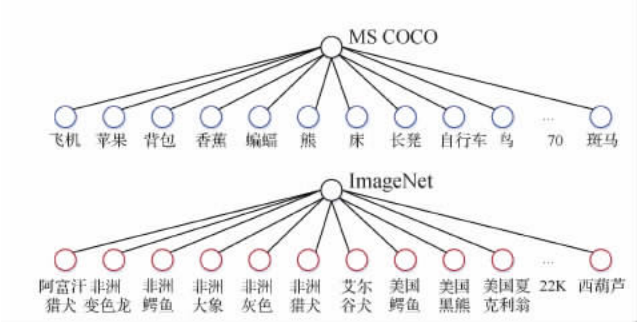


图 13　MS COCO 和 ImageNet 数据存储结构示意图

Fig. 13　Data storage structure of MS COCO and ImageNet

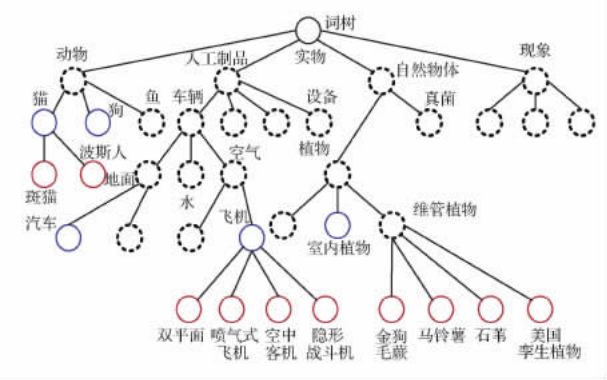


图 14　WordTree 数据存储结构示意图

Fig. 14　Data storage structure of WordTree

是研究人员通过实验分析发现删除 30% 的正确样本标签之后,Faster R-CNN 的性能只下降了 5%,因此,Wu 等人(2018)提出了一种简单有效的物体检测方法,即软采样(soft sampling)。软采样方法将 ROI 的梯度重设为与正实例重叠的函数,确保不确定的背景区域相对于难样本(hard negative)的权重更小。在 PASCAL VOC 数据集上的大量实验证明了软采样方法在不同标注率下的有效性,在 Open-ImagesV3(Krasin 等,2017)数据集上的实验也表明在缺少注释的真实数据集上,软采样优于标准检测基线方法,因此可以利用软采样的训练思想实现多类别物体检测。

3.2　网络结构的改进

R-FCN 算法(Dai 等,2016)通过引入位置敏感得分图(position-sensitive score map)解决了 Faster R-CNN 中 ROI 的重复计算问题,提升了检测速度。位置敏感得分图由 R-FCN 网络中的一个卷积层获得,该卷积层中卷积核的数量与位置敏感得分图的输出维度相同,都是 $p^2(C+1)$,其中,C 是物体类别数,p 是 ROI 输出尺寸,因此当检测类别较多时会急剧增加位置敏感得分图的计算量,从而导致算法的检测速度下降。针对这个问题,Singh 等人(2018)提出了 R-FCN-3000 算法用于满足多类别物体检测的需求。R-FCN-3000 将分类网络分成两条支路,一条支路的整体思想与 R-FCN 中的位置敏感得分图类似,用来进行大类别(super-class)检测,得到物体对应的大类别分数。大类别指物体类型相近的一些类,大类的数量是固定的,可以通过聚类(K 均值聚类)获得,该支路的计算量为 $p^2(K+1)$,其中,K 为物体大类数目,且 K 的数量远远小于物体类别数目,因此可以降低位置敏感得分图的计算量;另一条支路用来进行细粒度分类,由一个包含 C 个卷积核的卷积层组成,C 为所有物体的类别数,该部分大约只有 R-FCN 中 $p^2(C+1)$($p=7$ 时)的计算量的 1/49。最后两个分支的类别概率分数相乘就是目标的最终类别分数。此种做法在多类别物体检测方面,相对于 R-FCN 模型来说,可以减少很多计算量。解耦之后,即使物体类别 C 从 20 类增加为 3 000 类,所带来的计算量增加并不大,因此可以在保证检测类别的前提下维持检测速度。

R-FCN-3000 在解耦分类支路的基础上,提出了一种基于 R-FCN 的多类别目标检测架构,为大规模目标的检测提供了一种更加准确、高效的解决方案,但是 R-FCN-3000 是在具有边界框注释的强监督数据集上训练的,很难直接推广到具有图像级类别标签的类别,因此,Guo 等人(2019)提出了一种用于大规模半监督目标检测的层次结构和联合训练策略(hierarchical structure and joint training, HSJT)。HSJT 首先利用目标类别之间的关系,建立新的层次网络模型,进一步提高识别性能;其次将边界框级标记图像和图像级标记图像结合起来进行联合训练,提出一种联合训练生成大规模半监督目标检测的方法。实验结果表明,HSJT 提出的大规模半监督目标检测网络具有很好的多类别检测性能,在 ImageNet 数据集上的检测精度达到了 38.1%。

4　轻量化模型

深度网络模型中,上百层的网络结构会包含大量的参数,需要极大的存储空间和运行空间来完成检测任务。为了使基于深度学习的目标检测算法能

够从服务器迁移到移动端，迫切需要将基于深度学习的目标检测模型进行轻量化处理。

谷歌的 MobileNetv1(Howard 等，2017)从实际应用场景出发将标准目标检测网络结构的卷积分成两个部分：1)深度可分离卷积(depth-wise convolution)。深度可分离卷积与分组卷积(group convolution)类似，不同之处是分组卷积中一组卷积核负责对一组特征图进行卷积，而深度可分离卷积中一个卷积核只负责对一个特征图进行卷积，这样可以减少网络权值参数和模型计算量，轻量化检测模型；2)逐点卷积(pointwise convolution)。逐点卷积利用 1×1 卷积核将深度可分离卷积中得到的特征图进行融合，用来解决特征图之间信息不流畅的问题。

