目标检测算法研究综述

方路平[1],何杭江[1],周国民[2]
FANG Luping[1], HE Hangjiang[1], ZHOU Guomin[2]

1.浙江工业大学 信息工程学院,杭州 310023
2.浙江警察学院 计算机与信息技术系,杭州 310053
1.College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
2.Department of Computer And Information Technology, Zhejiang Police College, Hangzhou 310053, China

FANG Luping, HE Hangjiang, ZHOU Guomin. Research overview of object detection methods. Computer Engineering and Applications, 2018, 54(13):11-18.

Abstract: Object detection is an important problem in computer vision, which has critical research value in the field of pedestrian tracking, license plate recognition and unmanned driving. In recent years, the accuracy of image classification is greatly improved with deep learning, thus the object detection methods based on deep learning have gradually become mainstream. The development and present situation of object detection methods are reviewed, and a prospect is made. Firstly, the development, improvement and deficiency of the traditional algorithms and depth learning-based algorithms are summarized, and then compared. Finally, the difficulties and challenges of the object detection method based on deep learning are discussed, and the possible development direction is prospected.

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

摘 要: 目标检测是计算机视觉中一个重要问题,在行人跟踪、车牌识别、无人驾驶等领域都具有重要的研究价值。近年来,随着深度学习对图像分类准确度的大幅度提高,基于深度学习的目标检测算法逐渐成为主流。梳理了目标检测算法的发展与现状,并作出展望:总结了传统算法与引入深度学习的目标检测算法的发展、改进与不足,并就此做出对比;最后讨论了基于深度学习的目标检测算法所存在的困难与挑战,并就可能的发展方向进行了展望。

关键词: 目标检测;深度学习;计算机视觉;卷积神经网络;目标分类检测
文献标志码: A **中图分类号:** TP183 doi:10.3778/j.issn.1002-8331.1804-0167

1 引言

视觉,作为人类接收信息的主要方式之一,负责超过80%的信息获取。视觉计算理论创始人Marr[1]认为视觉的主要作用是将二维的图像通过计算进行三维重建,也就是对空间物体的识别和理解。和人类视觉基本功能一样,计算机视觉中物体的分类和检测,一直是一个重要问题。随着计算机技术的迅猛发展,目标检测已在人脸识别、行人跟踪、车牌识别、无人驾驶等领域获得广泛应用。

相比于图像分类,目标检测更具难度。目标检测,就是将目标定位和目标分类结合起来,利用图像处理技术、机器学习等多方向的知识,从图像(视频)中定位感兴趣的对象。目标分类负责判断输入的图像中是否包含所需物体(object),目标定位则负责表示目标物体的位置,并用外接矩形框定位。这需要计算机在准确判断目标类别的同时,还要给出每个目标相对精确的位置。

自目标检测的概念提出以来,国内外学者针对这个问题做出了不懈探索。传统的目标检测算法,多是基于滑动窗口的框架或是根据特征点进行匹配。自2012年AlexNet[2]在当年度ImageNet大规模视觉识别挑战赛中一举夺冠,且效果远超传统算法,将大众的视野重新带回到深度神经网络。2014年R-CNN[3]的提出,使得基于

基金项目: 国家自然科学基金(No.U1509219,No.81771481)。

作者简介: 方路平(1968—),男,硕士,教授,研究领域为机器学习、云计算、物联网技术及其应用,E-mail:flp@zjut.edu.cn;何杭江(1994—),男,硕士研究生,研究领域为机器学习、图形图像研究;周国民(1971—),男,硕士,副教授,研究领域为公安应用技术、图形图像研究。

收稿日期: 2018-04-16 **修回日期:** 2018-06-01 **文章编号:** 1002-8331(2018)13-0011-08

CNN[4]的目标检测算法逐渐成为主流。

深度学习的应用,使检测精度和检测速度都获得了改善。因此,笔者认为,目标检测算法可以根据是否应用深度学习,将之分为传统算法与基于深度学习的目标检测算法。

本文将对传统算法与基于深度学习的目标检测算法的主要算法进行论述,并分析相关算法的优缺点,结合现有的问题,对将来的发展作出预测。

2　传统算法

传统算法大致可以分为目标实例检测与传统目标类别检测两类:

(1)目标实例检测问题通常利用模板和图像稳定的特征点,获得模板与场景中对象的对应关系,检测出目标实例。目标实例检测关注的只是具体目标本身,图像中的其余对象都是无关量。

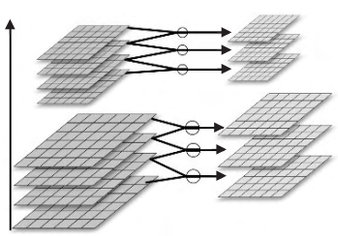
(2)传统目标类别检测则通过使用AdaBoost[5]算法框架、HOG[6]特征和支持向量机[7]等方法,根据选定的特征和分类器,检测出有限的几种类别。

