doi:10.3772/j.issn.1002-0470.2021.05.004

基于改进 Faster RCNN 的目标检测方法①

王宪保② 朱啸咏 姚明海

(浙江工业大学信息工程学院 杭州 310023)

摘 要 针对基于区域的目标检测算法中定位精度不高的问题,本文提出了一种分裂机制的改进 Faster RCNN 算法。该算法首先选择特征提取能力强的卷积神经网络(CNN)作为骨干网络进行特征的提取;然后通过 12 种不同 Anchors 产生候选目标区,以进一步提升检测的精确度;最后将得到的特征分别传送到两个子网络,分别实现分类与定位。分类网络以全连接结构为基础,定位网络则主要由卷积神经网络构成。本文在 Pascal VOC2007 和 Pascal VOC2012 以及吸尘袋图像集上对算法的有效性进行了验证。结果表明,提出的算法在对目标进行有效检测的同时,定位效果比 Faster RCNN 更加精确,实现了边界框的精准回归。

关键词 目标检测;卷积神经网络(CNN);定位精度;改进 Faster RCNN;分裂机制

0 引 言

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、机器学习等技术,识别图片中是否存在事先定义的类别目标物体,如果存在,返回该类别目标物体的空间位置以及空间范围,一般使用矩形边框进行标定的计算机视觉技术[1]。检测过程一般分为 2 个阶段,第 1 阶段通过目标分类判断输入的图像中是否存在目标物体,第 2 阶段负责将搜索到的目标物体使用边界框进行标注[2]。这要求计算机在准确判断目标类别的同时,还要给出每个目标的准确位置。

在目标检测算法中,图像以像素矩阵的方式存储,需要从中抽象出目标类别和边框位置相关的图像特征才可以进行目标检测[3]。传统目标检测算法,一般根据图像特征点进行匹配或是基于滑窗的框架。首先利用图像预处理方法对输入图像进行去噪、增强、裁剪等操作,之后采用滑动窗口方法对图像进行候选区域的筛选,再采用经典特征提取方法,例如方向梯度直方图(histogram of oriented gradient, HOG)[4],Sift[5],可变形零件模型(deformable parts model, DPM)[6]等对候选区域进行特征提取,最后使用 AdaBoost[7] 和支持向量机(support vector machine, SVM)[8]等机器学习算法对得到的特征进行分类,之后通过目标类别对目标进行边框回归。传统的目标检测模型对于不同特征需要选择适合的分类器,导致其泛用性差、鲁棒性不足。

2012 年之后,深度学习给机器学习领域带来了巨大的变革,计算机视觉技术也得以迅猛发展。传统图像特征提取方法的泛化性能不如深度学习方法,在卷积神经网络(convolutional neural network, CNN)提出以后,深度学习方法已经替代了手工特征方法。凭借 CNN 在提取图像高层特征上的优势,深度学习在目标检测领域取得了较大的成功。文献[9]和文献[10]分别提出了快速的基于区域的卷积网络(fast region based convolutional neural network, Fast RCNN)和更快的基于区域的卷积网络(faster region based convolutional neural network, Faster RCNN)算法。前者利用共享卷积减少了整体

① 国家自然科学基金(61871350),浙江省科技计划项目(2019C011123)和浙江省基础公益研究计划(LGG19F030011)资助项目。

② 男,1977 年生,博士,副教授;研究方向:模式识别,神经网络,图像处理;E-mail: wxb@zjut.edu.cn

(收稿日期:2020-04-03)