MobileNetv1 与一般的卷积方式存在差异，如图 15 所示，其中输入特征图有 M 个，输出特征图有 N 个，输入特征图尺寸为 D_F。一般卷积(standard convolution)中采用 N 个大小为 $D_K\times D_K$ 的卷积核进行卷积操作，如图 15(a)所示，卷积的计算量为

$$S_c = D_K \times D_K \times M \times N \times D_F \times D_F \tag{5}$$

需要的参数数目为

$$S_p = D_K \times D_K \times M \times N \tag{6}$$

深度可分离卷积中一个卷积核只负责一个通道，则需要有 M 个 $D_K\times D_K$ 的卷积核，具体信息如图 15(b)所示。逐点卷积为了达到输出 N 个特征图的操作，采用 N 个 1×1 的卷积核进行卷积，如图 15(c)所示，最终完成一次卷积的计算量为

$$S_C = (D_K \times D_K \times M + M \times N) \times D_F \times D_F \tag{7}$$

参数数目为

$$S_P = D_K \times D_K \times M + M \times N \tag{8}$$

由此可见，在 D_K，N 较大的情况下，MobileNetv1 中的深度可分离卷积和逐点卷积无论是参数数目还是计算速度上都有很大优势。

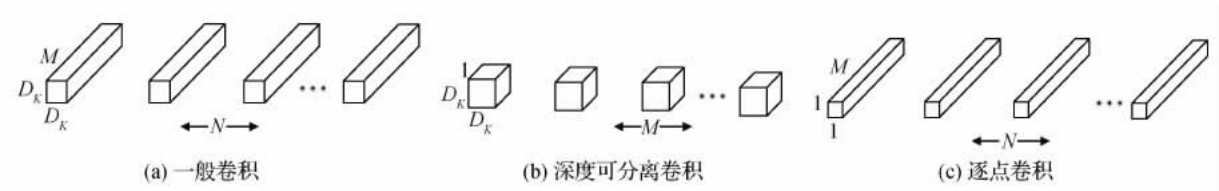


图 15 MobileNetv1 与一般卷积的区别

Fig. 15 The difference between MobileNetv1 and standard revolution

((a) standard convolution; (b) depth-wise convolution; (c) pointwise convolution)

MobileNetv1 利用深度可分离卷积减少了模型的运算量和参数量，但整体网络结构还是延续了 VGG 网络直上直下的特点。Sandler 等人(2019)在 MobileNetv1 的基础上提出了 MobileNetv2，主要包含两点改进：一是使用线性瓶颈(linear bottlenecks)去除小维度输出层后面的非线性激活层，从而保证模型的表达能力；二是利用倒置残差模块(inverted residual block)，该结构与传统残差模块(residual block)中维度先缩减再扩增的方式相反。实验结果表明 MobileNetv2 相对于 MobileNetv1 的两点改进极大地提升了模型效果，算法在 ImageNet 数据集上的准确率从 70.6% 提高到了 74.7%。

Face++团队提出的 ShuffleNetv(Han 等，2017)利用通道交换(channel shuffle)方法将各部分特征图的顺序有序地打乱，构成新的特征图以解决分组卷积(group convolution)带来的信息流通不畅和 MobileNetv1 中的逐点卷积带来的大量权值参数问题，最终达到轻量化检测模型的效果。其中通道交换是将各组的通道平均分为 g($g=3$)份，然后按照一定次序重新构成特征图，具体信息如图 16 所示。

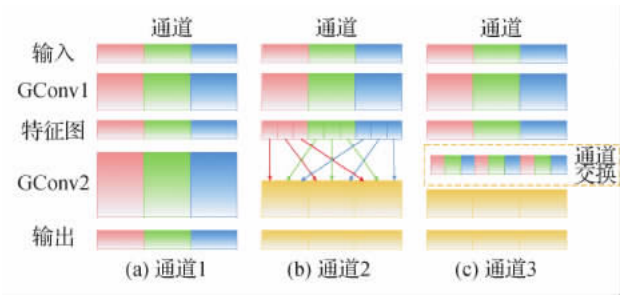


图 16 群卷积中的通道交换方法

Fig. 16 Channel shuffle with two stacked group convolutions

((a) channel 1; (b) channel 2; (c) channel 3)

大部分的模型加速和压缩文章在对比加速效果

时使用的指标都是每秒浮点运算次数(floating point operations per second,FLOPs),这个指标主要衡量卷积层的乘法操作,而内存访问消耗时间(memory access cost,MAC)的计算可能会导致模型即使在相同的 FLOPs 下,实际速度差别也会比较大。Ma 等人(2018)通过 4 个实验分析了影响 MAC 的 4 个主要因素。

第 1 个实验是关于卷积层的输入输出特征通道数对 MAC 指标的影响。假设一个 1×1 卷积层的输入特征通道数是 c_1,输出特征尺寸是 h 和 w,输出特征通道数是 c_2,用 B 表示卷积层的 FLOPs 值,则

$$B = hwc_1c_2 \tag{9}$$

因为是 1×1 卷积,所以输入特征和输出特征的尺寸是相同的,hwc_1 表示输入特征所需存储空间,hwc_2 表示输出特征所需存储空间,c_1c_2 表示卷积核所需存储空间,模型的 MAC 为

$$M_{AC} = hw(c_1 + c_2) + c_1c_2 \tag{10}$$

由均值不等式可得

$$M_{AC} \geqslant 2\sqrt{hwB} + \frac{B}{hw} \tag{11}$$

因此当 $c_1 = c_2$,即输入特征通道数和输出特征通道数相等时,在给定 FLOPs 前提下,MAC 达到取值的下界。

第 2 个实验是关于卷积的 group 操作对 MAC 的影响。同理可得

$$M_{AC} = hw(c_1 + c_2) + \frac{c_1c_2}{g} = hwc_1 + \frac{Bg}{c_1} + \frac{B}{hw} \tag{12}$$

第 3 个实验是关于模型设计的分支数量对模型速度的影响,实验结果显示在相同 FLOPs 的情况下,单卷积层的速度最快,因此模型支路越多对于并行计算越不利。

第 4 个实验是关于元素判断(element-wise)操作对模型速度的影响。通过对 ShuffleNetv1 和 MobileNetv2的几种层操作的时间消耗的分析,发现 element-wise 操作带来的时间消耗远比在 FLOPs 上体现的数值要多,因此要尽可能减少 element-wise 操作。

ShuffleNetv2(Ma 等,2018)是基于 ShuffleNetv1 中存在的问题和以上 4 个实验提出的优化网络结构,结构对比如图 17 所示。图 17(a)(b)是 ShuffleNetv1 的两种不同块结构,其区别是后者对特征图尺寸进行了缩小操作。图 17(c)(d)是 ShuffleNetv2 的两种不同结构。从图 17(a)和 17(c)的对比可以看出,图 17(c)在开始处增加了一个通道拆分(channel split)操作,将输入特征的通道拆分成 c 和 c',c' 采用 $c/2$,此处改进与第 1 个实验的分析相对应。基于第 2 个实验的分析,图 17(c)相对于图 17(a)取消了 1×1 卷积层中的 group 操作,同时前面的通道拆分可以看做变相的 group 操作。基于第 3 个实验的分析,图 17(c)将通道交换(channel shuffle)的操作移到了 concat 操作后面。基于第 4 个实验的分析,图 17(c)将图 17(a)中的 add(element-wise add)操作替换成 concat 操作。最终 ShuffleNetv2 的检测精度相比于原有的 ShuffleNetv1 从 29.9%提升到 31.8%,提升了 1.9%。

