卷积神经网络在目标检测中的应用综述

于进勇[1] 丁鹏程[2] 王 超[1]
(海军航空大学控制工程系 山东 烟台 264001)[1] (海军航空大学研究生五队 山东 烟台 264001)[2]

摘 要 深度学习作为机器学习的一个分支,在各个领域的应用越来越广,已经成为语音识别、自然语言处理、信息检索等方面的一个主要发展方向;其在图像分类、目标检测等方面更是不断取得新的突破。文中首先梳理了卷积神经网络在目标检测中的典型应用;其次,对几种典型卷积神经网络的结构进行了对比,并总结了各自的优缺点;最后,讨论了深度学习现阶段存在的问题以及未来的发展方向。

关键词 计算机视觉,目标检测,深度学习,卷积神经网络
中图法分类号 TP751 **文献标识码** A

Overview:Application of Convolution Neural Network in Object Detection

YU Jin-yong[1] DING Peng-cheng[2] WANG Chao[1]
(Department of Control Engineering,Naval Aeronautical University,Yantai,Shandong 264001,China)[1]
(Postgraduate Team No. 5,Naval Aeronautical University,Yantai,Shandong 264001,China)[2]

Abstract As a branch of machine learning,deep learning has obtained wide application in various fields,and has become a major development direction of speech recognition,natural language processing,information retrieval and other aspects. Especially in image classification and object detection,it has made new breakthroughs. This paper first sorted out the typical applications of convolution neural network in object detection. Secondly,this paper compared several typical convolutional neural network structures,and summed up their advantages and disadvantages. Finally,the existing problems and the future development direction of deep learning were discussed.

Keywords Computer vision,Object detection,Deep learning,Convolutional neural networks

1 引言

Gartner 是用来评估新科技可见度的一种工具,也是技术企业投资决策的重要风向标。图 1 所示为 2017 年发布的新兴科技技术成熟度曲线(Hype Cycle for Emerging Technologies),其中包含了各类新兴科技技术未来五到十年内可以帮助各企业在数字经济时代中生存并繁荣发展的情况。

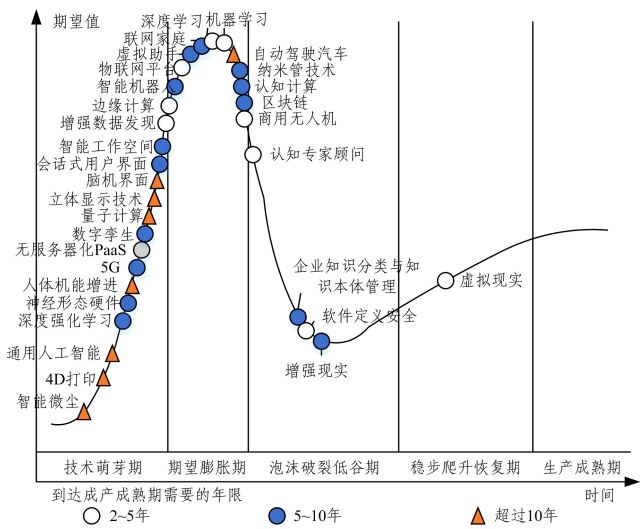

图 1 2017 年新兴科技技术成熟度曲线

从图 1 中可以看到,深度学习是目前最被看好的新兴科技技术之一。如今,深度学习迅猛发展,在语音识别、计算机视觉、自然语言处理以及信息检索方面不断取得新突破。归结其原因,主要是因为芯片处理能力的显著提高(如 GPU 单元加速)和深度学习算法的长足进步[1],使得深度学习在理论和应用上不断取得突破。

随着计算机数据资源和硬件的不断完善发展,各大企业和科研院所对深度学习的研究发展越来越重视。2013 年 1 月,互联网搜索引擎公司百度的年会上,百度成立了第一个研究院,其重点方向就是深度学习[2]。此外,微软公司有邓力团队,Google 公司有 Geoffrey Hinton 团队等[3]。高校团队中有多伦多大学的 Geoffrey Hinton 研究组、加拿大蒙特利尔大学的 YoshuaBengio 研究组、纽约大学的 YannLeCun 研究组[3]等。如今,卷积神经网络作为深度学习的重要分支,被广泛应用在自然语言处理[4-7]、医药发现[8]、灾难气候发现[9]以及人工智能程序中[10],但是应用最为成熟的还是在计算机视觉方面,特别是其近年来在目标检测任务中的应用研究得到了迅猛发展[11-13]。

如图 2 所示,目标检测是一个复杂的过程,不仅需要分辨出物体,还需要用边界框从背景中圈出它在图像中的具体位置,所以图像分类和图像定位成了目标检测评判的重要标准。此外,在现实生活中往往需要对多目标进行检测,很容易出现多目标遮挡、自身非刚体形变等问题,这对检测的精度提出了更高的要求。近年来,随着深度学习技术的发展,卷积神经网络在精度上显著优于传统方法,成为了最新的研究热点[14],

于进勇(1976-),男,博士,副教授,主要研究方向为深度学习、飞行器智能控制等;**丁鹏程**(1994-),男,硕士生,主要研究方向为深度学习,E-mail:632875352@qq.com(通信作者);**王 超**(1988-),男,硕士,讲师,主要研究方向为智能算法、飞行器控制与制导。

并细化出基于候选区域和基于回归方法两个分支。前者注重准确率,性能更好;后者是在适当降低准确率的情况下获得速度的最大收益。

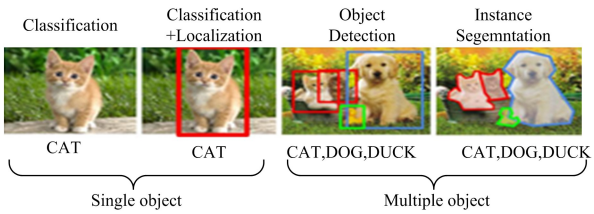

图 2 目标检测的组成

2 卷积神经网络在目标检测中的研究现状

2.1 传统目标检测的不足

传统目标检测方法一般分为 3 个阶段:首先,采用滑动窗口框架的方法,利用不同尺寸的滑动窗口框在给定图像的不同位置上选取候选区域;其次,对这些候选区域进行特征提取;最后,利用分类器进行识别。针对不同类别的物体,需要设计不同的特征和分类方法。比如,选用经典的 Harr[15] 特征和 Adaboosting[16] 分类器,利用滑动窗口搜索策略进行人脸检测;将梯度方向直方图(Histogram of Gradients,HOG)[17] 提取的特征经过支持向量机(Support Vector Machine,SVM)[18-19] 进行行人检测;针对一般性的物体检测,则采用 HOG 的特征加多尺度形变部件模型(Deformable Part Model,DPM)[20] 算法。特别是 DPM 算法,在传统目标检测中连续获得 VOC(Visual Object Class)2007－2009 年的检测冠军,2010 年其作者 Felzenswalb Pedro 也因此被 VOC 授予"终身成就奖"。

但是,这些算法都需要人工从原始输入中获取有关的目标特征信息,伴随着诸多的局限:1)可移植性差。针对特定的检测任务,需要人工设计不同的方法,对于不同的目标或者同一目标的不同形态,对设计者的经验有很高的要求。2)特征提取和分类训练分离是传统检测模型的通病,如果在设计过程中,人工特征的提取出现漏提现象,漏提的有用信息将无法从分类训练中恢复,进而影响检测结果。3)传统方法多采用滑动窗口进行遍历搜索,把图片尽可能分成各种尺度和大小的图片小块,然后对图片小块进行识别,对概率大的部分进行保留,将概率小的进行合并删减。这种方法的复杂度高,且存在大量冗余小块,严重影响运行速度,在现实中也难以工程实现,所以自从深度学习 2013 年在目标检测领域兴起后,便迅速取代了传统算法的地位。