网络的计算消耗,但由于其区域生成算法复杂度过高,使其训练及检测速度较慢。后者在前者的基础上结合全卷积网络[11],利用区域提议网络(region proposal network, RPN)替换原先的 Selective Search[12]以及 Edge Boxes[13]算法,共享特征映射图减少了重复的计算,加快了训练和检测的速度。Faster RCNN 为两阶段目标检测算法奠定基础,但 Faster RCNN 无法实现候选区域提取网络和特征提取网络之间的权值共享。文献[14]提出基于区域的全卷积神经网络(region based fully convolutional network, R-FCN),利用位置敏感得分图(position sensitive score maps)在一定程度上解决图像变形在图像分类中一致但在目标检测中不一致的问题,且检测速度高于 Faster RCNN。近年来,又出现了级联基于区域的卷积网络(cascade region based convolutional neural network, Cascade RCNN)[15]等两阶段目标检测器。文献[16]提出的只看一次(you only look once, YOLO)算法,其将目标检测问题定义为图像空间边框回归以及类别预测问题。YOLO 中仅包含单个网络,无需候选区域提取,大幅减少了计算量,虽然检测精度不及 Faster RCNN,但其检测速度是 Faster RCNN 的 10 倍,其将分类和定位结合的思想为后续研究提供了新思路。在 YOLO 的基础上单激发多盒探测器(single shot multibox detector, SSD)[17]结合 Anchor 机制并综合利用多尺度卷积层的信息实现对小目标检测精度的提升。之后文献[18]提出 RefineDet 在 SSD 基础上将信息由粗到细进一步提升回归框信息,同时利用特征融合提升小目标检测精度。YOLOv2[19]以及 YOLOv3 等[20-21]算法针对 YOLO 检测精度差的问题,前者增加 k-means 进行方框先验提升了检测效果,后者采用逻辑回归对方框置信度进行回归并对每个类别独立使用逻辑回归采用二分类交叉熵作损失函数,提升了多标签任务的检测效果。

在目标检测任务中,图像的空间信息对于对象分类是至关重要的,它提供了区域提议(region proposal)中是否涵盖完整的目标对象的信息。全连接头(fully connected head, FC Head)网络适合分类任务,因为它具有全局的空间敏感性。相反,边界框回归任务需要对象级别上下文来确定边界框偏移量。由于卷积可以提取对象级上下文信息,卷积更适合边界框回归任务。但是单一的 FC Head 网络或是卷积头(convolution head)网络都不足以同时处理分类和定位。

本文提出了一种分裂的快速区域卷积网络(divide faster region based convolutional neural network, DF RCNN)算法,将分类任务和边界框回归任务有效地分开,提升了目标检测中边界框定位的精度。利用全连接结构将局部特征进行有效的组合,充分利用特征金字塔提取的特征,获得更好的分类效果。Convolution head 因其丰富的上下文信息提供了更加精确的边界框定位。算法在吸尘袋图像集上进行了验证,结果证明了算法的有效性。

本文后续结构如下。第 1 节对 Faster RCNN 等相关工作进行了叙述;第 2 节对提出的 DF RCNN 算法进行了详细介绍;第 3 节通过实验设计,对 DF RCNN 算法做了进一步分析,验证了所提出的深度学习模型在该场景下的识别效果;第 4 节对全文工作进行了总结和展望。

1 Faster RCNN

1.1 概述

Faster RCNN 模型是在 Fast RCNN 的基础上,用借鉴注意力机制提出的 RPN 替代 Fast RCNN 中的选择性搜索,提供更加精确的 region proposals 的同时,降低了网络计算的冗余,提高了检测的速度。

1.2 网络结构

Faster RCNN 模型的结构如图 1 所示。模型主体分为特征提取网络、RPN、感兴趣区域池化(region of interest pooling, ROI Pooling)以及分类器几部分组成。特征提取网络一般选择 VGG16[22],其通过一系列不同大小的卷积层提取输入图像中不同层级的语义信息,形成特征映射图。

1.3 RPN

RPN 接收任意尺寸的输入图像并输出一组包含得分的矩形框作为 region proposals。使用共享卷积顶层的特征信息,通过 proposal 层在大小为 $w \times h$

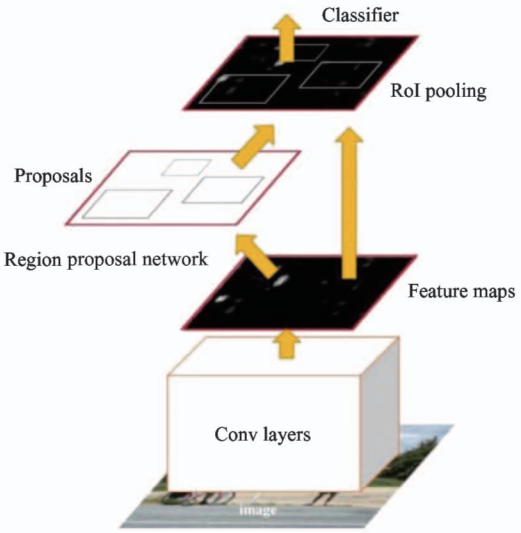


图 1　Faster RCNN 的网络结构[10]

