基于深度学习的目标检测算法综述*

周晓彦 王 珂 李凌燕
(南京信息工程大学江苏省气象探测与信息处理重点实验室 南京 210044)

摘 要：传统的目标检测算法及策略已经难以满足目标检测中数据处理的效率、性能、速度和智能化等各个方面要求。深度学习通过对大脑认知能力的研究和模仿以实现对数据特征的分析处理，具有强大的视觉目标检测能力，成为了当前目标检测的主流算法。首先回顾了传统目标检测的发展以及存在的问题；其次介绍以 R-CNN 为代表的结合 region proposal 和卷积神经网络(CNN)分类的目标检测框架(R-CNN、SPP-NET、Fast R-CNN、Faster R-CNN)；然后介绍以 YOLO 算法为代表的将目标检测转换为回归问题的目标检测框架(YOLO、SSD)；最后对深度学习的目标检测算法存在的问题做出总结，以及未来的发展方向。

关键词：深度学习；卷积神经网络；目标检测
中图分类号：TP183 **文献标识码**：A **国家标准学科分类代码**：250.2060

Review of object detection based on deep learning

Zhou Xiaoyan Wang Ke Li Lingyan
(Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: The traditional target detection algorithm and strategy has been difficult to meet the target detection of data processing efficiency, performance, speed and intelligence and other aspects. Depth learning through the study of brain cognitive ability and imitation to achieve the analysis of data characteristics of the treatment, with a strong visual target detectioncapabilities, has become the current target detection of the mainstream algorithm. Firstly, the developmentand problems of traditional target detectionare reviewed; Secondly, the target detection framework which combines region proposal and CNNclassification with R-CNN is introduced(R-CNN, SPP-NET, Fast R-CNN, Faster R-CNN); Then, the target detection framework is introduced, which is based on YOLO(YOLO, SSD)algorithm; Finally, this paper makes a summary of the problemsexisting in the target detection algorithm of deep learning and the development of the future.

Keywords: deep learning; convolutional neural network; object detection

0 引 言

目标检测(objection detection)是机器视觉中最常见的问题。是一种基于目标几何和统计特征的图像分割，它将目标的分割和识别合二为一，其准确性和实时性是整个系统的一项重要能力，近年来，目标检测在人工智能，人脸识别，无人驾驶等领域都得到了广泛的应用[1]。然而，在目标检测的过程中会受到各种各样干扰，比如角度、遮挡、光线强度等因素，这些因素会导致目标发生畸变，为目标检测增加了新的挑战。如图 1 所示传统目标检测方法的 3 个步骤：1)使用不同大小的滑动窗口框住待测图像中的某一部分作为候选区域，然后提取该候选区域相关的视觉特征，比如人脸检测常用的 Harr[2]特征；2)行人检测和普通目标检测常用的 HOG[1-3](histogram of orientation gradient)特征等；3)使用训练完成的分类器进行分类，比如常用的支持向量机[4](support vector machine，SVM)模型，Adaboost[5]、DPM[6]、RF[7](random forest)模型等。

图 1 传统目标检测的 3 个阶段

在传统的目标检测中，Felzenszwalb 等人提出了多尺度形变部件模型(deformable part model，DPM)。DPM 在 HOG 和 SVM 的基础之上进行性能延拓，充分的利用了 HOG 和 SVM 的优点，在图像处理、人脸识别等任务上取

收稿日期：2017-04
* 基金项目：国家自然科学基金(61201444)资助

得了重要突破，但是传统目标检测算有着两个主要的缺陷[8]：1）使滑动窗口策略进行区域选择时针对性不强，提高了时间复杂度和窗口冗余；2）手动设计的特征对于目标的多样性并没有很好的鲁棒性。DPM 模型的复杂度较高，目标检测的速度和准度较低。随着深度学习的崛起，目标检测的精度不断的提升。比如文献[9]中的算法在 VOC 2007 测试集合上的 mAP 只能 30％多一点，文献[10]中的 OverFeat 在 ILSVRC 2013 测试集上的 mAP 只能达到 24.3％。2013 年 R-CNN 诞生了，VOC 2007 测试集的 mAP 被提升至 48％，2014 年时通过修改网络结构又上升到了 66％，同时 ILSVRC 2013 测试集的 MAP 也被提升至 31.4％。Girshick 等人[11-13]提出的 R-CNN 在目标检测领域取得了突破，先后出现了 SPP-net[14]、Fast R-CNN[15]、Faster R-CNN[16]、R-FCN[17]、YOLO[18]、SSD[19] 等算法。这些创新算法都是把传统的计算机视觉领域和深度学习结合二为一，效果显著，比如选择性搜索（selective search）和图像金字塔（pyramid）等。因此，基于深度学习的目标检测算法的到了广大研究者们的关注，成为了机器学习领域的热点之一。