2.1　SIFT系列算法

2.1.1　SIFT算法

Lowe提出的SIFT[8]算法,通过查找不易受光照、噪声、仿射变换影响的特征点来匹配目标,是目前应用极为广泛的关键点检测和描述算法。

该算法通过使用高斯模糊实现尺度空间,高斯差分函数(Difference of Gaussian)进行极值检测,再通过对边缘主曲率的判定,筛除边缘响应的不稳定点,得到匹配稳定、抗噪能力强的关键点。最后利用方向直方图统计关键点邻域梯度和方向,获得描述符。



(a)高斯金字塔　(b)高斯差分金字塔

图1　高斯差分金字塔示意图

SIFT算法通过一系列方法,保证提取的特征具有平移、缩放及旋转不变等特性,对于光线、噪声、少量视角改变也具有一定的鲁棒性,针对部分遮挡也有不错的识别率。但是,SIFT算法存在复杂度高,检测速度慢,对模糊图像和光滑边缘很难提取有效特征点等问题。

2.1.2　PCA-SIFT算法

对于SIFT存在的问题,Ke等人提出了PCA-SIFT[9]算法。该算法在SIFT的基础上,对其最后一步做出了改进。引入主成分分析(PCA)方法,使用PCA替代直方图,来对描述子向量进行降维,以提高匹配效率。

相较SIFT,PCA-SIFT维数更少且灵活可变,检测速度约为SIFT的3倍。但降维损失部分信息,导致只对具有代表性的图像有较好效果,具有局限性。

2.1.3　SURF算法

SURF[10]算法也是一种基于SIFT的改进算法,Hessian矩阵是该算法的核心。该算法利用高斯滤波保证尺度无关性,并用盒式(box)滤波器替代高斯滤波器,简化计算。通过Hessian矩阵的构建,获取关键点定位。另外,在尺度空间中,不同于SIFT构建不同尺度的图像,SURF保持图像大小不变,只改变滤波器的大小,从减少了计算量。

简单来说,SURF算法利用近似的Hessian矩阵减少降采样过程,快速构建尺度金字塔,实现了目标检测速度的提高。

2.1.4　改进

除了上述算法以外,还有面向稠密特征提取的DAISY[11]描述子、可以较好处理大视角变化情况的ASIFT[12]算法、结合Fast[13]和Brief[14]的速度优势的ORB[15]算法。其中,ORB算法利用Fast检测特征点,Brief计算特征点的描述子,其特征点性能介于SIFT与SURF之间,运行效率却远快于SURF。

2.1.5　比较

对于SIFT、PCA-SIFT以及SURF这三种算法,笔者作出如下总结:

PCA-SIFT与SURF算法分别对SIFT的匹配过程做出简化,因此在特征点匹配精度上必然有所下降。其中SIFT算法提取的特征点最为丰富,在尺度、旋转等情况下都具有最好的性能,但高复杂度导致检测速度最慢,且对于模糊、光滑边缘的提取效果并不理想;PCA-SIFT使用PCA方法进行降维,减少计算的同时,产生信息丢失,因此整体性能在三种算法中比较一般;SURF合理利用积分图减少运算,小波变换、Hessian矩阵等方法基本不会降低精度,因此在获得好检测速度的同时,也保证了整体性能优于PCA-SIFT。

2.2　基于AdaBoost系列算法

2.2.1　AdaBoost算法

AdaBoost是一种是基于Boosting[16]的机器学习算法。初始时,设训练集中 n 个样本具有相同的权重。在每次训练后调整训练集中数据权重,增加错误样本的权重,使得下一个分类器能够对错误样本进行重点训练。经过 N 轮训练后,将 N 个弱分类器整合,根据各分类器的性能分配相应的权值,组成一个高准确率、低错误率的强分类器。

2.2.2　Viola-Jones算法

Viola-Jones[17-18]算法是第一种能实时处理且效果较好的人脸检测算法,此算法的提出标志着人脸检测进入实际应用阶段。

表1　传统目标检测算法对比

方法	分类	算法逻辑	优点	缺点	适用场合
SIFT	实例检测	提取平移、缩放、旋转不变的描述子用以匹配	检测特征丰富，具有优秀匹配效果	计算量大，检测速度慢	图像识别(无速度要求)、图像拼接、图像恢复等
PCA-SIFT	实例检测	PCA降维减少运算	检测速度获得改善	不完全仿射不变，检测精度不高	特征点匹配
SURF	实例检测	近似Hessian矩阵，积分图减少降采样	检测速度快，精度较高，综合性能好	过于依赖主方向的选取准确度	物体识别、3D重构
ORB	实例检测	Fast检测特征点，Brief计算特征点描述子	检测速度快，检测精度良好	不具备尺度不变性	实时视频处理
VJ	类别检测	Haar+AdaBoost检测目标，级联减少计算量	第一种能够实时检测的人脸检测算法	准确率一般，鲁棒性不足	人脸/物体检测