的特征图上的每个像素位置采集 k 个初始区域,总共获得 $w \times h \times k$ 个 Anchors。通过分类器对这些 Anchors 进行遴选,挑出可能含有目标对象的 Anchors 对其进行边界框回归,修正后的 Anchors 作为目标候选区域。

满足下面 2 个条件的 Anchors 在 proposal 层被标注为正样本,否则为负样本。

(1) Anchors 与目标对象的交并比(intersection over union, IOU)最大。

(2) Anchors 与至少一个目标的 IOU 不低于 0.7。

在 Anchors 完成标注后,根据式(1)作为 RPN 的损失函数对 RPN 进行训练。

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

式中,i 是 Anchors 的索引值,p_i 表示索引值为 i 的 Anchor 是目标对象的概率;p_i^* 代表真实值标签(ground truth, GT)的值(如果真实值是正样本则该值为 1,若为负样本则为 0);t_i 是索引值为 i 的 Anchor 的边界修正值,t_i^* 为真实标签的边界框修正值;L_{cls} 是分类损失函数,采用交叉熵损失函数实现,表示预测值与目标值的误差;L_{reg} 是边界框回归的损失函数,采用下面的 *SmoothL*1 损失函数实现,其中 x 表示输入;λ 是分类损失函数和边界框回归损失函数之间的平衡权重。利用该损失函数训练好的 RPN 可以预测目标 Anchors,并修正其边界得到最终的 ROI。ROI 作为 RPN 的输出被传输给 ROI Pooling 以及分类器以便做进一步的检测。

$$SmoothL_1(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{其他} \end{cases} \tag{2}$$

1.4　ROI Pooling

ROI Pooling 使用的是空间金字塔池化[23],将 region proposals 对应的特征图作为输入,输出固定长度的特征向量。特征向量经过全连接层后,分别输入到 Softmax 分类器和边界框回归器中进行目标类别的预测以及边界框的修正。最后,使用非极大值抑制(non-maximum suppression, NMS)消除重复的边界框以得到更加准确的检测结果。该部分的损失函数如下:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geqslant 1]L_{reg}(t^u, v) \tag{3}$$

式中,L_{cls} 为分类损失函数,L_{reg} 为边界框回归损失函数,p 是分类的预测值,u 是类别的索引,t^u 表示第 u 类边界框预测的修正值,v 是真实标签的修正值。

Faster RCNN 网络对 RPN 输出的目标特征向量使用全连接层进行特征组合,随后分别输入给分类器以及边界框回归器。但是全连接层对于这两种任务并非都合适。边界框回归任务需要更多的目标级别的上下文,卷积结构相较于全连接结构更能提取目标级别上下文,故卷积结构在边界框回归任务上比全连接结构更合适。本文就此提出一种新的检测模型来提升边界框回归的精度。

2　Divide Faster RCNN

本文基于上述的先验知识,将全连接结构应用于目标对象的分类,卷积结构应用于目标对象的定位,提出了可以有效提升目标检测的精度,实现有效的目标定位的改进 Faster RCNN 模型——DF RCNN。图 2 显示了本文的目标检测模型。首先利用骨干网络提取 RGB 空间下输入图像的多尺度特征,

并将多尺度特征构建成特征金字塔，实现多尺度特征映射图的构建。然后利用 Anchor 机制以及共享特征映射方式实现候选区域生成，对提出的候选目标区域使用感兴趣区域对齐方法（region of interest align，ROI Align）[24]将提出的候选区域映射至对应尺度的特征图上，最后将选择的特征输入基于全连接结构的分类网络和基于卷积结构的回归网络实现目标的类别确定以及目标的定位，得到目标检测的结果。