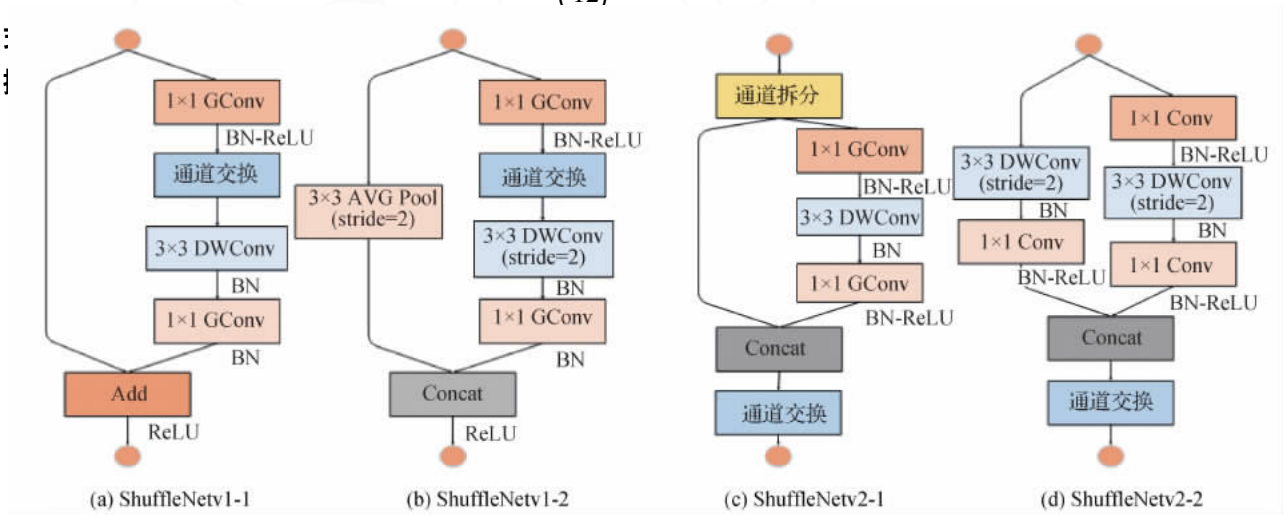


图 17　ShuffleNetv1 和 ShuffleNetv2 的构成模块

Fig. 17　Building blocks of ShuffleNetv1 and ShuffleNetv2

((a) ShuffleNetv1-1; (b) ShuffleNetv1-2; (c) ShuffleNetv2-1; (d) ShuffleNetv2-2)

谷歌的研究人员在结合了 MobileNetv1 的深度可分离卷积、MobileNetv2 的具有线性瓶颈的逆残差结构和 MnasNet(Tan 等,2019)的基于压缩和激活结构的轻量级注意力模型优点的基础上,提出了一个新的轻量级网络 MobileNetv3(Howard 等,2019),取得了更好的实验效果。该模型包含两大创新点:互补搜索技术组合和网络结构改进。

在互补搜索技术组合方面,MobileNetv3 首先使用神经网络搜索功能构建全局的网络结构,随后利用 NetAdapt 算法(Yang 等,2018)对每层的核数量进行优化,最终在计算和参数量受限的前提下自动搜索到最优的网络架构。对于全局的网络结构搜索,MobileNetv3 使用与 MnasNet 中相同的基于 RNN 的控制器和分级的搜索空间技术,并针对特定的硬件平台进行精度—延时平衡优化,在目标延时范围内进行搜索。而对于每层核数量的优化,MobileNetv3 利用 NetAdapt 方法对每一层按照序列的方式进行调优,在尽量优化模型延时的同时,保持精度,减小扩充层和每一层中瓶颈的大小。

在网络结构改进方面,针对 MobileNetv2 网络最后阶段计算量较大的问题,重新设计了 MobileNetv2 的网络结构,将最后一步的平均池化层前移,并移除最后一个卷积层,同时引入 h-swish(hard version of swish)激活函数来降低网络的计算量,提高模型检测速度。由于 swish 激活函数能够有效提高网络的精度,但是带来的计算量太大,因此提出了 h-swish(hard version of swish)激活函数,具体为

$$h_{\text{swish}}[x] = x\frac{R_{\text{eLU}}6(x+3)}{6} \tag{13}$$

式中,x 代表对应的输入,$R_{\text{eLU}}6$ 代表 ReLU 激活函数。这种非线性激活函数在保持精度的情况下带来了很多优势,不仅 $R_{\text{eLU}}6$ 在众多软硬件框架中都可以实现,而且量化时避免了数值精度的损失,运行速度快。

5 数据集和各算法性能比较

5.1 数据集

数据、计算力和算法是人工智能发展的 3 大要素,其中数据是人工智能发展的基础,任何研究都离不开数据的支撑。深度学习算法之所以可以得到广泛发展,很大一方面原因得益于大规模数据集的出现,目标检测算法能发展到现在,同样也离不开数据集的重要贡献。数据集不仅是衡量和比较目标检测算法性能的常用依据,同时也是将目标检测推向越来越复杂和越来越具有实用性方向的强大动力。

目标检测的相关研究中使用过很多的数据集,包括通用数据集和专用数据集。本文仅对相关通用数据集进行介绍。从 Caltech101(Lin 等,2014)开始,目标检测中使用的通用数据集主要包括 Caltech256(Griffin 等,2017)、PASCAL VOC 2007、PASCAL VOC 2012、Tiny Images(Torralba 等,2008)、Cifar10(Williams 和 Li,2018)、Sun(Xiao 等,2016)、ImageNet、Places(Zhou 等,2018a)、MS COCO 和 Open Images。现阶段目标检测算法中使用的主流数据集主要包括 PASCAL VOC 2007、PASCAL VOC 2012、MS COCO 和 ImageNet 等,相关数据集的详细信息见表 1。

5.2 各算法性能比较

目标检测中的性能指标主要包括准确率(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)。准确率表示所有样本中分类对的样本所占的比例,是正确预测类别的样本个数与样本总数的比值。准确率一般用来评估模型的全局准确程度,不能包含太多信息,无法全面评价一个模型性能。精确率指识别出的正样本中,正确识别的正样本个数所占的比例。召回率指测试集中所有正样本样例中被正确识别的正样本个数所占的比例。平均精度指 P-R 曲线下的面积,P-R 曲线显示的是分类器上精确率和召回率之间的权衡,P-R 曲线上的点为某一阈值下该模型的准确率和召回率,P-R 曲线是通过将阈值从高到低移动生成的。平均精度均值(mAP)是先对每一个类别都计算出平均精度(AP),然后再计算 AP 的平均值。目标检测算法中主要使用的性能指标是 mAP,本文所有的检测精度都是指 mAP。

上文对相关挑战中的很多算法以及相应的数据集和目标检测中的性能指标进行了详细介绍。下面将从骨干网络、输入图像尺寸、测试集、检测精度(mAP)、检测速度和单双阶段划分的角度整体介绍传统目标检测算法、主流目标检测算法的改进与优化算法、提高小目标物体检测精度算法和多类别目标检测算法的相关信息,具体内容见表 2。