2.2 图像分割

深度神经网络在图像分类、目标检测等方面要想取得巨大突破,第一步就是预测图像上的每个像素点,这个任务就是图像分割[21]。通过图像分割,我们希望可以预测图片上的每个像素点属于哪个部分,这个作为计算机视觉应用的第一步,非常关键。图像分割常用的数据集有 PASCAL VOC2012[22],Microsoft COCO[23],PASCAL-CONTEXT[24],Sift Flow[25] 等。一些传统的方法,比如基于最优阈值的 Otsu 法[26]、基于边缘检测的平均比值法(Ratio Of Average,ROA)[27]、基于模糊聚类的模糊 C-均值(Fuzzy C-Means,FCM)聚类算法[28] 等,都取得了一定的成果。Long 等[29] 于 2015 年提出的全卷积网络(Fully Convolutional Networks,FCN)为图像分割开创了新的途径。该方法训练了一个端到端的网络,用卷积层代替传统网络中的内积层,可以接受任意尺寸的图像输入,对每一个像素都可进行语义预测。最后,使得在数据集 PASCAL VOC 上的结果比 2012 年的算法结果提高了约 20%。在 FCN 的基础上,Chen 等[30] 加入随机场算法(Conditional Random Fields)[31] 来对 FCN 模型进一步细粒度优化,改进了图像分割边界的效果,在 PASCAL VOC2012 上达到了 71.6%的 IOU(Intersection Over Union)精度。Noh 等[32] 用一个与 FCN 网络对称的多层反卷积网络代替简单的双线性插值,可以更好地反映物体细节,提升了分割效果。针对图像边缘信息分割精度的问题,Zheng 等[33] 提出将 CRF 当作递归神经网络(Recurrent Neural Network,RNN)嵌入到 FCN 模型的 CRF-RNN 网络,将 VOC2012 的平均 IOU 提高到 74.7%。针对小数据量样本的过拟合问题,全卷积网络的 DenseNet 模型可以在无需预训练的基础上达到所需精度,并将模型缩小到原模型的 1/10,在医学图像、卫星图像等任务上具有广泛的应用前景[34]。

2.3 图像分类

图像分类是计算机视觉应用最为广泛的领域[35-38],相关的数据集有很多,比如 CIDAR-10/100[39],Caltech-101/256[40-41] 和 ImageNet[42]。Krizhevsky 等首次将卷积神经网络(Convolutional Neural Networks,CNN)应用于 ImageNet 大规模视觉挑战赛(ImageNet large scale visual recognition challenge,ILSVRC)[35],凭借 AlexNet 的 2-GPU 并行结构,使其训练的深度卷积神经网络在 2012 年比赛中荣获冠军,且 top5 错误率直降到 16.4%,使用额外数据更是达到 15.3%。而在 2010 年和 2011 年的最佳 top5 错误率还分别为 28.2% 和 25.8%。在 2013 年的比赛中,排名前 20 的算法全部都使用了深度学习,而 2013 年之后,参赛 ILSVRC 的队伍基本上全使用了深度学习算法。

如图 3 所示,一直到 2015 年,深度学习在图像分类中的错误率基本都以 4%左右的速度在减小,从而表明深度学习打破了传统机器学习算法在图像分类上的瓶颈,让图像分类问题得到了很好的解决。2014 年,通过用全局平局池化层替代全连接结构和采用 Inception Moudle 结构提高了参数利用率的 GoogleLeNet[43] 首次出现在 ILSVRC 2014 的比赛中,并以巨大优势夺冠。同年,牛津大学计算机视觉组(Visual Geometry Group)和 Google DeepMind 公司一起研发的深度卷积神经网络 VGGNet[44],全部使用了 3×3 的卷积核和 2×2 的池化核,将多个 3×3 卷积核以堆叠的方式呈现,使参数缩减近一半,又可让 3×3 的卷积核多次使用 ReLU 激活函数,使得 CNN 特征学习能力大大提升。2015 年,来自微软亚太研究院 residual networks 的 ResNet[45] 采用残差网络结构将错误率降至3.57%,而在 ImageNet 数据集上人眼能达到的错误率大概在 5.1%,这还是经过了大量训练的专家才能达到的成绩,一般人要区分 1000 种类型的图片还是比较困难的。2017 年,获得 CVPR(IEEE Conference on Computer Vision and Pattern Recognition)的最佳论文的 DenseNet[46] 针对深层网络梯度消失问题,尽量缩短层与层之间的连接,每层只学习较少的特征,同时确保每两层之间都有直接连接,实现特征的重用,提高网络计算的效率。通过融合 ResNet 和 DenseNet 的思想特点,2017 年 ImageNet 提出了双通道网络(Dual Path Network,DPN)模型[47],其在适当提高准确率的基础上,重点优化了模型大小和计算复杂度。在 ImageNet-1k 分类任务中,相比之前最好的 ResNet-101(64×4D),DPN 模型的规模缩减了 26%,计算量和内存消耗分别降低了 25% 和 8%,设计的 DPN-131 更是将训练速度提升了 2 倍。

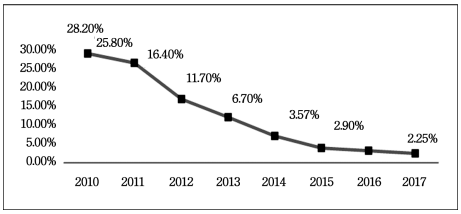

图 3　历年 ILSVRC 图像分类比赛中最佳算法的 top5 错误率

2.4　基于候选区域的目标检测

除了 ImageNet[42] 数据集外，PASCAL VOC[48]，SUN[49] 和 MS COCO 数据集[23] 等也是目标检测常用的公共数据集。

基于候选区域特征提取的深度学习，常用产生候选区域的方法有 Selective Search[50] 和 Edge Boxes[51]。特征提取方面，目标检测普遍采用卷积神经网络替代人工特征提取[52]，常用到的特征提取网络模型有 AlexNet[35]，GoogleNet[43]，VGG[44]，ResNet[45] 等。如 2013 年 ILSVRC 分类加定位得主 OverFeat[53] 就是在 AlexNet 的基础上，取得了 24.3% 的 mAP(mean Average Precision)，略高于传统算法的 22.581% mAP。而在 ILSVRC 2014 中，首次引入深度学习方法的 R-CNN[54] 就将 mAP 提高了大约一倍，达到了 43.933%。这是一种基于 AlexNet[35] 网络提取区域候选框的特征，利用 CNN 的正向传播，将最后一层得到的特征提取出来的网络。在R-CNN 的基础上，又出现了两种改进方案，即 Fast R-CNN[55] 和 DeepID-Net[56]，但这两种模型计算候选区域的时间成为整个模型运行时间的瓶颈。在提出设计一个全卷积网络的候选区域网后[29]，Ren 等[57] 提出了 Faster R-CNN 模型，其在提高候选区域质量的情况下，采用 RPN(Region Proposal Networks)网络计算候选框，提高了准确率，且将检测时间缩至原来的 1/10。

针对目标在变形、遮挡、逆光等难以区别情况下的样本，Shrivastava 等[58] 利用 Bootstrap[59] 的方法提出了困难样本在线挖掘(Online Hard Example Mining，OHEM)，在 Fast R-CNN 上加入 Hard ROI Sampler，对输入损失进行排序和误差回传，重新训练分类器，提高了网络的检测精度。

Yang 等在 SDP-CRC[60] 中提出了两个新的策略。一个是针对不同尺度的目标处理，采用基于不同候选区域的 Scale Department Pooling(SDP)，让小尺度目标在较前的网络层池化，大尺度目标在较后的网络层池化，从而得到全图的池化特征。另一个策略是采用 Cascaded Rejection Classifer(CRC)，只保留包含目标可能较大的候选区域，以提高候选区域的计算效率。