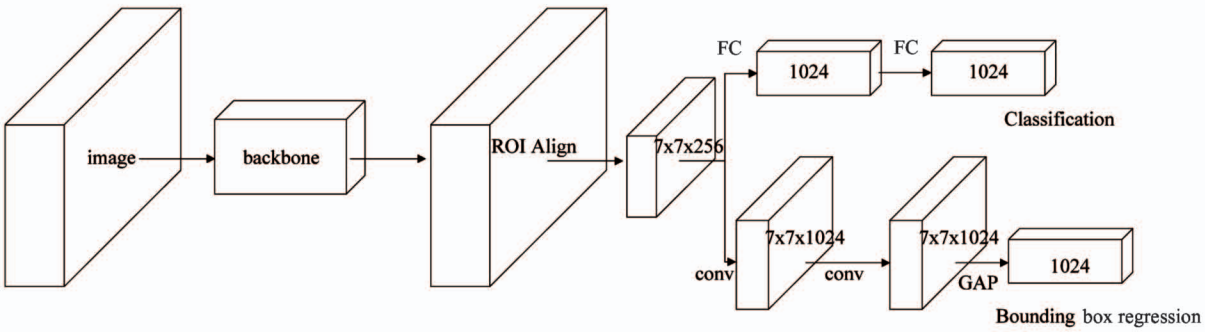


图 2　Divide Faster RCNN 网络结构

2.1　Divide Faster RCNN 的主干网络

DF RCNN 首先使用残差学习框架作为主干网络。残差结构是在标准的前馈卷积网络的基础上添加一个残差连接（skip connect）以便跳过一些层的连接，每完成一次 skip connect 便产生一个残差块，如图 3 所示。其中 Conv 表示卷积模块用于对输入进行卷积处理；BN 为批归一化操作，减小了网络内部协方差平移，加速深层网络训练；ReLu 为网络的激活函数。残差块的数学表达如下：

$$y = F(x, \{W_i\}) + x \tag{4}$$

式中，y 表示残差块的输出，x 是残差块的输入，$F(x, \{W_i\})$ 表示待学习的残差映射。

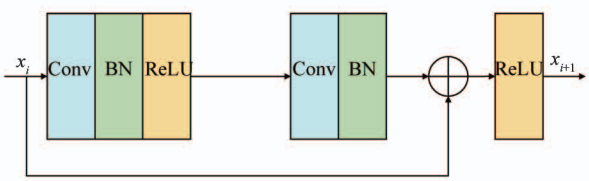


图 3　残差块结构

残差网络的 skip connect 方式降低了模型随着网络深度增加发生过拟合的可能性，使模型的深度可以构建得更深。一般而言，卷积层越深，高层语义越明显，更利于分类任务，但同时损失的信息也越多。该级别的损失对于大尺度目标的影响不大，但是对小尺度目标会造成严重的影响。这就是 Faster RCNN 对小物体检测效果相对较差的原因。特征金字塔（feature pyramid network，FPN）[25]提取不同层级卷积层的输出，可以同时捕捉浅层与高层的语义信息，提升了对小目标的检测能力。以残差网络为例，将残差网络中 *conv*2 、*conv*3 、*conv*4 、*conv*5 的输出分别用 C_2 、C_3 、C_4 、C_5 表示。FPN 首先对 C_5 使用一个 1×1 的降维层将其通道数从 2048 降为 256 得到特征映射图 CP_5，之后对 CP_5 进行上采样，同时对 C_4 使用一个降维层将其通道数从 1024 降至 256，采用 *Add* 函数将两者结合得到 CP_4。之后对 CP_4 进行上采样再 *Add* 上 C_3 经过降维层结果得到特征映射图 CP_3。对 CP_3 进行上采样再 *Add* 上 C_2 得到 CP_2。最后分别对 CP_2 、CP_3 、CP_4 、CP_5 使用一个 3×3 的卷积以降低上采样带来的混叠现象，最终得到 P_2 、P_3 、P_4 、P_5，构成了特征金字塔。

2.2　区域提议

目标候选区域推荐的数目和质量直接关系到目标检测的速度和精度。本文采用 RPN 方法进行目标候选区域的提议。RPN 通过 Anchor 机制直接在特征映射图上生成候选区域。RPN 共享了模型特征金字塔部分的卷积特征，降低了整体计算的冗余，有效提升了模型的检测速度。