1　基于 Region Proposal 的深度学习目标检测算法

卷积神经网络（CNN）是 Region Propoal 算法中的核心组成部分。卷积神经网络最早是由 Yann LeCun 教授提出来的，早期的卷积神经网络是用作分类器使用，主要用于图像的识别。然而卷积神经网络有 3 个结构上的特性：局部连接、权重共享以及空间或时间上的采样。这些特性使得卷积神经网络具有一定程度上的平移、缩放和扭曲不变性。在 2006 年 Hinton 提出利用深度神经网络从大量的数据中自动的学习高层特征。Region Propoal 在此基础之上解决了传统目标检测的两个主要问题。比较常用的 region proposal 方法有 Selective Search[20] 和 Edge Boxes[21]。此后，CNN 网络迅速发展，微软最新的 ResNet 和谷歌的 Inception V4[22-23] 模型的 Top-5 error 降到了 4％以内，所以目标检测得到候选区域后使用 CNN 对其进行图像分类的准确率和检测速度上都有提高。

1.1　R-CNN

如上所述，在初期的目标检测算法中采用的是滑动窗口策略提取特征，充分的利用了穷举法的特性进行遍历，Girshick 等人提出 R-CNN 采用的是 Selective Search，使用聚类的方法，对图像进行分个分组，得到多个候选框的层次组。R-CNN 的实现主要分为 4 个步骤：首先区域提名通过 Selective Search 从原始图片提取 2 000 个左右区域候选框；其次区域大小归一化把所有侯选框缩放成固定大小（原文采用 227×227 的尺寸）；然后用 CNN 网络进行特征提取；最后分类与回归在特征层的基础上添加两个全连接层，再用 SVM 分类来做识别，用线性回归来微调边框位置与大小，其中每个类别单独训练一个边框回归器。

R-CNN 在 PASCAL VOC2007 上的检测结果从 DPM HSC 的 34.3％直接提升到了 66％（mAP），但是 R-CNN 需要对 SS 提取得到的每个 proposal 进行一次前向 CNN 实现特征提取，因此计算量很大，无法实时更新。此外，由于全连接层[24-25]的存在，需要严格保证输入的 proposal 最终重新到相同尺度大小，这在一定程度造成图像畸变，影响最终结果。但是针对这些问题，SPP-NET 给出了很好的解决方案。

1.2　SPP-NET

SPP-NET 算法是 He 等人提出的，其主要思想是去掉了原始图像上的 crop/warp 等操作，换成了在卷积特征上的空间金字塔池化层（spatial pyramid pooling，SPP），引入 SPP 层，主要原因是 CNN 的全连接层要求输入图片是大小一致的，而实际中的输入图片往往大小不一，如果直接缩放到同一尺寸，很可能有的物体会充满整个图片，而有的物体可能只能占到图片的一角。传统的解决策略是对不同的图像位置进行裁剪，但是这些裁剪技术都可能会导致一些问题出现：1）crop 时会破坏物体的完整性；2）warp 时导致物发生严重的畸变。SPP-net 做目标检测的主要步骤如下：首先区域提名用 Selective Search 从原候选窗口；其次区域大小缩放 SPP-net 不再做区域大小归一化，而是缩放到 $\min(w,h)=s$，即统一长宽的最短边长度，然后特征提取利用 SPP-net 网络结构提取特征；最后分类与回归。利用 SVM 基于上面的特征训练分类器模型，用边框回归来微调候选框的位置。SPP-net 解决了 R-CNN 区域提名时 crop/warp 带来的偏差问题，提出了 SPP 层，使得输入的候选框可大可小，但其他方面依然和 R-CNN 一样，因而依然存在问题，这就有了后面的 Fast R-CNN。