Bell 等提出的 ION[61] 充分利用多尺度信息提取的内部信息和包含上下文的外部信息。其连接不同卷积层预测多尺度特征，可以提高小目标的检测精度；使用空间递归神经网络(Spatial Recurrnet Neural Network)[62]，可以充分利用视觉识别中的上下文信息，对遮挡目标有较好的检测效果。

He 等[63] 重新对齐 RoI(Region of Interesting)Pool，保证特征区域可以更精准地对应原始图像，设计了具有 RoIAlign 结构的 Mask RCNN 网络。此外，网络在 Faster R-CNN 的基卷积网络特征顶部加入全卷积分割子网，并行输出分割和检测结果，将原来的 2 个任务(分类＋回归)扩展成了 3 个任务(分割＋分类＋回归)，并以此获得了 OCO2016 的冠军。

2017 年，Lin 等[64] 在 CVPR 上提出的 FPN 使用多尺度、多层级特征金字塔网络自上而下横向融合多层网络特征，建立了高级语义特征图。在 COCO 数据上，使用了 FPN 的 Faster R-CNN 创造了单模型目标检测的冠军。这个会议上，还第一次将生成式对抗网络(Generative Adversarial Networks，GAN)[65] 应用到目标检测，提出了 Perceptual Generative Adversarial Networks(PGAN)[66]，其通过建立单个架构，将小 RoI 转换为大 RoI，缩小小目标与大目标的表示差异，提升了小目标的检测能力。

2.5　基于回归方法的目标检测

基于特征候选区域的目标检测分为两步，可以充分提取图像特征，实现精确检测。但是，这种结构的网络需要训练两个卷积，网络结构较为复杂，并且训练时间较长，更适合学术研究而非工业实践。网络结构更加简单、实时性更高的一步回归方法被应用在目标检测上。

Redmon 等[67] 提出只用一个神经网络进行目标检测的 YOLO 网络。YOLO 采用空间限制，直接求取整幅图像的边界框的置信度和类别的条件概率，可以更加“看清”图像局部信息，大大降低了背景的误检率。虽然准确率略有下降，但其实时性强，fast YOLO 更是达到了 155 帧/s。类似于 YOLO，Najibi 等[68] 综合了速度和精度，在 Fast R-CNN 的基础上把固定网络的检测框渐变为真实边界框，经过 5 次迭代得到的 G-CNN 网络取得了与 Fast R-CNN 相当的准确率，同时速度提高了 5 倍。随后提出的 SSD[69] 网络采用 6 个模块进行特征提取，结合 Hard Negative Mining 策略，其在速度和精度上都达到了当时的最高水平，并且得到了广泛的应用和推广。

针对回归方法目标检测准确率和召回率不高的问题，Redmon 等[70] 用批正则化、维度聚类、直接位置预测等方法对 YOLO 进行框架改进，提出了 YOLO-v2。在 VOC2007 数据集上，测试速度为 67 帧/s 时，mAP 为 76.8%，测试速度为 40 帧/s 时，mAP 为 78.6%，效果显著。同时，该作者还提出了一种联合训练分类和检测的方法，使 YOLO9000 可以同时在 COCO 和 ImageNet 上进行训练，并达到 9000 种目标的实时检测。

与 ION 相似，Ren 等[71] 也将上下文信息充分应用在网络中，结合循环神经网络(Recurrent Neural Networks，RNN)[72]，提出了 RRC 网络。通过在 SSD 网络里加入上下文信息，网络可以同时对大目标和小目标进行检测，在较高的 IOU 阈值下有着较高的性能，获得了 KITTI car 数据集困难样本检测的冠军。

2.6　视频目标检测

视频目标检测需要充分运用上下文信息，一些优秀的图像目标检测网络如 YOLO 和 SSD 等虽然在比较简单的视频中可以达到较好的效果，但是对复杂视频的处理结果不尽人意。

为了更好地对大规模视频分类进行经验评估，Karpathy 等[73] 融合单帧、不相邻两帧、相邻多帧以及多阶段相邻多帧的方法，把 Sports-1M 数据集的 100 万段 YouTube 视频数据用于卷积神经网络的训练，并提出了多分辨率的网络结构。相比于传统人工提取特征的方法，该模型在 Sports-1M 上的分类准确率从 55.3% 提升到 63.9%。Ji 等[74] 在空间和时序上运用三维卷积提取特征，将得到的多个相邻帧的运动信息用于行为识别，提出了 3D CNN 网络。该模型基于输入帧，生成多个特征图通道，将所有通道的信息结合，从而获得最后的特征表示。Baccouche 等[75] 提出了一种时序的深度学习模型，其可在没有任何先验知识的前提下学习分类人体行为。模型的第一步是将卷积神经网络拓展到三维，自动学习时空特征；然后使用 RNN 方法训练分类每个序列。该模型在 KTH 上的测试结果优于其他已知深度模型，KTH1 和 KTH2 上的精度分别为 94.39% 和 92.17%。

ImageNet 视频目标检测任务是 2015 年引入的尝试赛，类似于从静态图像中进行目标检测。该任务的数据集包括 30 个类别，这些类别是目标检测任务中 200 个类别的子集。T-CNN[76]从视频中获得的 tubelets 中集成了时域和上下文信息，显著改善了视频中目标检测的性能，赢得了 ILSVRC 2015 视频目标检测比赛的冠军。

Zhu 等[77]针对目标检测效率受限的问题，在实时视频处理过程中结合光流法[78]的思路，保证特征在帧与帧之间传播和复用。视频中并不是每一帧都适用于特征提取，但是每帧内容具有高度的相关性。网络设计了两个子网络，一个只在关键帧进行 CNN 特征的提取，并通过流场传播给其他帧；另一个根据不同的任务，设计不同的结构。

针对如何让机器从视频中找出感兴趣部分的起止时间点的问题，Shou 等[79]提出了卷积-逆卷积神经网络(Convolutional-De-Convolutional Networks，CDC)，其大大提升了对视频时序边界的定位精度。该网络的最大贡献是设计了一个可以同时进行时序增采样和空间降采样的 CDC 过滤器，使网络同时在时空级和粒度级的时序动态中推断高级动作语义并进行精度定位。在单一 GPU 下，CDC 网络视频的处理速度达到 500 帧/s。

3 几种经典的卷积神经网络结构

3.1 目标检测的基础

3.1.1 AlexNet[35]

我们常说的经典卷积神经网络，一般都认为从 AlexNet 开始。图像分类方面，ZFNet[80]，GoogleNet[43]和 VGGNet[44]都是基于 AlexNet 的改进，而目标检测的关键性突破 R-CNN[54]也是 AlexNet 的变形。2012 年，Hinon 的学生 Alex Krizhevsky[35]提出了深度卷积神经网络模型 AlexNet，并参加了当年的 ImageNet，获得了图像分类组的最好成绩，top-5 为 16.4%，大幅度降低了图像分类的错误率。与最早的深层卷积神经网络之一的 LeNet5[81]相比，AlexNet 增加了许多新的特征：1)包含更多的隐含层，它包含了 5 个卷积层和 3 个全连接层，并有 3 个卷积层后面连接了最大池化层。2)首次在 CNN 中采用 Dropout 解决过拟合问题。3)采用重叠的最大池化代替传统的平均池化，避免了平均池化的模糊化效果，又通过池化层之间的重叠覆盖效果提升了特征的丰富性。4)首次在 CNN 中用 ReLU 激活函数代替之前传统的 Sigmoid 或者 tanh 等，既可以加快收敛，大幅缩短训练时长，又可以成功解决梯度弥散问题。如今，除了神经网络的输出层一般还采用比较接近概率输出分布的 Sigmoid 函数，其他大部分隐含层都将 ReLU 作为主流。5)提出了 LRN 层，增强了反馈较大的数值，抑制了反馈较小的数值，增强了模型的泛化能力。6)如图 4 所示，采用双 GPU 两路训练结构，分别处理输入图像的上、下部分。