Viola-Jones检测算法(简称VJ算法)使用Haar特征来描述窗口，反映局部区域的明暗变化，并利用积分图的思路解决Haar特征提取时计算量大、重复的缺点。同时，引入级联的思想。如图2所示，VJ根据分类器的复杂程度和计算代价排列，分类代价越高的分类器需要分类的图像越少，减少分类工作量。

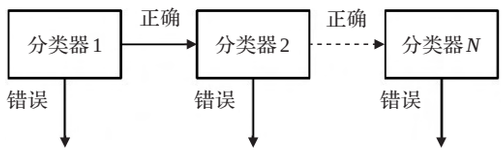


图2　级联示意图

概括地说，VJ算法利用Haar-like特征描述目标共有属性，利用积分图实现特征快速计算，使用级联分类器减少AdaBoost的计算量，快速检测出目标。

2.2.3　改进

Rainer Lienhart和Jochen Maydt将Viola-Jones检测器用对角特征进行扩展，形成了Haar[19]分类器。除此之外，也有将Stump函数改为决策树、使用RealBoost、GentleBoost等分类器的算法陆续提出。

2.3　总结

如表1所示，本文针对传统算法做出对比总结。总的来说，这些算法的目的都是在保证提取丰富、准确特征的前提下，快速地进行特征计算及预测。但传统算法提取的特征基本都是低层次、人工选定的特征，这些特征相对更直观、易理解，针对特定对象更有针对性，但不能很好地表达大量、多类目标。

3　基于深度学习的目标检测算法

自从AlexNet在比赛中使用卷积神经网络进而大幅度提高了图像分类的准确率，便有学者尝试将深度学习应用到目标类别检测中。卷积神经网络不仅能够提取更高层、表达能力更好的特征，还能在同一个模型中完成对于特征的提取、选择和分类。在这方面，主要有两种主流的算法：一类是结合region proposal、CNN网络的，基于分类的R-CNN系列目标检测框架(two stage)；另一类则是将目标检测转换为回归问题的算法(single stage)。

3.1　基于分类的检测算法

Region proposal(候选区域)[20]是通过Selective Search[21]等算法，根据图像中纹理、边缘、颜色等信息，检测较少区域的同时保证了较高的召回率。

3.1.1　OverFeat算法

OverFeat[22]是最先将深度学习应用到目标检测中的算法之一。

严格来说，OverFeat并没有使用region proposal，但其思路被后面的R-CNN系列沿用并改进。该算法通过多尺度的滑动窗口结合AlexNet提取图像特征，完成检测。在ILSVRC 2013数据集上的平均准确率(mean Average Precision，mAP)为24.3%，检测效果较传统算法有显著改进，但依旧存在较高错误率。

3.1.2　R-CNN算法

在Overfeeat提出后不久，Ross Girshick等人提出了R-CNN模型。如图3所示，R-CNN利用Selective Search获得候选区域(约2 000个)。随即对候选区域大小进行归一化，用作CNN网络的标准输入。再使用AlexNet获得候选区域中的特征，最后利用多个SVM进行分类以及线性回归微调定位框(Bounding-box)。

图3　R-CNN结构示意图

R-CNN将检测效果从OverFeat的24.3%大幅提升至31.4%(ILSVRC 2013数据集)，并在VOC2007数据集上获得58.5%的准确率(下文中如无特殊说明，皆为VOC2007数据集上检测结果)。但是，R-CNN对近2 000个候选区域分别做特征提取，而候选区域之间存在许多重复区域，导致大量且重复的运算，运行缓慢，平均每幅图片的处理时间为34 s。同时，对每一步的数据进行存储，极为损耗存储空间。另外，对候选区域进行归一化操作，会对最终结果产生影响。

3.1.3 SPP-Net

如图4所示,针对R-CNN对所有候选区域分别提取特征的缺点,SPP-Net[23]一次性对整张图片作卷积操作提取特征。使得特征提取从R-CNN的近2 000次变为提取1次整张图片特征,大大减少了工作量。

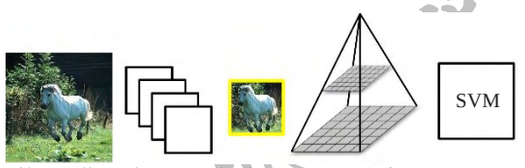


图4 SPP-Net结构示意图

另外,SPP-Net在最后一个卷积层后、全连接层前添加空间金字塔池化层(SPP层),提取固定尺寸的特征向量,避免对候选区域大小进行归一化的复杂操作。

以上两点改进使得SPP-Net的检测速度比R-CNN快38~102倍,并解决了候选区域归一化问题。SPP-Net虽然更换了卷积网络,但准确率相差无几。同时,SPP-Net依然没有解决R-CNN存储空间消耗的问题,确定候选区域、特征提取、对象分类、定位修正这些步骤依然是分离的。