1.3　Fast R-CNN

2015 年 Ren 等人提出 Fast R-CNN 是要解决 R-CNN 和 SPP-net 2000 个左右候选框带来的重复计算问题，其主要思想为：首先使用一个简化的 SPP 层 RoI[26]（region of Iinteresting）Pooling 层，操作与 SPP 类似；然后训练和测试是不再分多步执行，不再需要额外的硬盘来存储中间层的特征，梯度能够通过 RoI Pooling 层直接传播；此外，分类和回归用多任务[27]的方式一起进行；最后使用 SVD 分解全连接层的参数矩阵，压缩为两个规模小很多的全连接层。Fast R-CNN 的主要步骤如下：1）进行特征提取，利用神经网络算法对物体进行特征提取；2）利用滑动窗口策略等方法提取区域候选框，并把这些候选框全部投影到最后的特征层；3）针对特征层上的每个区域候选框进行 RoI Pooling 操作，得到固定大小的特征表示；4）再通过两个全连接层，分别用 softmax 多分类做目标识别，用回归模型进行边框位置与大小微调。Fast R-CNN 融合了 R-CNN 和 SPP-NET 的精髓，并且引入多任务损失函数使整个网络的训练和测试变得十分方便。其中对一个图像的损失函数表达式

如下:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \quad (1)$$

$$p_i = \begin{cases} 1, & \text{正 anchor} \\ 0, & \text{负 anchor} \end{cases} \quad (2)$$

式中:i 代表 mini-batch anchor 的索引;p_i 是目标预测的概率;t_i 是一个向量,表示预测的包围盒的 4 个参数化坐标;t_i^* 是与正 anchor 对应的 ground truth 坐标向量;N_{cls} 是 mini-batch 的大小;N_{reg} 是 anchor 位置的数量;λ 是平衡权重,归一化。

在 Pascal VOC2007 训练集上训练,在 VOC2007 测试的结果为 66.9%(mAP),如果使用 VOC2007 结合 2012 训练集训练,在 VOC2007 上测试结果为 70%(数据集的扩充能大幅提高目标检测性能)。使用 VGG16 每张图像总共需要 3 s 左右。但是 region proposal 的提取使用 Selective Search,目标检测时间大多消耗在这上面,无法满足实时应用,而且并没有实现真正意义上的端到端训练测试,显然 proposal 提取成为端到端算法性能的瓶颈。因此在 2016 年 Ren 等人提出 Faster R-CNN 算法该算法在原有的 Fast R-CNN 算法上引入了 RPN(region proposal network)网络,显著的缩小了 proposal 提取的时间。

1.4 Faster R-CNN

2016 年 Ren 等人提出新的 Faster-R-CNN 算法,该算法引入了 RPN 网络提取 proposals。RPN 网络是一个全卷积神经网络,通过共享卷积层特征可以实现 proposal 的提取,RPN 提取一幅像的 proposal 只需要 10 ms 左右,RPN 的核心思想是使用卷积神经网络直接产生 region proposal,使用的方法本质上就是滑动窗口。RPN 的设计比较巧妙,RPN 只需在最后的卷积层上滑动一遍,因为 anchor 机制和边框回归可以得到多尺度长宽比的 region proposal。Faster R-CNN 的主要步骤:1)特征提取,同 Fast R-CNN,以整张图片为输入,利用 CNN 得到图片的特征层;2)区域提名,在最终的卷积特征层上利用 k 个不同的矩形框(anchor box)进行提取,k 值通常取 9;3)分类与回归,对每个 Anchor Box 对应的区域进行 object/non-object 二分类,并用 k 个回归模型,每个模型对应不同的 Anchor Box。对候选框的位置进行微调,最后再进行分类。总之,Faster R-CNN 抛弃了滑动窗口这种策略,引入了 RPN 网络,使区域提取、分类、回归共用卷积特征,从而运算的速度大大提升。但是,Faster R-CNN 需要对大量的 Anchor Box 进行目标判定,然后再进行目标识别,随着 SPP-net、Fast R-CNN、Faster R-CNN 这么多算法的快速发展,基于深度学习目标检测的效果来越好,速度越来越快。

2 基于回归方法的深度学习目标检测算法

虽然 Faster R-CNN 算法是目前主流的目标检测算法之一,但是速度上并不能满足实时的要求。随后出现像 YOLO,SSD 这一类的算法逐渐凸显出其优略性,这类方法充分的利用了回归的思想,直接在原始图像的多个位置上回归,出目标位置边框以及目标类别。

