

**人工智能与机器学习课程报告**

**2020年春季学期**

**考 核 科 目： 人工智能与机器学习**

**学生所在院系： 汽车工程师学院**

**授 课 教 师： 龚家元 陈安庆**

**报 告 题 目：基于深度学习的目标检测算法综述**

**姓 名： 梅振东**

**学 号： 201702053**

目录

**[基于深度学习的目标检测算法综述 - 1 -](#_Toc32482)**

**[1 深度学习目标检测算法 - 2 -](#_Toc657)**

**[2 基于分类的检测算法 - 2 -](#_Toc15634)**

[2.1 OverFeat - 2 -](#_Toc6544)

[2.2 Ｒ-CNN - 2 -](#_Toc2561)

[2.3 SPP-Net - 3 -](#_Toc26910)

[2.4 FastＲ-CNN - 3 -](#_Toc13287)

[2.5 FasterＲ-CNN - 3 -](#_Toc28377)

**[3 基于回归的检测算法 - 4 -](#_Toc12559)**

[3.1 Deep neural network for object detection - 4 -](#_Toc636)

[3.2 可扩展的目标检测 - 4 -](#_Toc12918)

[3.3 基于记忆的检测模型 - 4 -](#_Toc12411)

[3.4 YOLO - 4 -](#_Toc23682)

[3.5 SSD - 5 -](#_Toc13914)

[3.6 YOLOv2 - 5 -](#_Toc24186)

[3.7 YOLOv3 - 5 -](#_Toc5839)

**[4 结语 - 6 -](#_Toc9887)**

**基于深度学习的目标检测算法综述**

目标检测主要是对输入图像中的物体类别和位置进行判断，实质上是图像分类和目标定位的结合，是计算机视觉领域中的一个重要研究方向。目标检测在人脸识别、无人驾驶等领域取得了广泛的应用。在深度学习兴起之前，传统的目标检测算法大多采用人工特征来进行检测，人工特征主要针对目标的某个具体特征进行检测。其中，Sobel、Prewitt、Ｒoberts、Canny和LoG等描述子主要用来做边缘检测;Harris、FAST、CSS和DOG主要用于角点检测等。人工设计特征的好坏对目标检测的精度起到了决定性作用。深度学习使用非线性模型将原始输入数据逐层转化为更高层次、更抽象的特征，具有强大的学习和表征能力，能够自动发现分类或检测任务所需要的特征。深度学习的发展极大地推动了目标检测，基于深度学习的目标检测获得了广泛的关注，成为了当今计算机视觉领域的研究热点之一。

1 深度学习目标检测算法

卷积神经网络AlexNet、GoogleNet、Ｒes-Net等在图像分类任务中取得了较好的成果。研究人员将这些卷积神经网络应用到目标检测中，相对于传统的目标检测算法，在速度和准确度上都有了极大提升。基于深度学习的目标检测算法主要分为2类:1)以Ｒ-CNN系列为代表的基于分类的目标检测框架(twostage);2)以YOLO、SSD算法为代表的基于回归的目标检测框架(onestage)。

2 基于分类的检测算法

基于分类的目标检测算法首先利用SelectiveSearch、EdgeBoxes等算法生成可能包含待检测目标的候选区域(Ｒegionproposal)，再对这些候选区域进一步进行分类以及位置校准得到最终的检测结果。

## 2.1 OverFeat

Sermanet等人在AlexNet的基础上做了改进，提出了OverFeat算法。OverFeat是最先将深度学习应用到目标检测中的算法之一。OverFeat利用卷积神经网络提取特征，可同时实现图像分类、定位以及目标检测任务，不同任务之间共享特征提取层，在ILSVＲC2013数据集上的mAP达到了24．3%。虽然OverFeat相对传统算法有显著改进，但共享特征层的表达能力不强，对小目标的检测效果不好，整体效果仍然差强人意。

## 2.2 Ｒ-CNN

Girshick等人提出的Ｒ-CNN算法首先采用Se-lectiveSearch方法，将输入图像分割成多个模块，再基于颜色、纹理等相似度对这些模块进行合并，最终得到大约2000个不同大小的候选区域，并将这些区域归一化到固定大小;然后使用卷积神经网络对这些候选区域分别进行特征提取，采用多个SVM对这些提取到的特征进行分类;最后对这些区域进行位置校准。Ｒ-CNN在VOC数据集上获得了58．5%的准确率，但是Ｒ-CNN存在以下2个问题:1)对每个候选区域进行特征提取会产生大量重复运算，无法做到实时更新;2)由于卷积神经网络的全连接层要求所有输入大小相同，Ｒ-CNN将图片缩放到相同大小，破坏了物体的长宽比，导致部分有用信息丢失，影响检测效果。