表1 目标检测的常用数据集

Table 1 Popular databases for object detection

名 称	图像总数	类别数目	图像尺寸/像素	起始时间/年	特 点
Caltech101(Lin 等,2014)	9 145	101	300×200	2004	训练图像相对较少; 每幅图像只有一个目标; 大多数图像中没有噪声; 有限的类内变化; 不太适用于实际评估。
PASCAL VOC 2007(He 等,2015)	9 963	20	375×500	2005	仅涵盖日常生活中常见的 20 个类别; 大量训练图像; 图像接近真实世界; 较大的类内变化; 目标在相应的场景中; 一幅图像中包含多个目标; 包含许多不同的样本; 创建标准化的先例。
PASCAL VOC 2012(He 等,2015)	11 540	20	470×380	2005	与 PASCAL VOC 2007 类似,是 PASCAL VOC 2007 数据集的升级版,包含图像数量更多。
TinyImages(Torralba 等,2008)	约 79 000 000	53 464	32×32	2006	图像数量、类别数量最多; 低分辨率图像; 未手动验证; 不太适合算法评估; 子集 CIFAR10 和 CIFAR100 常用于目标分类。
Caltech256(Griffin 等,2017)	30 607	256	300×200	2007	与 Caltech101 类似,目标类的数量比 Caltech101 多。
ImageNet(Nielsen,2018)	约 14 000 000	21 841	500×400	2009	图像和目标类别的数量较多; 比 PASCAL VOC 更具挑战性; 子集 ImageNet1000较流行; ILSVRC 挑战的基础。
SUN(Xiao 等,2016)	131 072	908	500×300	2010	大量场景类别; 每个对象类别的实例数显示长尾现象; 场景识别和目标检测的两个基准分别是 SUN397 和 SUN2012。
MS COCO(Havard 等,2017)	约 328 000	91	640×480	2014	更接近真实场景; 每幅图像包含更多目标实例和更丰富的目标注释信息; 用于大规模目标检测的下一个主要数据集。
Places(Zhou 等,2017)	约 10 000 000	434	256×256	2014	用于场景识别的最大标记数据集; 常用的 4 个子集是 Places365 Standard、Places365 Challenge、Places205 和 Places88。
Open Images(Krasin 等,2017)	约 9 000 000	约 6 000	变化	2017	约 900 万幅图像组成的数据集,图像已经用图像级标签和目标边界框进行了注释。

从表 2 可以得出以下结论:

1) 深度学习的利用使目标检测算法实现了质的提升。通过使用深度学习,在 PASCAL VOC 2007 数据集上的检测精度从以 DPM 为代表的传统目标检测算法的 29.2%,提高到以 R-CNN 为起点的基于深度学习的目标检测算法的 66%,提升了 36.8%。

2) 在主流目标检测算法的相关改进算法中,双阶段算法的检测精度较高,在 PASCAL VOC 2007 数据集上可以达到 80.5%,算法的改进可以使模型的检测速度从 0.5 帧/s 提升到 9 帧/s; 单阶段算法的检测速度较高,可以保持在 40 帧/s 以上,算法的改进使得单阶段目标检测算法的检测精度从 66.4% 提升到 78.6%; RefineDet 算法结合双阶段算法的精度优势以及单阶段算法的速度优势,使算法在检测速度接近单阶段算法条件下检测精度达到 81.8%。因此,需要结合单、双阶段算法的特点进行改进,使目标检测的精度和速度达到更好的平衡。

3) 在提高小目标物体检测精度的相关算法中,发现新的高性能骨干网络。如 ResNet-101、DenseNet-169、DetNet-59 可以将目标检测算法在 PASCAL VOC 2007 数据集上的检测精度提升到一个新的层次,分别达到 80.0%、80.9%、40.2%; Tri-

表 2　不同目标检测算法比较

Table 2　Comparisons of different object detection approaches

算法	骨干网络	输入图像尺寸/像素	测试集	mAP/%	检测速度/(帧/s)	阶段划分
DPM(Juan 和 Gwun 等,2013)	–	–	VOC 2007	29.2	–	–
R-CNN(Girshick 等,2014)	VGG16	1 000 × 600	VOC 2007	66	0.5	双
Fast R-CNN(Girshick 等,2015)	VGG16	1 000 × 600	VOC 2007	70	7	双
Faster R-CNN(Ren 等,2018)	VGG16	1 000 × 600	VOC 2007	73.2	7	双
Faster R-CNN(Ren 等,2018)	ResNet101	1 000 × 600	VOC 2007	76.4	5	双
HyperNet(Kong 等,2016)	VGG-16	1 000 × 600	VOC 2007	76.3	0.88	双
R-FCN(Dai 等,2016)	ResNet-101	1 000 × 600	VOC 2007	80.5	9	双
MaskR-CNN(He 等,2017)	ResNet-101	1 000 × 600	MS COCO	33.1	4.8	双
YOLOv1(Redmon 等,2016)	VGG16	448 × 448	VOC 2007	66.4	45	单
SSD(Liu 等,2016)	VGG16	300 × 300	VOC 2007	77.1	46	单
YOLOv2(Redmon 和 Farhadi,2017)	Darknet-19	544 × 544	VOC 2007	78.6	40	单
YOLOv3(Redmon 和 Farhadi,2018)	Darknet-53	608 × 608	MS COCO	33	51	单
RON(Kong 等,2017)	VGG-16	384 × 384	VOC 2007	75.4	15	单
RefineDet320(Zhang 等,2018c)	VGG-16	320 × 320	VOC 2007/2012	80.0/78.1	40.3	双
RefineDet512(Zhang 等,2018b)	VGG-16	512 × 512	VOC 2007/2012	81.8/80.1	24.1	双
DSSD321(Fu 等,2017)	ResNet-101	321 × 321	VOC 2007/2012	78.6/76.3	9.5	单
DSSD513(Fu 等,2017)	ResNet-101	513 × 513	VOC 2007/2012	81.5/80.0	5.5	单
STDN321(Zhou 等,2018b)	DenseNet-169	321 × 321	VOC 2007	79.3	40.1	单
STDN513(Zhou 等,2018b)	DenseNet-169	513 × 513	VOC 2007	80.9	28.6	单
DetNet(Li 等,2018)	DetNet-59	321 × 321	MS COCO	40.2	–	–
RFB Net300(Liu 等,2018b)	VGG-16	300 × 300	VOC 2007	80.5	83	单
RFB Net512(Liu 等,2018b)	VGG-16	512 × 512	VOC 2007	82.2	38	单
TridentNet(Li 等,2019)	ResNet-101	321 × 321	MS COCO	48.4	–	单
FPN(Seferbekov 等,2018b)	ResNet-50	1 000 × 600	MS COCO	35.8	5.8	双
DES300(Zhang 等,2018b)	VGG-16	300 × 300	VOC 2007/2012	79.7/77.1	76.8	单
DES512(Zhang 等,2018b)	VGG-16	512 × 512	VOC 2007/2012	81.7/80.3	31.7	单
NAS-FPN(Ghiasi 等,2019)	ResNet-50	1 024 × 1 024	MS COCO	44.2	92.1	双
Cascade R-CNN(Cai 和 Vasconcelos,2018)	ResNet-101	1 280 × 800	MS COCO	42.8	–	双
D-RFCN + SNIP(Singh 等,2018)	DPN-98	1 000 × 600	MS COCO	48.3	2	双
YOLO9000(Redmon 和 Farhadi,2017)	Darknet-19	544 × 544	VOC 2007	78.6	40	单
Soft Sampling(Wu 等,2018)	VGG-16	321 × 321	VOC 2007	79.3	–	双
R-FCN-3000(Singh 等,2018)	ResNet-101	1 000 × 600	VOC 2007	80.5	30	双
HSJT(Guo 等,2019)	ResNet-101	1 000 × 600	ImageNet	38.1	–	双