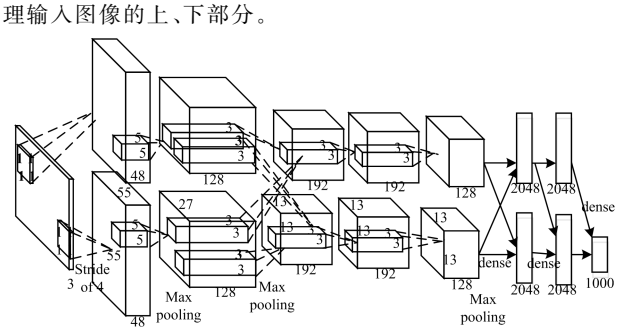

图 4 AlexNet 结构

3.1.2 OverFeat[53]

在 R-CNN[54]诞生之前，OverFeat 采用卷积神经网络 CNN，集分类、定位和检测于一身，并获得 2013 年分类加定位的冠军。这篇论文最大的特色是充分利用了卷积神经网络的特征提取，针对不同任务，无需从头训练整个网络，只需改变最后几层来处理提取的特征。具体来说，前 5 层和 AlexNet[35]相似，为卷积特征提取层；6 到 9 层针对分类和检测的任务分别设计，得出各个结果，进而达到预期结果。这种模型重用的思路也是后来 R-CNN 系列不断沿用和改进的经典做法。在图像分类上，采用卷积层代替全连接层，特别是第 6 层采用一个 5×5 的卷积核，以保证在测试过程中输入不同大小的图片时得到的不再是一个 1×1 的图片，而是一个与输入大小相关的预测值矩阵的图片。将每一列的最大值作为本尺度每个类别的概率值，最后将 12 个尺度的结果进行平均后作为最终结果。此外，在这个过程中还加入了 offset 池化处理，将得到的 9 种池化结果分别送入后面的网络层，得到 9 个预测结果，取其最大值作为每个类别的预测概率值。在定位任务中，把第 6 层到第 9 层设计成回归问题，在不同尺度上得到预测物体边框的 4 个角点坐标，将预测边界和真实边界之间的 L2 范数作为代价函数来训练网络。最后选用贪心算法进行合并，得到最高置信度的边框。

3.2 基于候选区域的目标检测

3.2.1 R-CNN[54]

Girshick 等主要致力于如何利用卷积神经网络进行目标定位和如何用小量检测数据来训练大容量模型的研究，提出了 R-CNN 网络模型，如图 5 所示。传统的目标检测主流方法是 DPM(Deformable Part Model)[82]，它是一种基于组件的检测算法。与 DPM 使用滑动窗口进行遍历搜索的方式相比，R-CNN 第一步采用选择性搜索[83]自底而上地生成候选区域，使用得分最高的 2000 个区域，每个候选区域用特定大小的 CNN 进行特征向量的提取，可以有效减少后面特征提取的计算量，能很好地应对尺度问题，在目标检测[50]和语义分割方面都取得了成功[84]。此外，CNN 在实现上采用 GPU 进行并行计算，计算效率明显优于 DPM 方法(实现上采用单 CPU 计算)。最后，利用线性 SVM 分类器、非极大值抑制和边界框回归策略使目标分类的准确性和定位的精确性得到进一步提升[85]。

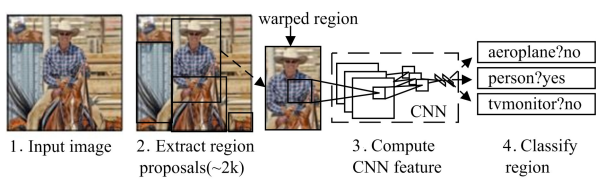

图 5 R-CNN 结构

针对第二个问题，先在大型辅助数据库(ILSVRC)上进行有监督的预训练，在小数据集(PASCAL)特定区域采用反向传播算法从特征层开始向后调整各层权重；接着，将特征层输出的高维特征向量和目标类别标签作为输入，训练支持向量机[85]。这种迁移学习是在数据不足的情况下，CNN 学习大量数据的有效手段。通过对检测进行微调，可以提高 mAP 8%的性能。这种微调的方法后来被深度学习广泛应用在目标检测的预处理上。

但这个网络同样存在很多问题：1)阶段繁多，从特征提取到微调卷积，再到线性 SVM 处理和边界框回归，连接性差；2)R-CNN 虽然不需要穷举，但是对 2000 个候选框分别进行

卷积操作，计算量大，且包含众多的重复计算；3）每次训练需要将候选框提取的特征保存到磁盘，空间开销大；4）将候选区域缩放到固定尺寸，导致出现物体不全或严重形变的情况，降低了检测的精度；5）基于上述情况，训练和测试时间开销大。在接下来的传统目标检测上，也主要针对上述问题进行发展与改进。

3.2.2　SPP Net[86]

由于存在全连接层，因此要求图片的输入尺寸必须一致。之前多采用裁剪和变形的方法来固定输入到全连接层图片的尺寸。但是，裁剪会导致物体不全，变形会导致图片拉伸后严重形变，从而影响检测精度。而在实际中，输入的图片往往大小不一，SPP Net 应用而生，它最大的特点是允许网络训练和测试不同大小的输入图片，且同时防止过拟合。SPP Net 引入了空间金字塔的概念，加入了空间金字塔池化（Spatial Pyramid Pooling，SPP）。这里，空间金字塔主要是为了用不同尺寸的方框来提取图片特征。如图 6 所示，用一个 4×4，2×2，1×1 大小的方框来分割特征图，可以得到 16＋4＋1＝21 种分割方式，将每个区域池化后可以得到 21 组特征，融合每块区域的特征，得到多个尺度特征的同时可以得到一个固定维度的输出，以满足全连接层的需要。这种组合池化的过程就是空间金字塔池化。

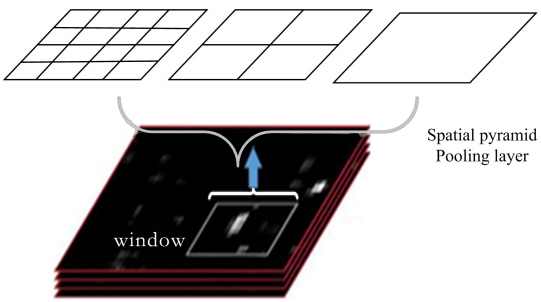

图 6　空间金字塔结构

通过加入 SPP 层，只需对图像进行一次卷积特征提取就可得到特征图，相比 R-CNN 每次需要 2000 个左右候选区域进行卷积提取，速度提高了 24～102 倍，精度也有一定提高。在训练时，使用分类任务得到的网络，只需微调全连接层，对于检测任务，与 R-CNN 一样，使用 SVM 和边界框回归，SVM 输入为全连接层，边界框回归使用 SPP 层。

SPP Net 对 R-CNN 最大的改进就在于 SPP 层的加入，所以训练也要经过多个阶段，特征也要存在磁盘中，离端到端的检测还差很多。特征提取是在 CPU 中进行，相对于 GPU 来说速度较慢。此外，网络微调只更新全连接层，检测精度提升不明显，不适合深层的网络结构。

3.2.3　Fast R-CNN[55]

相比于 R-CNN 和 SPP Net，Fast R-CNN 模型的主要亮点在于训练过程采用鲁棒性 L1 的多任务损失函数，实现单步骤完成，且不需要 SVM 训练分类器；相比于 SPP Net 只能更新全卷积层，Fast R-CNN 在一个批次里可以通过反向传播神经元偏导之和更新所有层参数，进而使 mAP 比 R-CNN 有很大的提升；不再需要磁盘作为特征缓存，比 R-CNN 训练速度快 9 倍，测试速度快 213 倍，与 SPP Net 相比，训练速度快 3 倍，测试速度快 10 倍。