## 2.3 SPP-Net

为了解决Ｒ-CNN只能检测固定大小图像的问题，He等人提出了SPP-Net。SPP-Net只需要对输入图像进行一次卷积得到整张图片的特征图(featuremap)，然后将不同大小的候选区域映射到特征图上。为了满足全连接层对输入大小一致的要求，SPP-Net采用空间金字塔池化层(SpatialPyramidPooling，SPP)，利用不同大小的池化窗口将这些映射到特征图上的候选区域统一到相同维度，最后利用SVM进行分类，同时利用边框回归来微调候选框的位置。空间金字塔池化层使得SPP-Net能够处理不同大小的图片，并且完整保存了输入图片的信息。同时由于SPP-Net只需要对图像进行一次卷积，减少了运算时间，其检测速度比Ｒ-CNN快38～102倍，在VOC2007和Caltech101数据集上取得了较好的成绩。但同Ｒ-CNN一样，在SPP-Net中，确定候选框、特征提取、训练SVM、训练边框回归模型这些步骤依然是单独进行的，导致训练时间较长;在训练时将所有的样本存储到磁盘上会占用大量空间;SPP-Net只能对空间金字塔池化层后的全连接层进行微调，无法同时更新前面的卷积层。

## 2.4 FastＲ-CNN

FastＲ-CNN继承了SPP-Net能够处理不同输入大小图片的优点，但是将SPP-Net多尺度空间金字塔池化层简化为池化窗口大小，可以任意设定单尺度的ＲOlPooling;同时引入了多任务学习模式，将多个步骤简化到一个卷积神经网络中;对卷积神经网络的全连接层输出做SVD分解得到2个向量，分别用于softmax分类以及边框回归。FastＲ-CNN检测时所有的特征都存放在显存中，释放了大量磁盘空间。FastＲ-CNN主要分为以下4个步骤:1)利用卷积神经网络提取图像特征。2)利用区域生成算法提取候选区域，并将这些区域映射到特征图上。3)利用ＲOlPooling将候选区域池化到固定大小的特征表示。4)分别应用softmax做多分类目标识别，应用边框回归对边框位置进行微调。FastＲ-CNN极大提升了检测速度，在精度上也有了一定提高，但是对候选区域的提取依然采用Se-lectiveSearch方法，消耗了大量时间，仍然无法达到实时性的要求。

## 2.5 FasterＲ-CNN

FasterＲ-CNN利用ＲPN网络提取候选区域，并将ＲPN和FastＲ-CNN融合到一个网络中共享卷积特征，实现了真正意义上的端到端训练。ＲPN是一个全卷积网络，其思想是通过在特征图上做窗口滑动，将锚点框(anchor)映射到原图，得到不同大小和比例的候选区域。ＲPN主要有2个输出:1)分类支路采用softmax得到anchor属于物体或背景的概率;2)边界框回归支路对anchor的位置进行校准得到4个坐标值。FasterＲ-CNN的主要步骤为:1)利用卷积神经网络对输入图片提取特征。2)利用ＲPN生成k个不同大小和比例的候选区域。3)ＲOIPooling利用ＲPN生成的候选区域和卷积神经网络提取到的共享特征图得到固定大小的候选区域特征图(proposalfeaturemap)。4)利用softmax对proposals进行具体类别的分类，利用边框回归获得物体的精确位置。FasterＲ-CNN利用ＲPN进一步提升了检测速度，同时提高了检测精度。由于FasterＲ-CNN经过2020年第5期曹燕，等:基于深度学习的目标检测算法研究综述65多次下采样丢失了物体细节信息，对小物体的检测效果仍然不好。

3 基于回归的检测算法

## 3.1 Deep neural network for object detection

Szegedy等人首次提出了利用深度神经网络(DeepNeuralNetwork，DNN)将目标检测当做回归问题处理。Szegedy将AlexNet最后的多分类层改为回归层，对每个目标输出二进制掩模(binarymask)，如果某个像素在给定物体的边界框内则置为1，否则为0。为了检测多个距离相近的目标，神经网络需要输出多个二进制掩模。为了精确定位，将二进制掩模的像素放大到与输入图像一致，同时使用定位器进一步提高定位准确度。