注　“–”表示未使用深度学习方法,测试集中 VOC 指 PASCAL VOC。

dentNet、NAS-FPN、Cascade R-CNN 等算法分别从增加视觉感受野、特征融合和级联卷积神经网络的角度,将在 MS COCO 数据集上的目标检测精度提高到 48.4%、44.2% 和 42.8%,有效提升了小目标物体的检测精度。从不同算法在 MS COCO、PASCAL VOC 测试集以及相同算法在 PASCAL VOC 2007、PASCAL VOC 2012 等不同检测集上的检测结果可以发现,检测类别数目越多,模型的检测精度越低。而 SNIP 算法通过修改模型的训练方式,将在 MS COCO 测试集上的目标检测精度提高到 48.3%。因此可以通过不断尝试从高性能骨干网络、增加视觉感受野、特征融合、级联卷积神经网络和修改模型的训练方式等角度来提升小目标物体检测精度。

4) 在实现多类别物体检测的相关算法对比中,发现 YOLO9000 和 R-FCN-3000 通过使用改变数据集的存储方式和网络分支的方法,不但实现了 9 000 和 3 000 个类别的物体检测,而且将在 PASCAL VOC 2007 数据集上的检测精度提高到 78.6% 和 80.5%,达到同等条件下少类别物体检测的精度要求,HSJT 算法更是将大规模数据集 ImageNet 上的检测精度提升到了 38.1%,极大提高了小目标物体的检效果。

综上,基于深度学习的目标检测算法正在从不同角度解决相关技术挑战,在 PASCAL VOC、MS COCO 和 ImageNet 数据集上的检测精度分别达到了 82.2%、48.3% 和 38.1%,检测速度也基本满足了实时性要求,取得了较好的实验效果。同时,本文利用轻量化检测模型的主流性能指标,如 FLOPs、检测精度和 GPU 速度对相关轻量化检测算法进行了总结和梳理,不同算法的详细信息如表 3 所示。

表 3 在 MS COCO 数据集上的目标检测结果

Table 3 Results of object detection on MS COCO dataset

模型	FLOPs/M							
	mAP/%				GPU 速度/(帧/s)			
	40	140	300	500	40	140	300	500
Xception(Chollet,2017)	21.9	29	31.3	32.9	178	131	101	83
ShuffleNetv1(Zhang 等,2018a)	20.9	27	29.9	32.9	152	85	76	60
MobileNetv2(Sandler 等,2019)	20.7	24.4	30	30.6	146	111	94	72
ShuffleNetv2(Ma 等, 2018)	22.5	29	31.8	33.3	188	146	109	87

从表 3 可以发现,随着算法网络结构的不断优化,轻量化检测模型的 FLOPs 已经可以减少到 500 M、300 M、140 M 和 40 M,检测图像的速度也在不断增加,在 ShufleNetv2 中已经可以达到 188 帧/s。同时,模型的检测精度也在不断提高,在 MS COCO 测试集上检测精度已经可以达到 33.3%。但是,无论检测精度、检测速度还是模型体量都距离移动端的真实需求有很大差距,因此目标检测算法应用到移动端需求中还需要不断地研究与发展。

6 待解决的问题与未来研究方向

目标检测作为计算机视觉领域中一个重要且具有挑战性的问题,受到了广泛关注。随着深度学习技术的显著发展,目标检测领域已经发生了巨大的变化,但是目前的检测效果与人性化的表现之间仍存在巨大差距。根据已有的研究方法和最新的研究思路,本文对基于深度学习的目标检测关键技术下一步待解决的问题与未来研究方向进行展望。

1) 如何使模型更适应特定场景下的目标检测需求。由于真实世界的图像存在着巨大变化,例如遮挡(Yan 等,2019; Sun 等,2019b)、图像模糊(Gao 等,2019; Zhang 等,2019a)、视点和光照变化、物体尺度、物体姿态(Kocabas 等,2019)、物体部分变形、噪音和扭曲等,这些特定的场景和因素为目标检测算法的实际应用提出新的挑战性,因此,需要通过上下文信息、选择性参数共享、数据增强(Zhang 等,2019b)、互补特征融合(Zhao 等,2019)等各种方法来不断提高模型的性能以适应特定场景下的目标检测需求。

2) 如何在先验知识缺失的条件下实现精确的目标检测。由于目标检测的最终任务是利用计算机从一些复杂场景下,快速准确地识别出数千种甚至更多类别中的目标对象或实例,因此,先验知识的丰富程度以及质量的好坏,将直接影响深度网络模型训练的质量。目前的主流方法是采用人工方式进行大规模的语料标注,例如,李飞飞等人建立的 ImageNet 数据集以及 PASCAL VOC、MS COCO 等通用的大规模标记数据集。但是这种有监督的学习方法,一方面人工标注的成本太大,同时也无法对所有的场景进行标注; 另一方面,标注好的数据对不同环境或场景下的新类别的目标无法实现自适应的精确识别。因此,近年来提出了迁移学习或强化学习等基于弱监督(Ren 等,2018; Karlinsky 等,2018)或无监督(Rahman 等,2018; Demirel 等,2018)的方法来训练目标检测模型,但如何结合先验知识或常识来实现不同

场景下的自适应目标识别已成为目前研究的核心。

3) 如何获取高性能的骨干网络。底层骨干网络的性能好坏直接影响到目标检测算法的性能质量,因此如何获取高性能的骨干网络将对后续的研究与实验结果产生重要影响。目前骨干网络的选择具有很大的随机性,特别是针对不同任务,骨干网络框架体现的性能也存在显著差异,因此目标检测领域中骨干网络的未来核心工作将围绕以下两点展开:(1) 提高骨干网络针对不同任务的专一性。虽然现有的骨干网络已经针对图像分类进行了优化,但是由于分类和检测之间的差异性导致学习过程产生偏差,因此迫切需要研究针对目标检测领域的专用骨干网络(Wang 等,2019b);(2) 提高骨干网络的处理效率。由于现有的骨干网络中参数的规模常常达到数百万,甚至数亿,需要消耗大量硬件资源进行训练,因此如何对网络进行压缩和加速(Cheng 等,2018; Wei 等,2018) 以满足网络处理效率的需求将是未来的核心研究方向。