3.1.4 Fast-RCNN

Fast R-CNN[24]算法在SPP-Net的基础上,将SPP层简化为ROI Pooling层,并将全连接层的输出作SVD分解,得到两个输出向量:softmax的分类得分以及Bounding-box外接矩形框的窗口回归。

这种改进将分类问题和边框回归问题进行了合并;用softmax代替SVM,将所有的特征都存储在显存中,减少了磁盘空间的占用;SVD分解则在几乎不影响精度的情况了,极大加快检测速度。

Fast R-CNN使用VGG16代替AlexNet,平均准确率达到70.0%,且训练速度较R-CNN提升9倍,检测速度达到每幅图片0.3 s(除去region proposal阶段)。Fast R-CNN依然使用Selective Search方法选取候选区域,这一步骤包含大量计算。在CPU上运行时,获取每张图片的候选区域平均需要2 s。由此可见,改进Selective Search是Fast R-CNN速度提升的关键。

3.1.5 Faster-RCNN

SPP-Net和Fast R-CNN从特征提取的角度,减少了工作量,但依然没有解决Selective Search选择候选区域速度慢的问题。Faster R-CNN[25]使用RPN网络(Region Proposal Networks)替代Selective Search算法,使目标识别实现真正端到端的计算。

如图5所示,RPN网络通过在特征图上做划窗操作,使用预设尺度的锚点框映射到原图,得到候选区域。RPN网络输入的特征图和全连接层中的特征图共享计算。RPN的使用,使Faster R-CNN能够在一个网络框架之内完成候选区域、特征提取、分类、定位修正等操作。

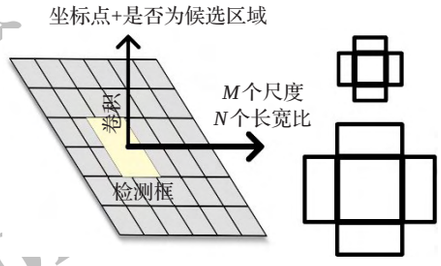


图5 RPN网络示意图

RPN使得Faster R-CNN在region proposal阶段只需10 ms,检测速度达到5 f/s(包括所有步骤),并且检测精度也得到提升,达到73.2%。但是,Faster R-CNN仍然使用ROI Pooling,导致之后的网络特征失去平移不变性,影响最终定位准确性;ROI Pooling后每个区域经过多个全连接层,存在较多重复计算;Faster R-CNN在特征图上使用锚点框对应原图,而锚点框经过多次下采样操作,对应原图一块较大的区域,导致Faster R-CNN检测小目标的效果并不是很好。

3.1.6 R-FCN

目标检测要包括两个问题:分类问题和检测定位问题。前者具有平移不变性,后者具有平移敏感性。

R-FCN[26]使用全卷积网络ResNet[27]代替VGG,提升特征提取与分类的效果;针对全卷积网络不适应平移敏感性的缺陷,该算法使用特定的卷积层生成包含目标空间位置信息的位置敏感分布图(Position Sensitive Score Map);ROI Pooling层后不再连接全连接层,避免重复计算。

R-FCN的准确率达到83.6%,测试每张图片平均花费170 ms,比Faster-RCNN快了2.5~20倍。但是R-FCN在得到Score map需要生成一个随类别数线性增长的channel数,这一过程虽然提升了目标检测精度,但减慢了检测速度,导致其难以满足实时性要求。

3.1.7 Mask R-CNN

Mask R-CNN[28]是一种在Faster R-CNN基础上加以改进的算法,增加了对实例分割的关注。该算法在分类和定位回归以外,加入了关于实例分割的并行分支,并将三者的损失联合训练。

实例分割要求实例定位的精准度达到像素级,而Faster R-CNN因为ROI Pooling层的等比例缩放过程中引入了误差,导致空间量化较为粗糙,无法准确定位。Mask R-CNN提出双线性差值RoIAlign获得更准确的像素信息,使得掩码(mask)准确率提升10%到50%;Mask R-CNN还使用ResNeXt[29]基础网络,在COCO数据集上的检测速度为5 f/s,检测准确性从Fast R-CNN的19.7%提升至39.8%。

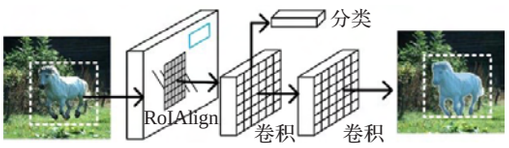


图6　Mask R-CNN结构示意图

Mask R-CNN在检测精度、实例分割方面都达到目前最高的层次。其后一些算法在性能上有所提升，但基本维持在同一水平。但是该算法的检测速度依旧难以满足实时要求，并且实例分割目前也还面临着标注代价过于昂贵的问题。