从图 7 可以看到，Girshick 的具体做法是首先将图片单输入改为双输入，分别输入图片集合和 RoI（Region of Interesting）集合；接着通过一系列卷积层后，采用一个简化的 SPP 层——RoI Pooling 层代替最后一个最大池化层，这个网络层可以把不同大小的输入映射到一个固定尺度的特征向量；最后将 RoI Pooling 层得到的特征通过两个全卷积层后，将 SVM 分类和边界回归纳入卷积神经网络中组成多任务模型。

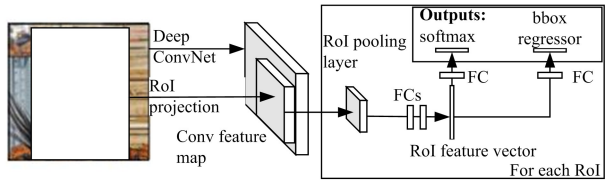

图 7　Fast R-CNN 结构

3.2.4　Faster R-CNN[57]

Fast R-CNN 使用选择性搜索提取候选特征，但是这部分只能在 CPU 上运行，存在速度瓶颈，难以达到实时。鉴于此，使用 RPN（Region Proposal Networks）网络计算候选框的方法被提出。Faster R-CNN 引入 RPN 网络，将区域提取、分类、回归共用卷积特征，这两点也是 Faster R-CNN 最大的特点。

RPN 采用全卷积层来代替之前的全连接层，可以使网络不受输入图片大小的限制，用卷积在得到的特征层上利用 k 个不同的矩形框进行区域提取，同时可以输出所有矩形框的特征。这个网络主要包含两个分支，即分类层和回归层，如图 8 所示。分类层用于判断该区域为前景还是后景，回归层预测区域的中心以及对应的坐标$[x,y,w,h]$，对应的损失函数分别为 softmax loss 和 L1 loss。

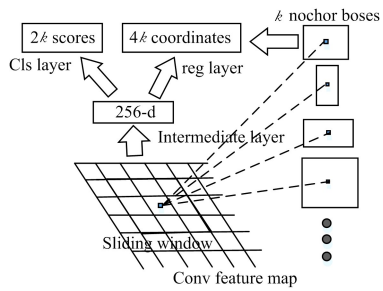

图 8　RPN 结构

为了更好地将全卷积网络的 RPN 与 Fast R-CNN 相结合，Ren 交替训练，实现特征共享。1）用 ImageNet 预训练模型初始化 RPN，端对端地对候选区域提取进行微调；2）将 RPN 得到的候选区域作为输入，独立训练检测网络 Fast R-CNN；3）固定共享卷积层，用检测网络初始化 RPN，并只微调 RPN 部分，开始进行卷积层共享；4）继续固定共享卷积层，只微调 Fast R-CNN 部分。两个网络共享卷积层，实现 RPN 和 Fast R-CNN 网络的统一，使得网络快速收敛，大幅提高网络速度。

3.2.5　R-FCN[87]

目标检测既要分类，又要检测，但是这两者之间存在矛盾。ImageNet 分类结果明显优于检测，主要是分类可以充分利用平移不变性的特点提高准确率。检测需要辨出物体在图中的位置，必须要考虑平移变化。随着网络深度的增加，卷积神经网络对位置的敏感度越来越低，对目标检测的准确率产生了显著的影响。针对上述问题，R-FCN 使用全卷积网络，实现计算的全程共享，并提出位敏得分图（position-sensitive score maps）来解决矛盾。

R-FCN 算法在 ResNet-101[45] 的基础上去掉最后的全局平局池化层和全连接层，保留前 100 层进行卷积，经 ImageNet 预训练[88]得到一个 2048 维的特征图。为了降低输出维度，附加一个随机初始化的 1024 维的 1×1 卷积层，将得到感

兴趣的特征响应图像进行位敏得分。

如图 9 所示，卷积后的图像把目标分成 $k\times k$ 个部分，并映射到一张位敏得分图上，每个得分图对应目标一小块，得到 $k^2(C+1)$(C 类目标加一个背景类)个输出通道。经过位敏 RoI 池化层后，得到 $C+1$ 个类别的置信度。最后将 softmax 用于类别判定，同时生成一个 $4\times k\times k$ 个回归边界框位置的地图，平均投票后得到边界参数。最终在数据集 VOC07 和 VOC12 上以 170 ms/张的速度分别得到 83.6% 和 82% 的 mAP。

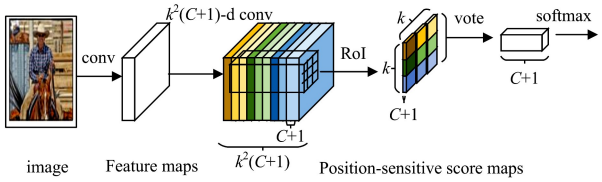

图 9 R-FCN 结构

3.3 基于回归方法的目标检测

3.3.1 YOLO[67]

先进行选择性搜索或者 RPN 网络生成候选区域，再进行分类和回归操作，决定着速度总会有一个上限。为了更好地实现实时性，直接从原始图片到物体位置和类别的 YOLO 被提了出来。它通过设计一个独立的网络，将目标检测作为一个回归问题，直接从像素求得边界框和置信度，正如网络名字一样，You Only Look Once。

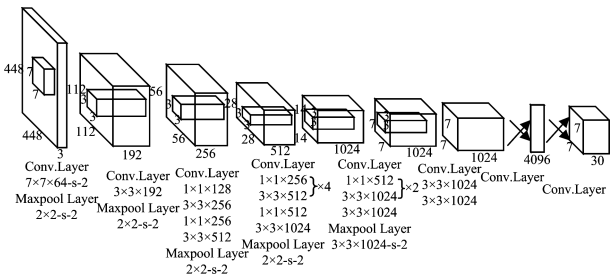

图 10 YOLO 结构

如图 10 所示，YOLO 的结构取自 GoogLeNet[43] 和 Network in Network[89]，共包含 24 个卷积层和 2 个全连接层，在 GoogLeNet 分类网络的基础上，用一个 1×1 缩小和 3×3 卷积层还原来代替 Inception 结构。YOLO 的卷积层用于特征提取，全连接层用于预测图像位置和类别概率。此外，为了增强模型的效果，网络将输入图像的维度从 224×224 扩展到 448×448，同时将边界框的位置信息和置信度也都进行了归一化处理。

通过端对端的一步处理，检测速度得到了极大的提高，在 VOC 2007 测试数据集上可以达到 45 帧/s，设计的小型 Fast YOLO 更是达到 155 帧/s，实时视频的滞后性只有 25ms，第一次让目标检测达到了实时性。在精度上，YOLO 在训练和评估过程中输入全图信息，达到其他实时处理算法的 2 倍，可以很好地避免如 R-CNN 或 Fast R-CNN 等易将局部背景误检成物体的情况。最后，只有一个端对端 CNN 模型，容易优化，通用性强，在非自然图像中的物体检测率远高于其他算法。

但是，作为第一个端对端物体检测网络，网络速度的提升是以牺牲精度为代价的，特别是小物体的检测，损失函数中小 IOU(Intersection Over Union)的误差影响明显，降低了定位准确度。另外，因为只能预测同一区域的一个物体，当物体在图像中占比较小或一个区域包含群体物体时，只能检测出其中一个物体，召回率低。

3.3.2 SSD[69]

针对 YOLO 在整个特征图中只用 7×7 网格对目标回归导致精度不足的问题，SSD 综合了 YOLO 和 Faster R-CNN，以特征分层提取作为主体设计思路进行回归和分类。低层次特征图的语义分割质量高，可以包含更多特征的细节，适合学习小尺度目标；高尺度特征语义分割更加光滑，适合学习大尺度目标。结构上，SSD 在单一回归网络的基础上，针对全图各位置，使用多尺度区域特征进行目标回归，既保证了 YOLO 端对端快速的特点，又保留了 Faster R-CNN 对小目标精确的特性，在 VOC2007 达到了 72.1% mAP，速度达到了 58 帧/s，获得当时目标检测比赛的冠军。