2.1 YOLO

2016 年 Redmon 等人提出的 YOLO 算法是一个可以一次性预测多个 Box 位置和类别的卷积神经网络,YOLO 算法的网络设计策略延续了 GoogleNet[28] 的核心思想,真正意义上实现了端到端的目标检测,且发挥了速度快的优势,但其精度有所下降。然而在 2016 年 Redmon 等人提出的 YOLO9000[29] 算法是在原先 YOLO 算法的速度上提高了其准确度。主要有两方面的改进:1)在原有的 YOLO 检测框架上进行了一系列的改进,弥补了检测精度的不足;2)提出了目标检测和目标训练合二为一的方法。YOLOv2 算法的训练网络采用降采样的方法在特定的情况下可以进行动态调整,这种机制可以使网络预测不同大小的图片,让检测的速度和精度之间达到平衡。表 1 是 YOLOv2 和其他网络在 VOC2007 上的对比。

表 1 YOLOv2 和其他网络在 VOC2007 上的对比

Detection Frameworks	Train	mAP	FPS
Fast R-CNN	2007+2012	70	0.5
FasterR-CNNVGG-16	2007+2012	73.2	7
Faster R-CNN ResNet	2007+2012	76.4	5
YOLO	2007+2012	63.4	45
SSD300	2007+2012	74.3	46
SSD500	2007+2012	76.8	19
YOLOv2 288×288	2007+2012	69	91
YOLOv2 352×352	2007+2012	73.7	81
YOLOv2 416×416	2007+2012	76.8	67
YOLOv2 480×480	2007+2012	77.8	59
YOLOv2 544×544	2007+2012	78.6	40

由表 1 可以看出 YOLOv2 算法在高分辨率图片检测中超出了实时检测速度的要求,达到了先进的水平。

2.2 SSD

2016 年 Liu 等人提出 SSD 算法,该算法结合 YOLO 的回归思想以及 Faster R-CNN 的 anchor 机制做到了速度与准确率并存。最初的 YOLO 算法是在 7×77×7 的框架下识别物体,用这种框架检测小物体时,准确率会下降。在 SSD 算法中就去掉了 YOLO 算法中的全连接层,所以对任意大小的物体都可以检测,性能基本不。对 SSD 的测试集进行训练和训练使用候选区域及用来池化的标准检测器之间最大的不同之处在于,ground truth 需要被赋予一组固定集合检测输出中某一个特定输出。当这个赋值确定之后,损失函数和后向传播就能够实现端到端的应用。总之,SSD 结合了 YOLO 中的回归思想和 Faster R-CNN 中的

anchor 机制，使用全图各个位置的多尺度区域特征进行回归，既保持了 YOLO 速度快的特性，也保证了窗口预测的跟 Faster R-CNN 一样比较精准。

3　结　　论

基于深度学习的目标检测算法总体上分为两类：1）基于区域提名的 R-CNN 算法系列；2）无需区域提名的 YOLO、SSD 算法系列。R-CNN 系列目标检测算法框架和 YOLO 目标检测算法框架给了我们对目标检测的研究提供了两种基本框架。在此基础之上研究者们提出了一些提高目标检测性能的方法：1）难分样本挖掘[30]（hard negative mining）；2）多层特征融合[31]；3）使用上下文信息。除此之外，各种开源深度学习框架层出不穷，其中包括 Tensorflow、Caffe、Keras、Theano，等等。这些开源框架加速了深度学习的发展，为目标检测设计更加合理的网络结构，提升回复式神经网络检测效率，实现多尺度多类别的目标检测提供了高效的学习工具。总的来说，基于深度学习的目标检测仍然是一个具有挑战性的课题，挑战性主要体现在以下两个方面：鲁棒性和计算复杂性[32]。当前，随着大数据和人工智能时代的到来，该课题提供了新的机遇和挑战，因此值得展开更深入的研究。

参考文献

[1] 王顺飞，闫钧华，王志刚. 改进的基于局部联合特征的运动目标检测方法[J]. 仪器仪表学报，2015，36(10)：2241-2248.

[2] VIOLA P，JONES M. Rapid object detection using a bo osted cascade of simple features [C]. IEEE Computer Society Confere nce on Computer Vision and Pattern Recognition，2001：I-511-I-518 .

[3] 蔡娟，李东新. 基于优化 k 均值建模的运动目标检测算法[J]. 国外电子测量技术，2016，35(12)：20-23.

[4] VAPNIK V N. The nature of statistical learning theory [C]. The Nature of Statistical Learning Theory. Berlin：Springer，1995：988-999.