3.1.8　改进

HyperNet[30]提出结合底层、中间层以及高层的Hyper Feature，在小物体的处理上获得更好的效果；A-Fast-RCNN[31]则引入GAN[32]，生成高难度的样本以提高网络对遮挡、形变的适应性；Light Head R-CNN[33]针对Faster R-CNN以及R-FCN的head部分，减少密集计算。尽量保持在精度的同时减少了计算量，达到102 f/s(Xception网络)。

3.1.9　小结

如图7所示，从R-CNN开始，研究者将目标检测的问题关注点集中到分类上，采用“region proposal+CNN feature+SVM”的思路，利用了CNN网络，大大提高了检测的精度；后面的SPP-Net、Fast-RCNN、Faster-RCNN等基本沿用了这一思路，在检测效率上进行改进；但Faster-RCNN只能达到5 f/s，就实时性而言略有不足。随后的R-FCN虽然有所提升，但效果依然无法令人满意。对此，研究者提出了另一种新思路，直接将目标检测转化到回归上，用一张图片得到bounding box以及类别。

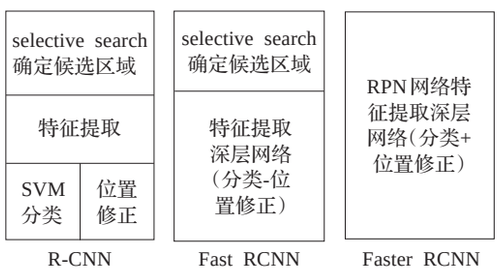


图7　R-CNN、Fast R-CNN、Faster R-CNN对比

3.2　基于回归的检测算法

3.2.1　YOLO

从R-CNN到Faster-RCNN，目标检测始终遵循“region proposal+分类”的思路，训练两个模型必然导致参数、训练量的增加，影响训练和检测的速度。由此，YOLO[34]提出了一种“single-stage”的思路。

如图8所示，YOLO将图片划分为$S\times S$的网格(cell)，各网格只负责检测中心落在该网格的目标，每个网格需要预测两个尺度的bounding box和类别信息，一次性预测所有区域所含目标的bounding box、目标置信度以及类别概率完成检测。

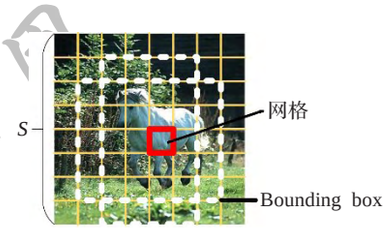


图8　YOLO网格划分示意

YOLO采用以cell为中心的多尺度区域取代region proposal，舍弃了一些精确度以换取检测速度的大幅提升，检测速度可以达到45 f/s，足以满足实时要求；检测精度为63.4%，较Faster R-CNN的73.2%，差距较大。

YOLO在极大提高检测速度的情况下，也存在以下问题：(1)因为每个网格值预测两个bounding box，且类别相同，因此对于中心同时落在一个网格总的物体以及小物体的检测效果差，多物体环境下漏检较多；(2)由于YOLO关于定位框的确定略显粗糙，因此其目标位置定位准确度不如Fast-RCNN；(3)对于外型非常规的物体检测效果不佳。

3.2.2　SSD

Faster-RCNN检测检测精度高但检测速度慢，YOLO检测精度不高但检测速度快，SSD[35]则结合两者的优点，在YOLO的基础上借鉴了RPN的思路，在保证高精度检测的同时，兼顾检测速度。

如图9所示，因为不同层的特征图具有对应大小的感受野，特定层的特征图只需要训练对应尺度的对象检测。因此，SSD结合高层和底层的特征图，使用多尺度区域特征进行回归。

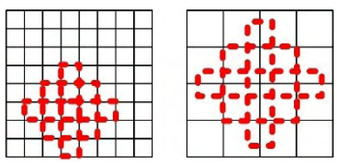


图9　特征图感受野示意图

SSD300的mAP能达到73.2%，基本与Faster R-CNN(VGG16)持平，而检测速度达到59 f/s，比Faster R-CNN快6.6倍。但是SSD具有以下问题：(1)小目标对应到特征图中很小的区域，无法得到充分训练，因此SSD对于小目标的检测效果依然不理想；(2)无候选区域时，区域回归难度较大，容易出现难以收敛等问题；(3)SSD不同层的特征图都作为分类网络的独立输入，导致同一个物体被不同大小的框同时检测，重复运算。

3.2.3　YOLOv2以及YOLO9000

YOLOv2[36]通过在每一个卷积层后添加batch normalization、多尺度训练，加入K-mean维度聚类等方式，使得检测速度和精度的再次提升。该算法能够在76.8%正确率的同时达到67 f/s的检测速度，78.6%的正确率时达到40 f/s。该算法性能基本代表目前业界的最