如图 11 所示，SSD 在是 VGG16[44] 的基础上，设计了 6 个模块进行特征提取。VGG16 前 5 层卷积为第一模块，用两个卷积层分别代替 VGG16 的两个全连接层作为第二模块，之后新加 4 层卷积网络作为后面的 4 个模块，进行更高层次语义信息的提取。为了避免生成过多负样本边界框而影响检测的准确率，网络在非极大值抑制整合高度重叠边界框的基础上，使用 Hard Negative Mining 策略，根据置信损失进行排序，只训练最高训练损失的负样本子集，保证正负样本比例为 1∶3。

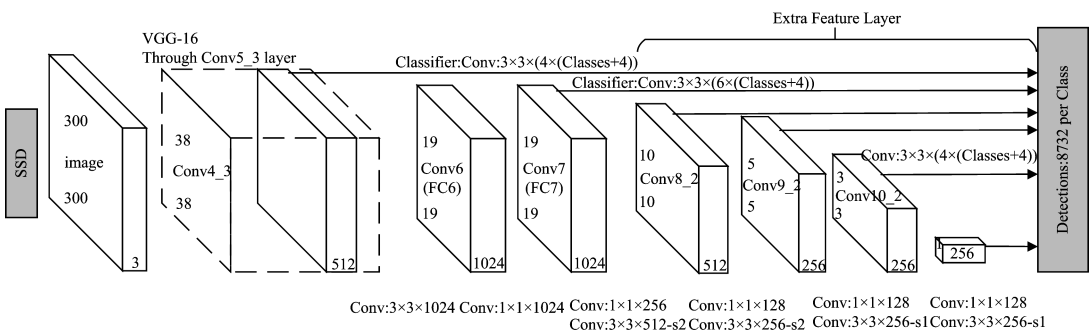

图 11 SSD 结构

3.4 视频目标检测

图像目标检测标准 mAP 依据每个检测窗口对目标标定是否精准来判定。如果检测窗口判断类别结果与真实目标相同，且给定窗口的重叠面积大于 0.5，则认定其为正例。而视频目标检测主要评判模型的时序一致性，即目标跟踪轨迹是否精准。如果检测得到的跟踪轨迹与真实目标给定轨迹同类且重叠面积大于 0.5 的数量超过一定比例，则认定其为正例。最后，求取序列所有窗口得分的平均值作为最终轨迹跟踪得分。

针对时序一致性问题，比较经典的解决方案是 Kang

等[76]提出的 T-CNN（Tubelets with Convolutional Neural Networks）网络，其在单帧图像检测的基础上，应用了多上下文抑制（MCS）和运动指导传播（MGP）算法。上下文抑制用来降低误检，即一个视频段中只可能出现几个类别的目标，且这几个目标之间有共现关系。通过对视频段上的检测结果进行分析，对所有检测窗口重新进行排序，选出得分较高的类别，剩余那些得分较低的类别可认为是误检，对其得分进行抑制，从而使高置信度的目标置信度更大，低置信度的目标置信度更低。运动指导传播针对单帧检测结果容易出现漏检的情况，借助光流信息将当前帧的检测位置和置信度传给邻近帧，提高目标的召回率。最后，加入目标跟踪算法，以实现视频目标的高识别率和跟踪率。

4　深度学习的问题与发展方向

近年来，随着深度学习的飞速发展，目标检测取得了丰硕的成果。可以预见到，基于卷积神经网络等深度学习方法的研究仍将是目标检测的重要方向。这其中有众多方面值得一起努力，也有一些问题等待科研工作者来解决。

4.1　多样性数据集不足

深度学习是一种专项的算法，针对不同的工程需要不同的训练数据集。从 2016 年至今的趋势看来，对于有监督的深度学习算法，每个类别只要有 5000 个样本，机器就能达到令人满意的表现；若有 1000 万以上的样本，机器就能达到甚至超越人类的水平。但是，针对许多特定问题，人工信息采集和标注任务量繁重，特别是针对一些小概率的困难样本的收集，需要人工进行信息处理和变形，这一定程度上限制了深度学习的大量应用。同时，当数据集过小时，很容易出现过拟合或者欠拟合的现象，从而影响模型的使用。因此，如何提高小样本的学习率是未来深度学习发展的一项重要研究内容。

此外，相比于目标分类的数据集，目标检测数据集的标注难度更大，特别是视频目标检测。将图像分类学习的知识尽可能多地应用在目标检测，或者将图像分类的数据集经过一定处理后直接作为目标检测的数据集，对目标检测数据集的多样性将有极大帮助。

4.2　深度学习理论有待完善

深度学习模型会向着结构更深、规模更大的方向发展。纵观这些年 ImageNet 大规模视觉识别挑战（ILSVRC）也能明显看到，参数越来越多，层数越来越多。探求有效降低计算复杂度的方法，实现参数与层数的完美融合，需要在理论和实验上不断改进。同时，深度学习的数学理论不够完善，现阶段模型的改进很大程度上依靠设计者的经验，理论基础落后于实践，不利于深度学习整体理论框架的提炼和升华。比如，现阶段有许多研究者针对多目标遮挡和小尺寸目标检测上效果较差的不足，提出了各种方法，这些方法各有利弊，但并没有从理论上对这两个问题进行合理的解释，无法形成规范统一的研究思路。

此外，如今的深度结构的层次模型虽然比浅度模型在结构上有了突破，但未能完全匹配类似生物皮层的信息处理结构。比如，现有的主流深度结构并未充分考虑到时间序列对学习的影响，而现实中，信息数据的学习往往不是静态的，既要能随着时间联系上下文，又需要针对不断扩充的在线数据集进行学习。必须加强深度学习的在线学习能力，在大数据的适应能力和层次结构上不断取得新突破，提出一些新的更加有效且容易做理论分析的算法模型，以更好地应用于计算机视觉、语音、自然语言等方面。

深度学习在理论和工程应用上也有很多不同。学术界可以不断追求目标检测精度的极限，比如 R-CNN 等后续的基于候选区域的检测方法可以一定程度上忽略算法对资源和时间的限制。但是，最后深度学习落实于工程实践，必须充分考虑资源的有效性和实时性，无论是放在云端上，还是嵌入手机或电脑等设备中，都必须要降低计算的复杂度，以保证网络的快速有效运行，才能真正让深度学习推动人类科学的不断进步。

4.3　非监督学习需要大力发展

传统的深度学习方法，重点在于监督学习。但是，在人类和动物的学习过程中，无监督学习占据主动地位。人类在进化过程中对世界内在结构的了解，是通过自己的观察和学习进行的，而不是被告知客观事物和规律。虽然近些年来，越来越多的学者将研究重点放在无监督学习上，且取得了一定的成果，但是其对特征的高效表达能力远不如监督学习算法。如何使机器像人与动物一样通过观察获取常识的无监督学习能力，将是未来深度学习的一个重要发展方向[90]。

4.4　结合机器学习和其他传统方法

结合近两年内 ICCV（IEEE Inernational Conference on Computer Vision），CVPR（IEEE Conference on Computer Vision and Pattern Recognition），ECCV（Europeon Conference on Computer Vision）三大计算机视觉顶级会议和其他论文可以看到，越来越多的研究者将机器学习和其他传统方法与深度学习算法融合应用到目标检测等领域，比如 2017 年融入 RNN 的 RRC 网络[71]和结合 DPM[82]的 Deformable CNN[91]网络对图像有着显著的操作效果。此外，迁移学习、集成学习、增强学习等方法也都在融合深度学习算法的过程中获得了突破，这也将是未来深度学习发展的一个显著趋势。