## 3.2 可扩展的目标检测

Erhan等人提出的神经网络模型对每个盒子预测一组边界框以及一个分数，表示这个盒子包含物体的可能性。在训练过程中采用聚类方法来得到物体的典型位置以便于与真实框进行匹配，随后的YOLOv2也采用了类似的思想来得到用于训练的anchor。

## 3.3 基于记忆的检测模型

Li等人通过模拟大脑的记忆和预测机制设计了一种检测模型，首先在静态图上滑动固定大小的窗口以生成序列，接着利用卷积神经网络来提取这些序列的特征，最后长短期记忆以适当的顺序接收这些序列特征。Li等人提出的模型集成了特征学习和序列学习，可应用于预测监督场景中对象潜在的位置。

## 3.4 YOLO

YOLO的网络结构延续了GoogleNet的核心思想，利用卷积层来提取图片特征，利用全连接层来得到最后的检测结果。YOLO将输入图像分成S×S个网格，每个网格负责检测中心落在该网格的物体。YOLO的端到端训练和预测使得其具有较快的检测速度，可以达到45f/s，然而YOLO的检测精度却只有63．4%;在YOLO算法中虽然每个网格预测B个边框，但是只预测一组类别概率值，并且最后只选择IOU最高的边框作为输出结果。如果一个网格中包含多个物体，YOLO只能检测出其中一个物体。

## 3.5 SSD

SSD借鉴了FasterＲ-CNN中anchor的机制，在特征图上生成一系列不同大小和长宽比的默认框，同时采用困难样本挖掘方法对负样本进行采样使得训练过程中正负样本的比例维持在1:3。SSD对每一个默认框输出相应的位置偏移以及类别置信度。值得注意的是，SSD将背景类别也当做了一个特殊的类别，也就是如果检测的类别为C类，那么每一个默认框需要预测C+1个类别置信度值，最后选择置信度最高的类别作为默认框所属的类别。SSD同时采用了特征金字塔的方式来做多尺度融合预测。SSD的检测速度达到了59f/s，mAP为73．9%，相对YOLO来说有了较大提升，但是对小物体的检测效果仍然不理想。

## 3.6 YOLOv2

为了得到更快、更准确的检测效果，YOLOv2在YOLO的基础上做了许多方面的改进:1)在每一层卷积层后加入了BN层加快整个网络的收敛速度;2)使用高分辨率图像来训练网络模型。目前许多目标检测算法采用在ImageNet数据集上得到的预训练模型作为基础特征提取网络，但许多预训练模型的输入图像大小低于256×256。比如YOLO采用224×224的图像来训练分类网络，但检测时采用448×448的图像，导致模型需要适应分辨率的变化。YOLOv2首先利用224×224的图像来预训练分类网络，再用448×448的图像来微调网络模型，最后检测时采用448×448的图像，这样网络就不需要花时间适应不同的分辨率。除此之外，YOLOv2还用了K-means聚类方法来确定anchor的大小、利用Darknet-19作为基础的特征提取网络等。与YOLOv2在同一时期提出的YOLO9000结合检测和分类数据集，在2类数据集上进行训练，可实现9000多类物体的检测。

## 3.7 YOLOv3

YOLOv3在YOLO9000的基础上主要做了如下改进:1)借鉴FasterＲ-CNN的思想。对每个候选框增加了“物体性”的预测，判断其属于前景或者背景的概率，不同的是，YOLOv3对每个真实框只分配一个先验边界框。2)采用多尺度融合预测。YOLOv3借鉴FPN的思想，利用多尺度特征图，提取细粒度的特征，增加对小物体的检测效果。3)采用Darknet-53作为主干网络。Darknet-53综合了YOLOv2的主干网络Darknet-19和ＲesNet的优点，在ImageNet上的Top-5准确率可媲美ＲesNet-152，同时减少了计算量，进一步提高了速度。YOLOv3虽然相对YOLO9000提升了对小物体的检测效果，但对大物体的检测效果却略有下降。

# 4 结语

近年来，目标检测算法蓬勃发展，基于深度学习的目标检测框架仍是目标检测任务的主流方向。目前大部分目标检测算法采用anchor机制，在特征图上产生大量密集的anchor，再对这些anchor进行进一步的分类和位置校准。但大量anchor容易产生正负样本不平衡的问题，同时针对不同数据集需要设置不同尺度与长宽比的anchor。如何设计合适的anchor还需要进一步研究。一些研究人员尝试使用anchor-free的方法来检测目标，但这一类方法对主干网络的要求相对较高。除此之外，如何优化目标检测框架，设计更好的损失函数、更适合目标检测任务的网络来更好地提取特征仍旧是未来目标检测算法的重点研究方向。