先进水平。

同文还提出了YOLO9000,该算法采用wordTree层次分类,混合检测数据、识别数据集,在分类和检测数据集上同时训练,实现9 418类的检测。

无论是YOLO系列还是SSD算法,都沿用R-CNN系列算法先在大数据集上进行分类预训练,再在小数据集上fine-tune的方法。但fine-tune预训练模型有以下问题:(1)预训练模型,往往无法迁移到如医疗图像等特定数据上;(2)预训练模型结构基本固定,难以修改;(3)预训练样本和最终检测目标有所区别,得到的模型未必是检测目标的最佳模型。

3.2.4 改进

针对预训练模型的问题,DSOD[37]算法提出一种从零训练网络的方法,达到媲美fine-tune模型的效果。DSOD基于SSD算法,在特征融合部分引入DenseNet[38]思想,减少了参数量;mAP为77.7%,与SSD300相当;检测速度为17.4 f/s,较SSD300的46 f/s尚有较大差距。

R-SSD[39]算法则在SSD的基础上,增加不同层之间特征图的关联,避免同一物体重复框的问题;同时增加特征金字塔中特征图的数量,改善小物体的检测效果。该算法mAP为80.8%,略高于SSD。但是特征图的增加,导致计算量增加,检测速度降低,仅为16.6 f/s。

3.3 其他

除了基于回归和region proposal的两种思路以外,还有放弃bounding box改用概率模型的Locnet[40];基于RNN[41]的RRC[42];使用深度可分离卷积的网络结构,用较小参数量达到大型网络相当精度的MobileNets[43]以及SSDLite[44]等方法。

另外,近两年还出现了一些针对特定场景特定目标的检测。Lea等人提出了用于特定动作分割和检测的网络[45],Dong[46]和Chen等[47]则关注于3D目标检测。这些算法更多地结合场景信息、空间信息与时间信息,而不是单纯的定位分类。

表2对基于深度学习的目标检测模型性能做出对比,表3对模型实时性与优缺点做出总结。

表2 基于深度学习目标检测模型性能对比

模型	基础网络	ILSVRC 2013	VOC2007	检测速度/(f·s^{-1})
OverFeat	AlexNet	24.3	—	—
R-CNN	AlexNet	31.4	58.5	34 s/张
SPP-Net	ZF-5	—	59.2	2
Fast R-CNN	VGG16	—	70.0	3
Faster R-CNN	VGG16	—	73.2	5
R-FCN	ResNet-101-C4	—	83.6	6
Mask R-CNN	ResNeXt	—	—	5
YOLO	Custom	—	63.4	45
SSD	VGG16	—	73.2	59
YOLOv2 416	DarkNet-19	—	76.8	67
DSOD300	DS/64-192-48-1	—	77.7	17.4
R-SSD	VGG16	—	80.8	16.6

4 思考与展望

基于深度学习的目标检测在检测精度以及检测速度上,较传统方法获得了极大的提高,但依然面临这一些问题:

(1)对于小数据量,目前的框架可能无法得到好的结果。目前的算法,大多使用了迁移学习,也就是现在现有的大数据集中进行训练,再将训练好的“半成品”做fine-tune操作。若目标数据不在ImageNet等数据集中,训练效果要视目标与大数据集相关程度而定。DSOD算法虽然设计了一种从零开始训练的网络,也取得了不错的效果,但是其检测速度尚有待提升。

(2)深度学习解释性差,特别是在更深的层次上,很多时候只能依靠测试和经验来猜测其有效或无效的原因,对于中间的过程缺少明确的解释,更像是一个黑盒。

(3)计算强度大。GPU的使用,提升了计算机的运算能力,但是很多操作依然过于庞大。如何简化、复用计算的同时,尽可能保证准确率,可能会是一个可以创新的点。

表3 基于深度学习目标检测模型优缺点对比

模型	实时性	优点	缺点
OverFeat	否	最早使用CNN进行特征提取	图像滑窗,时间、空间开销大
R-CNN	否	确定候选区域,CNN提取特征,SVM分类,性能比传统算法显著提高	对每个候选区域都做特征提取,时间、空间开销大
SPP-Net	否	整张图片提取特征,加快速度;SPP层,避免候选区域归一化	空间开销大
Fast R-CNN	否	同时完成定位和分类,节省空间	候选区域选取方法计算复杂,
Faster R-CNN	较差	真正完成端到端训练测试	模型复杂,小目标检测不佳,空间量化粗糙
R-FCN	较差	定位精度更高	模型复杂,计算量大
Mask R-CNN	较差	实例分割准确、检测精度更高	实例分割代价昂贵
YOLO	优秀	网络简单,检测速度优异	定位准确度低,小目标、多目标检测效果不佳
SSD	优秀	网络简单,检测准确度获得高	模型难收敛,小目标检测效果不佳
YOLOv2 416	优秀	允许用户在精度和速度之间调整	使用预训练,难迁移
DSOD300	较好	不需要预训练	检测速度一般
R-SSD	较好	小目标检测效果较好	模型计算复杂,检测速度一般