4) 如何获得更加丰富的图像语义信息。在目标检测过程中,图像语义信息越丰富,模型学习到的特征就会越多,检测效果就越好。但是,如何能够充分挖掘现有网络结构中图像的语义信息是一个关键性问题,目前的解决方案大致分为以下两类:(1) 利用高清表示,即针对深度神经网络的深层次特征提取导致特征图的尺寸缩减、分辨率降低、模型的检测精度降低等问题,可以通过利用沙漏型对称网络结构、反卷积、空洞卷积和并行多分辨率分支等方法获得具有高清表示信息的特征图来提升特征图空间分辨率,从而提高目标检测的精度;(2) 利用图像语义理解,图像的语义理解可以帮助模型更加高效和深层次地学习图像信息,特别是解决像素级对象实例分割问题,即通过像素级对象实例的精确分割,获取更加细粒度的图像实例特征,从而实现对图像语义的更深层次的理解。

5) 如何提高深度学习模型的可解释性。对于大多数的深度学习网络构架,神经元规模很容易达到几万的量级,参数的规模也大到百万乃至上亿,虽然研究者通过各种方式得到了准确度较高的算法,但它们实际上是一堆看上去毫无意义的模型参数和拟合度非常高的判定结果,而模型本身意味着知识。研究者希望知道模型究竟从数据中学到了哪些知识,从而做出最终决策,因此需要不断提高模型的可解释性。未来的核心研究方向将是利用胶囊网络(Verma 和 Zhang,2018; Rosario 等,2019) 等新结构来提高深度学习模型的可解释性。卷积神经网络中的池化层可以用于缩小网络规模和计算需求,但同时也会破坏信息特征,使得只能从网络结构中学习到图像的统计特征,而无法学习到更高级的语义信息。针对卷积神经网络存在的这种缺陷,类似于胶囊网络的这种新型结构可以学习图像中特定实体的各种特征,从而使得模型可以在不同的视角下使用更少的数据实现更好的泛化。

6) 如何自动化生成或设计最优的网络架构。基于多层神经网络的深度学习是一门计算密集型技术,设计一个好的网络结构不仅依赖于较高的领域知识与硬件资源,同时也依赖于不同设计者的个人经验和灵感。如何利用机器学习的方法来自动化地学习和训练出一个最优的网络架构是目前最新的研究热点方向之一。Zoph 和 Le(2019) 开创了通过神经网络架构搜索的方式来实现深度学习的自动化这一全新的领域,证明了利用强化学习算法可以发现更好的网络架构; Real 等人(2019) 进一步研究证明了利用神经进化算法也可以得到类似的结果,但是这两种搜索方法都要使用 GPU 训练几千小时,因此如何优化并降低这种计算负担成为关键。Zhao 等人(2019) 提出从一个简单的网络架构开始搜索,通过功能保留的操作逐步增加搜索的宽度和深度。Pham 等人(2018) 提出构造包含搜索空间中所有架构的过参数化架构,然后对这个大型架构中的小部分内容进行采样和训练,训练完成后,抽样得到的架构可以共享训练权重,从而大大降低神经网络架构搜索的工作量。目前,神经架构搜索已经在图像识别和检测(Ghiasi 等,2019) 中展现出很强的能力,不论是可微架构搜索的速度,还是基于强化学习搜索的准确度,自动架构搜索的性能已经超越了手动设计的版本,因此,如何利用神经网络架构自动化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的关键。

7 结 语

本文根据目标检测领域中存在的核心问题与关键的技术挑战,从基于主流目标检测算法的性能改进与优化、提高小目标检测精度、多类别物体检测和

轻量化检测模型 4 个方向对基于深度学习的目标检测算法进行综述,并对比分析了通用数据集以及相关算法在主流数据集上的实验结果,同时对该领域中如何使模型更适应特定场景下的检测需求、如何在先验知识缺失的条件下实现精确的目标检测问题、如何获取高性能的骨干网络、如何获得更加丰富的图像语义信息、如何提高深度学习模型的可解释性、如何自动化实现网络最优架构等 6 个研究热点方向进行了分析与展望。

参考文献(References)

Bodla N, Singh B, Chellappa R and Davis L S. 2017. Soft-NMS — improving object detection with one line of code//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5562-5570 [DOI: 10.1109/ICCV.2017.593]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE [DOI: 10.1109/CVPR.2018.00644]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Cheng Y, Wang D, Zhou P and Zhang T. 2018. Model compression and acceleration for deep neural networks: the principles, progress, and challenges. IEEE Signal Processing Magazine, 35(1): 126-136 [DOI: 10.1109/MSP.2017.2765695]

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Dai J F, Li Y, He K M and Sun J. 2016. R-FCN: object detection via region-based fully convolutional networks [EB/OL]. (2016-05-20) [2019-06-20]. https://arxiv.org/pdf/1605.06409.pdf

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]

Demirel B, Cinbis R G and Ikizler-Cinbis N. 2018. Zero-shot object detection by hybrid region embedding [EB/OL]. (2018-05-16) [2019-06-20]. https://arxiv.org/pdf/1805.06157.pdf

Divvala S K, Efros A A and Hebert M. 2012. How important are "deformable parts" in the deformable parts model? //Fusiello A, Murino V and Cucchiara R, eds. Computer Vision – ECCV 2012. Workshops and Demonstrations, . Berlin, Heidelberg: Springer: 31-40 [DOI: 10.1007/978-3-642-33885-4_4]

Everingham M, Eslami S M A, van Gool L, Williams C K I, Winn J and Zisserman A. 2015. The PASCAL visual object classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136 [DOI: 10.1007/s11263-014-0733-5]

Fu C Y, Liu W, Ranga A, Tyagi A and Berg A C. 2017. DSSD: deconvolutional single shot detector [EB/OL]. (2017-01-23) [2019-06-20]. https://arxiv.org/pdf/1701.06659.pdf

Gao H Y, Tao X, Shen X Y and Jia J Y. 2019. Dynamic scene deblurring with parameter selective sharing and nested skip connections [EB-OL]. [2019-06-20]. http://jiaya.me/papers/deblur_cvpr19.pdf

Ghiasi G, Lin T Y, Pang R M and Le Q V. 2019. NAS-FPN: learning scalable feature pyramid architecture for object detection [EB/OL]. (2019-04-16) [2019-06-20]. https://arxiv.org/pdf/1904.07392.pdf

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Girshick R, Iandola F, Darrell T and Malik J. 2015. Deformable part models are convolutional neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE: 437-446 [DOI: 10.1109/CVPR.2015.7298641]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Griffin G, Holub A and Perona P. 2017. Caltech-256 object category dataset [EB/OL]. [2019-06-20]. https://authors.library.caltech.edu/7694/1/CNS-TR-2007-001.pdf

Guo Y, Li Y L and Wang S J. 2019. Hierarchical structure and joint training for large scale semi-supervised object detection [EB/OL]. (2019-05-30) [2019-06-20]. https://arxiv.org/pdf/1905.12863.pdf

Han G X, Zhang X and Li C R. 2017. Single shot object detection with top-down refinement//Proceedings of 2017 IEEE International Conference on Image Processing. Beijing, China: IEEE: 3360-3364 [DOI: 10.1109/ICIP.2017.8296905]

Havard W, Besacier L and Rosec O. 2017. SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO dataset [EB/OL]. (2017-07-26) [2019-06-20]. https://arxiv.org/pdf/1707.08435.pdf

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.