本文对 RPN 的部分参数进行了调整，将 Anchors 的基础尺寸调节为（32，64，128，256），长宽比调整为（1，0.5，2），总共得到 12 种不同大小的 An-

chors。本文采用 NMS 方法,通过比较 Anchors 与 Ground Truth 之间的 IOU,删除 IOU 在 0.3 ~ 0.5 之间的 Anchors,同时将 IOU ≥ 0.5 的 Anchors 作为正样本,IOU < 0.3 的 Anchors 作为负样本。最后根据 Anchors 的得分大小,选取得分最高的 200 个 Anchors 作为最终的候选框。

2.3 ROI Align

ROI Pooling 可以从各个 ROI 中对特征进行进一步的提取,但是 ROI Pooling 层引入了两次量化操作,降低了 ROI 与其特征之间的一致性。同时给边界框回归带来了算法本身的偏移。ROI Align 采用双线性插值法,在 ROI 上先进行分割,然后在分割得到的每一块小区域中,采样 K (K 一般取 4)个点,即对该小区域进行等分得到 4 个子区域,之后每个子区域利用双线性插值得到中心点的像素值,之后取这 4 个子区域中心像素值的最大值作为这个小区域的像素值。该方法有效地避免了 ROI Pooling 中的量化过程,有利于 ROI 和特征之间的一致性,提升分类精度,避免量化带来的边界框误差。经过 ROI Align 处理的特征图的大小被固定为 7 × 7 × 256。

由于特征金字塔输出了多个不同层级的特征图,需要根据特征图尺度的不同来选择不同层级的金字塔的输出,具体选择应用如下公式。

$$k = k_0 + \log_2\left(\frac{\sqrt{wh}}{224}\right) \tag{5}$$

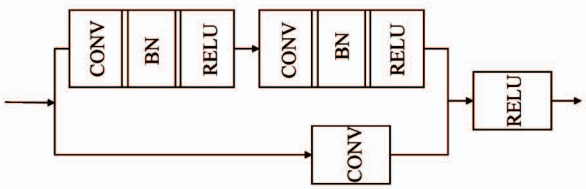
式中,$k_0 = 5$,w 和 h 表示对应 ROI 区域的宽和高,224 对应特征金字塔第 5 层的尺度。

2.4 分裂检测网络

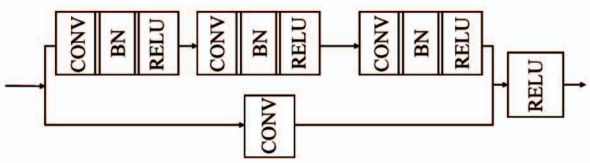
与其他主流两阶段方法不同,DF RCNN 将分类与回归问题分开,提出将目标检测中的分类任务与边界框回归任务分别传输给全连接结构与卷积结构以完成目标对象的准确分类以及精确定位。

全连接结构将分布式表示特征映射到样本标签空间,聚合了卷积网络提取的多种不同特征,降低了特征空间位置对分类的影响。本文采用的全连接结构包含 2 层全连接层,每层的神经节点数目为 1024,第 1 层全连接层将 ROI Align 输出的 7 × 7 × 256 张量降维到 1 × 1024,第 2 层全连接层的神经节点数目也是 1024。两层全连接层大大增加了网络的非线性能力,提升了模型的复杂度,增加了模型对复杂图像的识别能力。

全连接结构聚合的特性使其忽略了特征出现的空间信息,因此不适合边界框的回归。卷积网络对空间信息保留能力以及目标级别上下文提取能力更强,更适合边界框回归任务。卷积结构采用堆叠的子块构建,如图 4 所示。卷积子块 1 首先利用 3 × 3 的卷积对 ROI Align 的输出进行进一步的特征提取,后续 1 × 1 的卷积核将输入的通道数从 256 提升到 1024,增加了特征的丰富性。卷积子块 2 利用了瓶颈块结构,使用一个 1 × 1 的卷积核将输入的维度先降低到 256,再利用 3 × 3 的卷积进一步提取特征,之后通过 1 × 1 的卷积将输出的维度恢复为 1024。因为全局平均池化层 (global average pooling, GAP)[26] 不像 FC 层需要大量训练调优参数,且抗过拟合能力更强,故本文采用 GAP 层代替 FC 层,实现对卷积结构输出向量的进一步降维,防止过拟合。