[5] FERREIRA A J，FIGUEIREDO M A T. Boosting algorithms：A review of theory，methods，and applications [C]. Ensemble Machine Learning：Methods and Applications，2012.

[6] FELZENSZWALB P F，GIRSHICK R B，MCALLESTER D，etal. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2010，32(9)：1627-1645.

[7] BREIMAN L. Machine learning，Springer link[J]. Machine Learning，2001，45(1)：5 - 32.

[8] ROSENBERG C，HEBERT M，SCHNEIDERMAN H. Semi-sup ervised self-training of object detection models[C]. IEEE Workshop on Applications of Computer Vision/ IEEE Workshop on Motion and Video Computing，2005：29-36.

[9] SZEGEDY C，TOSHEV A，ERHAN D. Deep neural networks for object detection[C]. Advances in Neural Infor mation Processing Systems，2013.

[10] SERMANET P，EIGEN D，ZHANG X，et al. Overfeat：Integrated recognition，localization and detection using convolutional networks [C]. ICLR，2014.

[11] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object de tection and semantic segmentation[C]. ImageNet Large-Scale Visua Recognition Challenge Work Shop，ICCV，2013.

[12] 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 (CVPR)，2014.

[13] GIRSHICK R，DONAHUE J，DARRELL T，et al. Region-Based convolutional networks for accu ate object detection and segmentation [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015.

[14] HE K，ZHANG X，REN S，et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]. ECCV，2014.

[15] GIRSHICK R. Fast R-CNN[C]. ICCV，2015.

[16] REN S，HE K，GIRSHICK R，et al. FasterR-CNN：Towards real-time object detection with region proposal networks [C]. Advances in Neural Information Processing Systems，2015.

[17] DAI J F，LI Y，HE K M，et al. R-FCN：Object detection via region-based fully convolutional networks [C]. Conference on Neural Information Processing Systems (NIPS)，2016.

[18] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection[C]. CVPR，2016.

[19] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single shot multibox detector[J]. 2015，arXiv：15 12. 02325.

[20] SANDE K，UIJLINGS J R R，GEVERS T，et al. Segmentation as selective search for object recognition[C]. IEEE International Conference on Comuter Vision，2011：1879-1886.

[21] ZITNICK C L，DOLLÁR P. Edge boxes：Locating

object proposals from edges [C]. European Conference on Computer Vision, 2014.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Image net classification with deep convolutional neural networks [C]. International Conference on Neural Information Processing Systems, 2012:1097-1105.

[23] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Incept ion-4, inception-resnet and the impact of residual connections on learning[J]. Computer Vision and Pattern Recognition, 2016, arXiv:1602.07261.

[24] AGRAWAL P, GIRSHICK R, MALIK J. Analyzing the performance of multilayer neural networks for object recognition[C]. Computer Vision-ECC V2014. Springer International Publishing, 2014:329-344.

[25] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-ada ptation of feature detectors [J]. Computer Science, 2012, 3(4): 212-223.

[26] WANG J, WU Y, LIANG Z, et al. Lane detection b ased on random hough transform on region of in terestin [C]. IEEE International Conference on Information and Automation, 2010:1735-1740.

[27] SCHÖLKOPF B, PLATT J, HOFMANN T. Multi-task fe ature learning[C]. Conference on Advances in Neural Information Processing Systems, 2006:41-48.

[28] ZHONG Z, JIN L, XIE Z. High performance off line handwritten Chinese character recognition using GoogLeNet and directional feature maps [C]. International Conference on Document Analy s is and Recognition, 2015:846-850.

[29] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[J]. Computer Vision and Pattern Recognition, 2016, arXiv:1612.08242.

[30] HENRIQUES J F, CARREIRA J, CASEIRO R, et al. Beyond hard negative mining: Efficient detecto rlearning via block-circulant decomposition[C]. IEEE International Conference on Computer Vision, 2014: 2760-2767.

[31] KONG T, YAO A, CHEN Y, et al. HyperNet: Towards accurate region proposal generation and joint object detection[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 845-853.

[32] 任克强, 高晓林. 基于五帧差和二维 Renyi 熵的运动目标检测[J]. 电子测量与仪器学报, 2015,29(8): 1179-1186.

作者简介

周晓彦,工学博士,副教授,主要研究方向为模式识别、信号处理等。

王珂,工学硕士,主要研究方向为目标检测,模式识别。E-mail:459506810@qq.com