(4)对于场景信息、语义信息等视频中原有信息的利用不充分，造成一些有效信息的损失。

(5)无论是R-CNN系列还是SSD等算法，始终无法在小目标检测问题上获得令人满意的效果。就目前算法而言，为保证检测速度，通常减少特征金字塔的图像，以减少计算量，但这必然导致小目标在特征图上得不到充分训练；如R-SSD增加特征图数量，损失了检测速度。此问题与问题(3)有一定相通之处。

针对上述问题以及近两年研究趋势，结合周俊宇等人在此方向上所做的研究[48]，本文对目标检测算法未来的发展方向做出讨论：

(1)更多更全面的数据集。目前有两种解决思路：一种是人工手动标注，对于小数据量而言，操作简单且能保证较高正确率，但对大数据量以及物体分割要求精准标注的数据时，力有不逮；另一种是使用平行视觉方法。张慧等人[49]在展望中提到了王坤峰等[50]提出的平行视觉思路，旨在利用人工场景模拟实际场景，通过计算实验对模型进行设计和评估，平行执行在线优化视觉系统。平行视觉如果实现，那么将解决标注数据集不足的问题，促进目标检测发展。

(2)更多的计算共享。不论是R-CNN系列还是基于回归的检测算法，都是为了让不同的ROI之间的计算量得到更多的共享，以达到加快运算的目的。

(3)RNN思想的尝试。视频本身是包含上下文信息的，这是人类做出某些判断的依据。深度学习是一种类人的“学习”方式，结合深度学习中的循环神经网络思想是一种较可能实现的思路。另外，结合具体场景及语意信息，真正去“理解”场景也是一种思路。

(4)更具体的应用场景。高红红等人[51]针对监控视频光线暗、画质差的特点，提出一种基于背景分类的检测算法；赵燕熙等人[52]采用自底向上和自顶向下相结合的方式，消除动态背景的干扰，取得较好的效果；Wang等人[53]提出一种利用卷积网络检测视频中显著目标的模型；Li等人[54]提出一种检测小型交通标志的网络；Dong[46]和Chen等人[47]探索如何将自然环境中的目标检测转换为3D；Dave等人[55]更关注对目标具体动作的识别。可以看出，目标检测，特别是基于深度学习的目标检测，正在向着更具体、更实际的场景发展。

(5)“新”神经网络的应用。从AlexNet到VGG再到ResNet和ResNext，基础网络的改进，也是目标检测效果不断提升原因之一。早在2011年，被誉为“神经网络之父”的Hinton就提出capsule的概念。他在2017年的论文中提出了捕捉空间结构信息的capsule概念[56]，用向量输出代替标量输出，改善CNN网络各特征之间联系缺失，需要大量数据集的问题。

5 结束语

从最初的人为寻找特征到最近的基于深度学习的目标检测算法，可以看出对于目标检测的要求始终是快速、精准以及适用范围广。就目前来说，传统的目标检测方法仍在使用，且在一段时间内仍会有一定市场。传统的目标检测技术对数据量要求少，在针对数据来源不够丰富的项目时，可能会取得比深度学习更好的效果。但是将深度学习应用到目标检测中是可以预见的主流趋势。特别是随着硬件设备性能的提升，一定范围内的运算量处理将不会再成为实时检测的掣肘。

如何利用上下文关联信息、场景信息和语义信息，将会是接下来目标检测的一个重要研究方向。假使平行视觉的思路切实可行，那么数据集标注困难、数据量不足的问题，将获得较好的解决。另外，如何更好解决与训练集关联性不大的小数据集检测问题，也是一个比较重要的研究方向。Hinton的capsule能否获得比传统CNN更好的效果，也需要进行进一步的研究。

参考文献：

[1] Marr D.Vision：A computational Investigation into the human representation and processing of visual information[M].[S.l.]：W H Freeman and Company，1982.

[2] Krizhevsky A，Sutskever I，Hinton G E.ImageNet classification with deep convolutional neural networks[J].Communications of the ACM，2012，60(2).

[3] Girshick R，Donahue J，Darrell T，et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition，2013：580-587.

[4] Lecun Y，Boser B，Denker J S，et al.Backpropagation applied to handwritten zip code recognition[J].Neural Computation，1989，1(4)：541-551.

[5] Freund Y，Schapire R E.Experiments with a new boosting algorithm[C]//Proceedings of Thirteenth International Conference on International Conference on Machine Learning，1996：148-156.

[6] Dalal N，Triggs B.Histograms of oriented gradients for human detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition，2005：886-893.