(a) 卷积子块 1



(b) 卷积子块 2

图 4 构成卷积结构的子块

2.5 网络的整体训练

DF RCNN 的损失函数分为分裂网络的损失函数以及 RPN 网络的损失函数。将其写为一个总的损失函数进行训练,其总损失函数为

$$L_{total} = w_{fc}L_{fc} + w_{conv}L_{conv} + L_{rpn} \tag{6}$$

式中 L_{fc}、L_{conv} 和 L_{rpn} 分别为全连接结构、卷积结构以及 RPN 网络的损失函数,w_{fc} 和 w_{conv} 是 L_{fc} 和 L_{conv} 的

权重平衡系数。本文采用带动量的随机梯度下降法(SGD)进行网络的权重更新,利用动量可以有效避免损失函数陷入局部最优值且加快整体收敛的速度。

利用迁移学习的方法,首先利用在 COCO 数据集上预训练的参数对模型进行初始化,之后冻结网络中的部分参数,使用目标数据集进行对模型剩余部分的参数微调使其在更短时间内有效地适应目标任务。

3 实验

3.1 数据集

本文首先使用 Pascal VOC 2007 和 2012 数据集进行一般性测试。Pascal VOC 2007 和 2012 数据集包含了 20 种不同物体。训练数据涵盖了各种场景的图片,每张图片都有对应的标签文件。然后本文自主构建了吸尘器尘袋图像集。数据集分为 2 部分,数据集 1 和数据集 2。数据集 1 包含 8 种不同圆形开口的吸尘器尘袋,数据集 2 包含 2 种不同椭圆形开口的吸尘器尘袋。表 1 和表 2 给出了数据集 1 和数据集 2 的具体信息。图 5 和图 6 分别给出了对应数据集的部分示例图片。

3.2 评价标准

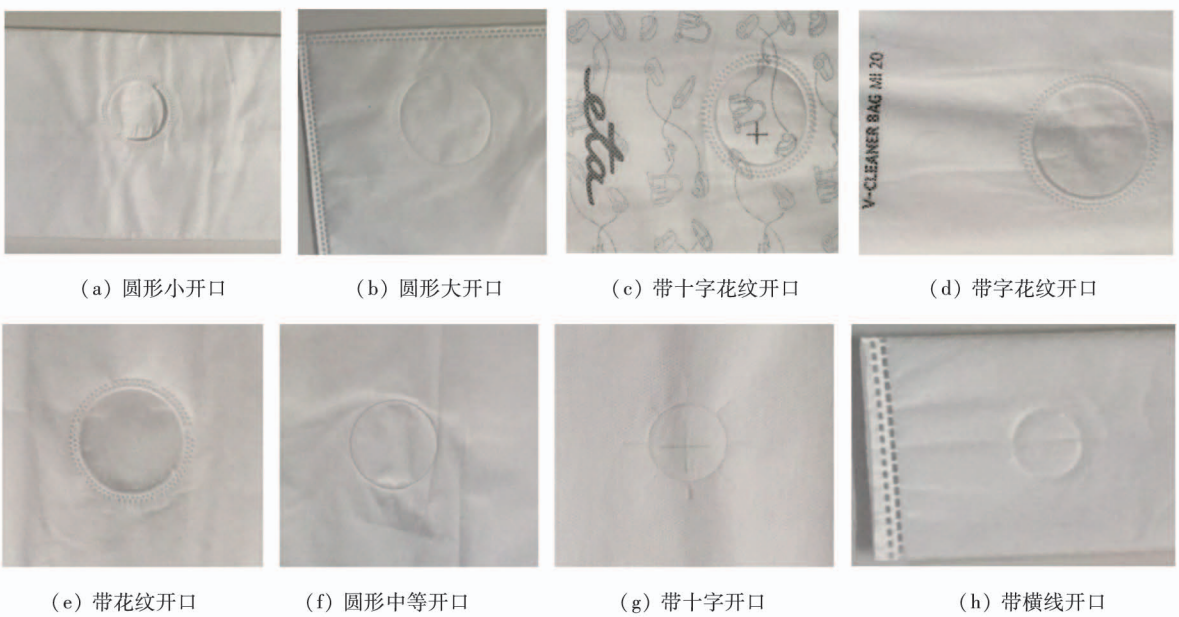
本文采用各类平均精度(mean average precision, mAP)以及中心偏移距离作为主要评价指标,同时使用收敛速度以及目标检测速度作为辅助评价指标,并与 Faster RCNN、Mask RCNN 两种主流的两阶段目标检测模型作比较。