4.5　计算机硬件需要同步发展

由于训练过程中存在大量的矩阵乘法、并行处理等，深度学习训练需要很大的计算量，计算机硬件性能的提升成了深度学习发展推广的基本条件之一。以著名的 GPU 厂商英伟达（Nvidia）为例，其近几年推出的 Tesla 和 Titan X 系列显卡处理器数量成百上千计，使得训练效率大幅提高。与同价位的 CPU 相比，其速度提升了约几十倍。2017 年战胜柯洁的 AlphaGo 更是大幅改进了 TPU，在减少 TPU 使用数量的同时仍可大幅提高计算效率。谷歌近年来宣称自己的 AutoML 方法可以自动找到最好的架构，但仍然需要超过 800 个 GPU 全天候运行数周。借鉴人脑的工作模式，开发出与神经网络算法相匹配的全新处理器架构，将是未来深度学习发展的一个新热点。

结束语　目标检测是深度学习一个非常热门的研究方向，利用卷积神经网络的卷积层、池化层等基本结构可以让网络自己学习和提取特征，并达到所需目的。这种特性可以省略很多复杂的建模过程，训练完毕后可以直接运用到实际物体。一方面，深度学习应用非常广泛，值得在工业、金融、医疗等各个领域广泛推广。另一方面，深度学习还属于起步阶段，值得不断地去探索和发展。有理由相信，未来深度学习一定可以在目标检测等人工智能领域中取得更大的突破。

参 考 文 献

[1]　LI H，ZHAO R，WANG X. Highly Efficient Forward and Back-

ward Propagation of Convolutional Neural Networks for Pixel-wiseClassification[J]. Computer Science, arXiv: 1412, 4526, 2014.

[2] 李彦宏. 2012 百度年会主题报告:相信技术的力量[R]. 北京:百度,2013.

[3] 张建明,詹智财,成科扬,等. 深度学习的研究与发展[J]. 江苏大学学报(自然科学版),2015,36(2):191-200.

[4] SHEN Y, HE X, GAO J, et al. Learning semantic representations using convolutional neural networks for web search[C]// International Conference on World Wide Web. ACM, 2014: 373-374.

[5] GREFENSTETTE E, BLUNSOM P, FREITAS N D, et al. A Deep Architecture for Semantic Parsing[J]. Computer Science, 2014, 30(5): 1-15.

[6] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[J]. arXiv: 1404. 2188, 2014.

[7] KIM Y. Convolutional Neural Networks for Sentence Classification[J]. arXiv: 1408. 5882, 2014.

[8] WALLACH I, DZAMBA M, HEIFETS A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery[J]. Mathematische Zeitschrift, 2015, 47(1): 34-46.

[9] LIU Y, RACAH E, PRABHAT, et al. Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets[J]. arXiv: 1605. 01156, 2016.

[10] CLARK C, STORKEY A. Teaching Deep Convolutional Neural Networks to Play Go[J]. arXiv: 1412. 3409, 2014: 1766-1774.

[11] FUHL W, SANTINI T, KASNECI G, et al. PupilNet: Convolutional Neural Networks for Robust Pupil Detection[J]. Revista De Odontologia Da Unesp, 2016, 19(1): 806-821.

[12] ZHANG X, ZOU J, HE K, et al. Accelerating Very Deep Convolutional Networks for Classification and Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(10): 1943.

[13] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Simultaneous Detection and Segmentation[M]// Computer Vision-ECCV 2014. Springer International Publishing, 2014: 297-312.

[14] 张慧,王坤峰,王飞跃. 深度学习在目标视觉检测中的应用进展与展望[J]. 自动化学报,2017,43(8):1289-1305.

[15] LIENHART R, MAYDT J. An extended set of Haar-like features for rapid object detection[C]// International Conference on Image Processing. IEEE, 2002: 900-903.

[16] 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(CVPR 2001). IEEE, 2003: 511-518.

[17] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR 2005). IEEE, 2005: 886-893.

[18] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[19] LIN C F, WANG S D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464.

[20] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. Computer, 2014, 47(2): 6-7.

[21] 卢宏涛,张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理,2016,31(1):1-17.

[22] EVERINGHAM M, ESLAMI S M A, GOOL L V, et al. The Pascal, Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[23] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[M]// Computer Vision-ECCV 2014. Springer International Publishing, 2014: 740-755.

[24] MOTTAGHI R, CHEN X, LIU X, et al. The Role of Context for Object Detection and Semantic Segmentation in the Wild[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014: 891-898.

[25] LIU C, YUEN J, TORRALBA A. Nonparametric scene parsing: Label transfer via dense scene alignment[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR 2009). IEEE, 1972: 1972-1979.

[26] OTSU N. A thresholding selection method from gray-level histogram[J]. IEEE Transactions on Systems Man & Cybernetics, 1979, 9(1): 62-66.

[27] BOVIK A C. On detecting edges in speckle imagery[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1988, 36(10): 1618-1627.

[28] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. Plenum, 1981.

[29] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Computer Vision and Pattern Recognition. IEEE, 2015: 3431-3440.

[30] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4): 357-361.

[31] KOLTUN V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc., 2011: 109-117.

[32] NOH H, HONG S, HAN B. Learning Deconvolution Network for Semantic Segmentation[C]// IEEE International Conference on Computer Vision. IEEE, 2015: 1520-1528.

[33] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional Random Fields as Recurrent Neural Networks[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2015: 1529-1537.

[34] JEGOU S, DROZDZAL M, VAZQUEZ D, et al. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation[C]// Computer Vision and Pattern Recognition Workshops. IEEE, 2017: 1175-1183.

[35] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Image Net classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc., 2012: 1097-1105.

[36] HE K, ZHANG X, REN S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[J]. arXiv: 1502: 01852, 2015: 1026-1034.

[37] XIE G S, ZHANG X Y, SHU X, et al. Task-driven feature pooling for image classification[C]// IEEE International Conference on Computer Vision(ICCV). IEEE, 2015.

[38] WU R, WANG B, WANG W, et al. Harvesting Discriminative Meta Objects with Deep CNN Features for Scene Classification [C]// 2015 IEEE International Conference on Computer Vision

(ICCVA). IEEE，2015：1287-1295.

[39] KRIZHEVSKY A. Learning Multiple Layers of Features from Tiny Images[J]. Handbook of Systemic Autommune Diseases，2009，1(4)：1-58.

[40] LI F F，FERGUS R，PERONA P. Learning Generative Visual Models from Few Training Examples：An Incremental Bayesian Approach Tested on 101 Object Categories[C]// Conference on Computer Vision and Pattern Recognition Workshop(CVPRW'04). IEEE，2005：178-178.

[41] GRIFFIN G，HOLUB A，PERONA P. Caltech-256 Object Category Dataset[R]. California Institute of Technology，2007.

[42] DENG J，DONG W，SOCHER R，et al. ImageNet：A large-scale hierarchical image database[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2009). IEEE，2009：248-255.

[43] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2014：1-9.

[44] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv：1409.1556，2014.

[45] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE，2016：770-778.

[46] HUANG G，LIU Z，WEINBERGER K Q. Densely Connected Convolutional Networks[C]// CVPR. 2016.

[47] CHEN Y，LI J，XIAO H，et al. Dual Path Networks[J]. arXiv：1707.01629，2017.

[48] EVERINGHAM M，GOOL L V，WILLIAMS C K I，et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision，2010，88(2)：303-338.

[49] XIAO J，HAYS J，EHINGER K A，et al. SUN database：Large-scale scene recognition from abbey to zoo[C]// Computer Vision and Pattern Recognition. IEEE，2010：3485-3492.

[50] 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.

[51] ZITNICK C L，DOLLÁR P. Edge Boxes：Locating Object Proposals from Edges[C]// European Conference on Computer Vision. Springer，Cham，2014：391-405.

[52] 温捷文，战荫伟，凌伟林，等. 实时目标检测算法 YOLO 的批再规范化处理[J]. 计算机应用研究，2018，35(11)：1-2.