[7] Vapnik V N.The nature of statistical learning theory[J].Technometrics，1997，8(6)：1564.

[8] Lowe D G.Distinctive image features from scale-invariant keypoints[J].International Journal of Computer Vision，2004，60(2)：91-110.

[9] Ke Y，Sukthankar R.PCA-SIFT：A more distinctive representation for local image descriptors[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition，2004：506-513.

[10] Bay H，Tuytelaars T，Gool L V.SURF：Speeded up robust

features[J].Computer Vision & Image Understanding, 2006,110(3):404-417.

[11] Tola E,Lepetit V,Fua P.DAISY:An efficient dense descriptor applied to wide-baseline stereo[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010,32(5):815-830.

[12] Morel J M,Yu G.ASIFT:A new framework for fully affine invariant image comparison[J].SIAM Journal on Imaging Sciences,2009,2(2):438-469.

[13] Rosten E,Drummond T.Machine learning for high-speed corner detection[C]//Proceedings of European Conference on Computer Vision,2006:430-443.

[14] Calonder M,Lepetit V,Strecha C,et al.BRIEF:Binary robust independent elementary features[C]//Proceedings of European Conference on Computer Vision,2010: 778-792.

[15] Rublee E,Rabaud V,Konolige K,et al.ORB:An efficient alternative to SIFT or SURF[C]//Proceedings of IEEE International Conference on Computer Vision,2012: 2564-2571.

[16] Schapire R E.The strength of weak learn ability[C]// Proceedings of the Second Annual Workshop on Computational Learning Theory,1989:197-227.

[17] Viola P,Jones M.Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2003:511-518.

[18] Viola P,Jones M J.Robust real-time face detection[J]. International Journal of Computer Vision,2004,57(2): 137-154.

[19] Lienhart R,Maydt J.An extended set of Haar-like features for rapid object detection[C]//Proceedings of International Conference on Image Processing,2002:900-903.

[20] Hosang J,Benenson R,Dollar P,et al.What makes for effective detection proposals?[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2016,38(4):814.

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

[22] Sermanet P,Eigen D,Zhang X,et al.OverFeat:Integrated recognition,localization and detection using convolutional networks[J].Eprint Arxiv,2013.

[23] He K,Zhang X,Ren S,et al.Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,37(9):1904-1916.

[24] Girshick R.Fast R-CNN[C]//Proceedings of ICCV 2015, 2015.

[25] Ren S,He K,Girshick R,et al.Faster R-CNN:Towards real-time object detection with region proposal networks[C]//Proceedings of International Conference on Neural Information Processing Systems,2015:91-99.

[26] Dai J,Li Y,He K,et al.R-FCN:Object detection via region-based fully convolutional networks[C]//Proceedings of NIPS 2016,2016.

[27] He K,Zhang X,Ren S,et al.Deep residual learning for image recognition[C]//Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[28] He K,Gkioxari G,Dollár P,et al.Mask R-CNN[C]// Proceedings of ICCV 2017,2017.

[29] Xie S,Girshick R,Dollar P,et al.Aggregated residual transformations for deep neural networks[C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,2016:5987-5995.

[30] Kong T,Yao A,Chen Y,et al.HyperNet:Towards accurate region proposal generation and joint object detection[C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,2016: 845-853.

[31] Wang X,Shrivastava A,Gupta A.A-Fast-RCNN:Hard positive generation via adversary for object detection[C]// Proceedings of CVPR 2017,2017.

[32] Goodfellow I J,Pouget-Abadie J,Mirza M,et al.Generative adversarial networks[J].Advances in Neural Information Processing Systems,2014,3:2672-2680.

[33] Li Zeming,Peng Chao,Yu Gang,et al.Light-Head R-CNN: In defense of two-stage object detector[C]//Proceedings of CVPR 2017,2017.

[34] Redmon J,Divvala S,Girshick R,et al.You only look once:Unified,real-time object detectiono[C]//Proceedings of CVPR 2015,2015:779-788.

[35] Liu W,Anguelov D,Erhan D,et al.SSD:Single shot multibox detector[C]//Proceedings of European Conference on Computer Vision,2016:21-37.

[36] Redmon J,Farhadi A.YOLO9000:Better,faster,stronger[C]// Proceedings of CVPR 2016,2016.

[37] Shen Z,Liu Z,Li J,et al.DSOD:Learning deeply supervised object detectors from scratch[C]//Proceedings of CVPR 2017,2017:1937-1945.

[38] Huang G,Liu Z,Maaten L V D,et al.Densely connected convolutional networks[C]//Proceedings of CVPR 2016, 2016.

[39] Jeong J,Park H,Kwak N.Enhancement of SSD by concatenating feature maps for object detection[C]//Proceedings of CVPR 2017,2017.

[40] Gidaris S,Komodakis N.LocNet:Improving localization accuracy for object detection[C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,2016:789-798.

(下转33页)