表 1 数据集 1 中每一类包含的图片数目

类别	圆型小开口	圆形大开口	带十字花纹开口	带字花纹开口	带花纹开口	圆形中等开口	带十字开口	带横线开口
数目	75	75	73	73	77	75	77	75

表 2 数据集 2 中每一类包含的图片数目

类别	跑道形椭圆	花纹椭圆
数目	76	74



(a) 圆形小开口 (b) 圆形大开口 (c) 带十字花纹开口 (d) 带字花纹开口

(e) 带花纹开口 (f) 圆形中等开口 (g) 带十字开口 (h) 带横线开口

图 5 数据集 1 图片示例

(a) 跑道形椭圆

(b) 花纹椭圆

图 6　数据集 2 图片示例

GT 的边框记作 box_{gt}，格式为 $(y1_{gt}, x1_{gt}, y2_{gt}, x2_{gt})$；预测出的边界框记作 box_{pred}，其格式与 box_{gt} 相同为 $(y1_{pred}, x1_{pred}, y2_{pred}, x2_{pred})$。两者之间的距离记作 $d = dis(box_{gt}, box_{pred})$，即为预测边界框的中心偏移距离。

$$d = dis(box_{gt}, box_{pred}) = \sqrt{(y_{gt} - y_{pred})^2 + (x_{gt} - x_{pred})^2} \tag{7}$$

其中，$y_{gt} = (y1_{gt} + y2_{gt})/2$，$x_{gt} = (x1_{gt} + x2_{gt})/2$，$y_{pred} = (y1_{pred} + y2_{pred})/2$，$x_{pred} = (x1_{pred} + x2_{pred})/2$。

本文的实验总共分为 3 部分。第 1 部分采用 Pascal VOC 2007 数据集中的 Pascal VOC 2007trainval 子数据集以及 Pascal VOC 2012 数据集中的 Pascal VOC 2012trainval 子数据集进行模型评估。第 2 及第 3 部分采用吸尘器尘袋数据集 1 和 2 来评估模型。

3.3　参数设置

本文算法采用 Python 实现，版本为 3.6，基于 TensorFlow1.12 框架。实验设备为 i5-9400，16 GB 内存，GTX1060(6 GB)的个人计算机。实验主要分为以下几个部分，第 1 个实验比较了在一般性数据集

上 Divide Faster RCNN 网络与 Faster RCNN 之间的性能差异，验证 Divide Faster RCNN 在通用检测方面的有效性能。第 2 个实验用于比较 Divide Faster RCNN 网络与 Faster RCNN 网络在吸尘器尘袋数据集 1 上的表现，验证 Divide Faster RCNN 对边界框回归精度的提升。在实验中，由于实验设备的限制，将 batch size 设置为 1，训练步数设置为 200，总共训练 100 个周期。第 3 个实验用于验证算法的性能，检验增加类别后算法的鲁棒性，在原始数据集 1 上增加扩展数据集 2 作为该实验的数据集。实验中 Faster RCNN 算法采用文献[10]中的默认配置，Divide Faster RCNN 采用带动量的 SGD 算法，动量设置为 0.9，学习率初始化为 0.0001，输入图像的大小在第 1 个实验缩放为 512×512，第 2 个和第 3 个实验缩放为 1024×1024。

3.4　算法比较

实验 1　表 3 给出了本文方法与现有主流两阶段目标检测算法分别在 Pascal VOC2007trainval 和 2012trainval 上的测试结果。实验表明，在 Faster RCNN 中加入特征金字塔且对不同任务使用不同网络头结构之后，各类平均精度分别提高了 3.9% 和 5.7%，实验表明了本文方法在一般性问题中优于 Faster RCNN。