[53] SERMANET P，EIGEN D，ZHANG X，et al. OverFeat：Integrated Recognition，Localization and Detection using Convolutional Networks[J]. arXiv：1312.6229，2013.

[54] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2014：580-587.

[55] GIRSHICK R. Fast R-CNN [C]// IEEE International Conference on Computer Vision. IEEE Computer Society，2015：1440-1448.

[56] OUYANG W，LOY C C，TANG X，et al. DeepID-Net：Deformable deep convolutional neural networks for object detection [C]// Computer Vision and Pattern Recognition. IEEE，2015：2403-2412.

[57] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. MIT Press，2015：91-99.

[58] SHRIVASTAVA A，GUPTA A，GIRSHICK R. Training Region-Based Object Detectors with Online Hard Example Mining [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2016：761-769.

[59] SUNG KK. Learning and example selection for object and pattern detection [M]. Massachusetts Institute of Technology，1996.

[60] YANG F，CHOI W，LIN Y. Exploit All the Layers：Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers[C]// Computer Vision and Pattern Recognition. IEEE，2016：2129-2137.

[61] BELL S，ZITNICK C L，BALA K，et al. Inside-Outside Net：Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2016：2874-2883.

[62] BYEON W，BREUEL T M，RAUE F，et al. Scene labeling with LSTM recurrent neural networks [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2015：3547-3555.

[63] HE K，GKIOXARI G，DOLLÁR P，et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，PP(99)：1.

[64] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2017：936-944.

[65] GOODFELLOW I J，POUGETABADIE J，MIRZA M，et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems，2014，3：2672-2680.

[66] LI J，LIANG X，WEI Y，et al. Perceptual Generative Adversarial Networks for Small Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2017：1951-1959.

[67] REDMON J，DIVVALA S，GIRSHICK R，et al. You Only Look Once：Unified，Real-Time Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2016：779-788.

[68] NAJIBI M，RASTEGARI M，DAVIS L S. G-CNN：An Iterative Grid Based Object Detector[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2016：2369-2377.

[69] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single Shot MultiBoxDetector[M]// Computer Vision-ECCV 2016. Springer International Publishing，2016：21-37.

[70] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger [J]. arXiv：1612.08242，2016：6517-6525.

[71] REN J，CHEN X，LIU J，et al. Accurate Single Stage Detector Using Recurrent Rolling Convolution[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2017：752-760.

[72] LIPTON Z C，BERKOWITZ J，ELKAN C. A Critical Review of Recurrent Neural Networks for Sequence Learning[J]. arXiv：1506.00019，2015.

[73] KARPATHY A，TODERICI G，SHETTY S，et al. Large-Scale

Video Classification with Convolutional Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2014:1725-1732.

[74] JI S,YANG M,YU K. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2012,35(1):221-231.

[75] BACCOUCHE M,MAMALET F,WOLF C,et al. Sequential deep learning for human action recognition[C]// International Conference on Human Behavior Unterstanding. Springer-Verlag,2011:29-39.

[76] KANG K,LI H,YAN J,et al. T-CNN:Tubelets with Convolutional Neural Networks for Object Detection from Videos[J]. arXiv:1604.02532,2016.

[77] ZHU X,XIONG Y,DAI J,et al. Deep Feature Flow for Video Recognition[J]. arXiv:1611.07715,2016.

[78] 潘光远. 光流场算法及其在视频目标检测中的应用研究[D]. 上海:上海交通大学,2008.

[79] SHOU Z,CHAN J,ZAREIAN A,et al. CDC:Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2017:1417-1426.

[80] ZEILER M D,FERGUS R. Visualizing and Understanding Convolutional Networks[C]// European Conference on Computer Vision. Springer,Cham,2014:818-833.

[81] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[82] FELZENSZWALB P,GIRSHICK R,MCALLESTER D,et al. Visual Object Detection with Deformable Part Models[C]// Computer Vision and Pattern Recognition. IEEE,2010:2241-2248.

[83] GU C,LIM J J,ARBELAEZ P,et al. Recognition using regions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2009:1030-1037.

[84] CARREIRA J,SMINCHISESCU C. CPMC:Automatic Object Segmentation Using Constrained Parametric Min-Cuts[M]. IEEE Computer Society,2012.

[85] 王万国,田兵,刘越,等. 基于 RCNN 的无人机巡检图像电力小部件识别研究[J]. 地球信息科学学报,2017,19(2):256-263.

[86] HE K,ZHANG X,REN S,et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[C]// European Conference on Computer Vision. Springer,Cham,2014:346-361.

[87] DAI J,LI Y,HE K,et al. R-FCN:Object Detection via Region-based Fully Convolutional Networks[J]. arXiv:1605.06409,2016.

[88] RUSSAKOVSKY O,DENG J,SU H,et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision,2015,115(3):211-252.

[89] LIN M,CHEN Q,YAN S. Network In Network[J]. arXiv:1312.44003v3,2013.

[90] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436.

[91] DAI J,QI H,XIONG Y,et al. Deformable Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE,2017:764-773.

(上接第 11 页)

[101]NORI F,DEYPIR M,HADI M,et al. A new sliding window based algorithm for frequent closed itemset mining over data streams[J]. Journal of Systems & Software,2013,86(3):615-623.

[102]DONG J,HAN M. BitTableFI:An efficient mining frequent itemsets algorithm[J]. Knowledge-Based Systems,2007,20(4):329-335.

[103]SONG W,YANG B,XU Z. Index-BitTableFI:An improved algorithm for mining frequent itemsets[J]. Knowledge-Based Systems,2008,21(6):507-513.

[104]BAYARDO R J. Efficiently mining long patterns from databases[C]// ACM SIGMOD International Conference on Management of Data. ACM,1998:85-93.

[105]AGARWAL R C,AGGARWAL C C,PRASAD V V V. Depth first generation of long patterns[C]// ACM SIGKDD International Conference on Knowedge Discovery and Data Mining. ACM,2000:108-118.

[106]BURDICK D,CALIMLIM M,FLANNICK J,et al. MAFIA:A Maximal Frequent Itemset Algorithm[C]// International Conference on Data Engineering. IEEE Computer Society,2001:443.

[107]GOUDA K,ZAKI M J. Efficiently Mining Maximal Frequent Itemsets[C]// IEEE International Conference on Data Mining. IEEE,2002:2405-2409.

[108]ZOU Q,CHU W W,LU B. SmartMiner:A Depth First Algorithm Guided by Tail Information for Mining Maximal Frequent Itemsets[C]// IEEE International Conference on Data Mining,2002(ICDM 2003). IEEE,2002:570-577.

[109]宋余庆,朱玉全,孙志挥,等. 基于 FP-Tree 的最大频繁项目集挖掘及更新算法[J]. 软件学报,2003,14(9):1586-1592.

[110]颜跃进,李舟军,陈火旺,等. 基于 FP-Tree 有效挖掘最大频繁项集[J]. 软件学报,2005,16(2):215-222.

[111]秦亮曦,史忠植. SFP-Max-基于排序 FP-树的最大频繁模式挖掘算法[J]. 计算机研究与发展,2005,42(2):217-223.

[112]JU S,CHEN C. MMFI:An Effective Algorithm for Mining Maximal Frequent Itemsets[C]// International Symposiums on Information Processing. IEEE Computer Society,2008:144-148.

[113]钱雪忠,惠亮. 关联规则中基于降维的最大频繁模式挖掘算法[J]. 计算机应用,2011,31(5):1339-1343.

[114]ZHAO Z G,WANG F,WAN J. Maximal frequent itemsets mining algorithm based on OWSFP-tree[J]. Computer Engineering & Design,2013,34(5):1687-1680.

[115]YANG P,PENG H,ZHOU X,et al. FP-MFIA:improved algorithm for mining maximum frequent itemsets based on frequent-pattern tree[J]. Journal of Computer Applications,2015,35(3):775-778.