2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

He, K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He Y H, Zhang X Y, Savvides M and Kitani K. 2018. Softer-NMS: rethinking bounding box regression for accurate object detection [EB/OL]. (2018-09-23) [2019-06-20]. https://arxiv.org/pdf/1809.08545.pdf

Hoffman J, Guadarrama S, Tzeng E, Hu J, Donahue J, Girshick R, Darrell T and Saenko K. 2014. LSDA: large scale detection through adaptation [EB/OL]. (2014-07-18) [2019-06-20]. https://arxiv.org/pdf/1407.5035.pdf

Howard A, Sandler M, Chu G, Chen L C, Chen B, Tan M X, Wang W J, Zhu Y K, Pang R M, Vasudevan V, Le Q V and Adam H. 2019. Searching for MobileNetV3 [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1905.02244.pdf

Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, Andreetto M and Adam H. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2019-06-20]. https://arxiv.org/pdf/1704.04861.pdf

Huang G, Liu Z, van der Maaten L and Weinberger, K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Juan L and Gwun O. 2013. A comparison of SIFT, PCA-SIFT and SURF. International Journal of Image Processing (IJIP), 3(4): 143-152.

Karlinsky L, Shtok J, Harary S, Schwartz E, Aides A, Feris R, Giryes R and Bronstein A M. 2018. RepMet: representative-based metric learning for classification and one-shot object detection [EB/OL]. [2019-06-20]. https://arxiv.org/abs/1806.04728

Kocabas M, Karagoz S and Akbas E. 2019. Self-supervised learning of 3D human pose using multi-view geometry [EB/OL]. (2019-04-09) [2019-06-20]. http://arxiv.org/pdf/1903.02330.pdf

Kong T, Sun F C, Yao A B, Liu H P, Lu M and Chen Y R. 2017. RON: reverse connection with objectness prior networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 5244-5252 [DOI: 10.1109/CVPR.2017.557]

Kong T, Yao A B, Chen Y R and Sun F C. 2016. Hypernet: towards accurate region proposal generation and joint object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 845-853 [DOI: 10.1109/CVPR.2016.98]

Krasin I, Duerig T, Alldrin N, Ferrari V, Abu-El-Haija S, Kuznetsova A, Rom H, Uijlings J, Popov S, Kamali S, Malloci M, PontTuset J, Veit A, Belongie S, Gomes V, Gupta A, Sun C, Chechik G, Cai D, Feng Z, Narayanan D and Murphy K. 2017. OpenImages: a public dataset for large-scale multi-label and multi-class image classification [EB/OL]. [2019-06-20]. https://github.com/openimages.

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/Nature14539]

Li J N, Liang X D, Wei Y C, Xu, T F, Feng J S and Yan S C. 2017. Perceptual generative adversarial networks for small object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 1951-1959 [DOI: 10.1109/CVPR.2017.211]

Li Y H, Chen Y T, Wang N Y and Zhang Z X. 2019. Scale-aware trident networks for object detection [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1901.01892.pdf

Li Z M, Peng C, Yu G, Zhang X Y, Deng Y D and Sun J. 2018. DetNet: a backbone network for object detection [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1804.06215.pdf

Lin M, Chen Q and Yan S C. 2014. Network in network [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1312.4400.pdf

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu L, Ouyang W L, Wang X G, Fieguth P, Chen J, Liu X W and Pietikäinen M. 2019a. Deep learning for generic object detection: a survey [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1809.02165.pdf

Liu P J, Zhang H Z, Zhang K, Lin L and Zuo W M. 2018a. Multi-level wavelet-CNN for image restoration//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, UT, USA: IEEE: 773-782 [DOI: 10.1109/CVPRW.2018.00121]

Liu S T, Huang D and Wang Y H. 2018b. Receptive field block net for accurate and fast object detection [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1711.07767.pdf

Liu T, Zhao Y, Wei Y C, Zhao Y F and Wei S K. 2019b. Concealed object detection for activate millimeter wave image. IEEE Transactions on Industrial Electronics, 66(12): 9909-9917 [DOI: 10.1109/TIE.2019.2893843]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Ma N N, Zhang X Y, Zheng H T and Sun J. 2018. ShuffleNet V2:

practical guidelines for efficient CNN architecture design [EB/OL]. (2018-07-30) [2019-06-20]. https://arxiv.org/pdf/1807.11164.pdf

Nielsen F Å. 2018. Linking ImageNet WordNet synsets with wikidata [EB/OL]. (2018-03-05) [2019-06-20]. https://arxiv.org/pdf/1803.04349.pdf

Ning X F, Zhu W and Chen S F. 2017. Recognition, object detection and segmentation of white background photos based on deep learning//Proceedings of the 32nd Youth Academic Annual Conference of Chinese Association of Automation. Hefei, China: IEEE: 182-187 [DOI: 10.1109/YAC.2017.7967401]

Pham H, Guan M Y, Zoph B, Le Q V and Dean J. 2018. Efficient neural architecture search via parameter sharing [EB/OL]. [2019-06-20]. http://arxiv.org/pdf/1802.03268.pdf

Qi C, Ouyang W L, Li H S, Wang X G, Liu B and Yu N H. 2017. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism//Proceedings of 2017 IEEE International Conference on Computer Vision: 4836-4845

Rahman S, Khan S and Porikli F. 2018. Zero-shot object detection: learning to simultaneously recognize and localize novel concepts//Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer: 547-563 [DOI: 10.1007/978-3-030-20887-5_34]

Real E, Aggarwal A, Huang Y P and Le Q V. 2019. Regularized evolution for image classifier architecture search [EB/OL]. [2019-06-20]. http://arxiv.org/pdf/1802.01548.pdf

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1804.02767.pdf

Ren M Y, Triantafillou E, Ravi S, Snell J, Swersky K, Tenenbaum J B, Larochelle H and Zemel R S. 2018. Meta-learning for semi-supervised few-shot classification [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1803.00676.pdf

Rosario V M D, Borin E and Breternitz Jr, M. 2019. The multi-lane capsule network (MLCN) [EB/OL]. [2019-06-22]. https://arxiv.org/pdf/1902.08431.pdf

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2019. Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1801.04381v1.pdf

Seferbekov S, Iglovikov V, Buslaev A and Shvets A. 2018. Feature pyramid network for multi-class land segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, UT, USA: IEEE: 272-2723 [DOI: 10.1109/CVPRW.2018.00051]

Shan Y H, Lu W F and Chew C M. 2019. Pixel and feature level based domain adaptation for object detection in autonomous driving [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1810.00345.pdf

Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1409.1556.pdf

Singh B, Li H D, Sharma A and Davis L S. 2018. R-FCN-3000 at 30fps: decoupling detection and classification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 1081-1090 [DOI: 10.1109/CVPR.2018.00119]

Sun K, Xiao B, Liu D and Wang J D. 2019a. Deep high-resolution representation learning for human pose estimation [EB/OL]. (2019-02-25) [2019-06-20]. https://arxiv.org/pdf/1902.09212.pdf

Sun Y F, Xu Q, Li Y, Zhang C, Li Y K, Wang S J and Sun J. 2019b. Perceive where to focus: learning visibility-aware part-level features for partial person re-identification [EB/OL]. (2019-04-01) [2019-06-20]. http://arxiv.org/pdf/1904.00537.pdf

Szegedy C, Ioffe S, Vanhoucke V and Alemi A. 2016. Inception-v4, inception-ResNet and the impact of residual connections on learning [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1602.07261.pdf

Tan M X, Chen B, Pang R M, Vasudevan V, Sandler M, Howard A, and Le Q V. 2019. MnasNet: platform-aware neural architecture search for mobile [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1807.11626.pdf

Torralba A, Fergus R and Freeman W T. 2008. 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11): 1958-1970 [DOI: 10.1109/TPAMI.2008.128]

Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M. 2013. Selective search for object recognition. International Journal of Computer Vision, 104(2): 154-171 [DOI: 10.1007/s11263-013-0620-5]

Verma S and Zhang Z L. 2018. Graph capsule convolutional neural networks [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1805.08090.pdf

Wang R J, Li X and Ling C X. 2019a. Pelee: a real-time object detection system on mobile devices [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1804.06882.pdf

Wang X D, Cai Z W, Gao D S and Vasconcelos N. 2019b. Towards universal object detection by domain attention [EB/OL]. [2019-06-

20]. https://arxiv.org/abs/1904.04402.pdf

Wang X Y, Han T X and Yan S C. 2009. An HOG-LBP human detector with partial occlusion handling//Proceedings of the 12th IEEEInternational Conference on Computer Vision. Kyoto, Japan: IEEE: 32-39 [DOI: 10.1109/ICCV.2009.5459207]

Wei Y, Pan X Y, Qin H W, Ouyang W L and Yan J J. 2018. Quantization mimic: towards very tiny CNN for object detection [EB/OL]. [2019-06-20]. http://openaccess.thecvf.com/content_ECCV_2018/papers/Yi_Wei_Quantization_Mimic_Towards_ECCV_2018_paper.pdf

Williams T and Li R. 2018. An ensemble of convolutional neural networks using wavelets for image classification. Journal of Software Engineering and Applications, 11(2): 69-88 [DOI: 10.4236/jsea.2018.112004]

Wu Z, Bodla N, Singh B, Najibi M, Chellappa R and Davis L S. 2018. Soft sampling for robust object detection [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1806.06986.pdf

Xiao J X, Ehinger K A, Hays J, Torralba A and Oliva A. 2016. SUN database: exploring a large collection of scene categories. International Journal of Computer Vision, 119(1): 3-22 [DOI: 10.1007/s11263-014-0748-y]

Yan Y C, Zhang Q, Ni B B, Zhang W D, Xu M H and Yang X K. 2019. Learning context graph for person search [EB/OL]. (2019-04-03) [2019-06-20]. http://arxiv.org/pdf/1904.01830.pdf

Yang, T J, Howard A, Chen B, Zhang X, Go A, Sandler M, Sze V and Adam H. 2018. NetAdapt: platform-aware neural network adaptation for mobile applications [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1804.03230.pdf

Zagoruyko S, Lerer A, Lin T Y, Pinheiro P O, Gross S, Chintala S and Dollár P. 2016. A MultiPath network for object detection [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1604.02135.pdf

Zhang K, Zuo W M and Zhang L. 2019a. Deep plug-and-play super-resolution for arbitrary blur kernels [EB/OL]. (2019-03-29) [2019-06-20]. http://arxiv.org/pdf/1903.12529.pdf

Zhang X Y, Zhou X Y, Lin M X and Sun J. 2018a. ShuffleNet: an extremely efficient convolutional neural network for mobile devices//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 6848-6856 [DOI: 10.1109/CVPR.2018.00716]

Zhang Z, He T, Zhang H, Zhang Z Y, Xie J Y, Li M and Services A W. 2019b. Bag of freebies for training object detection neural networks [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1902.04103.pdf

Zhang Z S, Qiao S Y, Xie C, Shen W, Wang B and Yuille A L. 2018b. Single-shot object detection with enriched semantics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 5813-582 [DOI: 10.1109/CVPR.2018.00609]

Zhang S F, Wen L Y, Bian X, Lei Z and Li S Z. 2018c. Single-shot refinement neural network for object detection//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [s.l.]: IEEE: 4203-4212

Zhao Q J, Sheng T, Wang Y T, Tang Z, Chen Y, Cai L and Ling H B. 2019. M2det: a single-shot object detector based on multi-level feature pyramid network [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1811.04533.pdf

Zhou B L, Lapedriza A, Khosla A, Oliva A and Torralba A. 2018a. Places: a 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1452-1464 [DOI: 10.1109/TPAMI.2017.2723009]

Zhou P, Ni B B, Geng C, Hu J G and Xu Y. 2018b. Scale-transferrable object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 528-537 [DOI: 10.1109/CVPR.2018.00062]

Zhou X Y, Gong W, Fu W L and Du F T. 2017. Application of deep learning in object detection//Proceedings of 2017 IEEE/ACIS 16th International Conference on Computer and Information Science. Wuhan, China: IEEE: 631-634 [DOI: 10.1109/ICIS.2017.7960069]

Zoph B and Le Q V. 2019. Neural architecture search with reinforcement learning [EB/OL]. [2019-06-20]. https://arxiv.org/pdf/1611.01578.pdf

作者简介

赵永强,1996 年生,男,硕士研究生,主要研究方向为深度学习、计算机视觉和模式识别。

E-mail: yongqiang1210@stu.xjtu.edu.cn

饶元,通信作者,男,副教授,博士生导师,主要研究方向为深度学习、机器学习、计算机视觉。

E-mail: raoyuan@mail.xjtu.edu.cn

董世鹏,男,硕士研究生,主要研究方向为计算机视觉与模式识别。E-mail: 624566671@qq.com

张君毅,男,博士研究生,主要研究方向为深度学习与图像识别。E-mail: zhangjunyi0806@xjtu.edu.cn