实验 2　本文采用端到端的训练方法在吸尘器尘袋数据集 1 上完成模型的训练。并将其与现有主流两阶段目标检测算法进行比较。表 4 展示了在尘

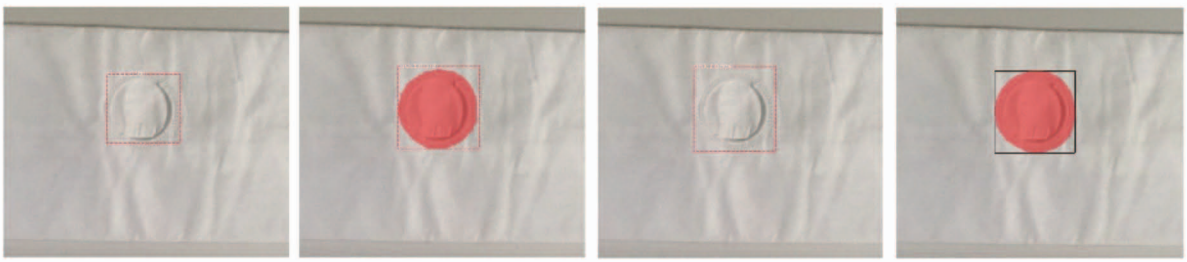
表 3　在 Pascal VOC2007trainval 和 2012trainval 上的测试结果

	Backbone	Anchor boxes	mAP/% (VOC2007trainval)	mAP/% (VOC2012trainval)
Faster RCNN	VGG16	12	68.2	65.4
Faster RCNN	Resnet50	12	69.7	66.8
Faster RCNN + FPN	Resnet50	12	70.5	68.1
本文算法	Resnet50	12	73.6	72.5

表 4 在吸尘器尘袋测试集上的检测准确率以及中心偏移距离

	Backbone	Anchor boxes	mAP/%	中心偏移距离/像素	fps
Faster RCNN	VGG16	12	90	24.34	1.12
Faster RCNN	Resnet50	12	92	20.62	1.23
Mask RCNN	Resnet50	12	94	12.72	0.95
本文算法	Resnet50	12	94	11.76	1.02

袋数据集 1 上本文算法与现有模型之间的定量比较结果,其中 backbone 为模型采用的特征提取网络的结构,Anchor Boxes 为在 RPN 网络中设置的 Anchors 的种类数目,fps 表示每秒可以完成检测的图像的数目,中心偏移距离的单位是像素。图 7 给出了部分的检测结果。



(a) Faster RCNN 检测结果 (b) Mask RCNN 检测结果 (c) 本文算法检测结果 (d) Ground Truth

图 7 在数据集 1 上检测的结果(示例)

从表 4 中可以看出,本文模型在中心偏移距离方面优于现有的模型。Faster RCNN 由于 ROI Pool 的缘故,在中心偏移距离方面表现最差。Mask RCNN 引入了 ROI Align 以及掩码(mask)分支,目标定位能力得到了改进,中心偏移距离比 Faster RCNN 提高了 7.9 个像素。本文的模型比 Mask RCNN 提高了 0.96 个像素。不难看出,去卷积结构相较于全连接结构对于目标定位精度确实有提升。分裂机制将卷积结构作为边界框回归的特征降维器,提高了特征的空间信息保留,提高了目标的定位精度。

从表 4 中可以看出,本文模型的分类精度要优于 Faster RCNN 模型。本文模型采用的 backbone 为残差网络,残差网络对特征的提取能力优于 VGG16,更强的特征保留带来了 mAP 的提升。同时本文算法构建了特征金字塔,提高了对小目标的检测精度,降低了模型的漏报率,提高了模型的召回率,在相同精确率情况下,更高的召回率带来了 mAP 的提升。特征金字塔的使用,使得本文模型比 Faster RCNN 分类性能更强。

在检测速度方面,本文模型由于引入了分裂机制,增加了模型的复杂度,相较于结合 FPN 技术的 Faster RCNN,本文增加了全卷积结构,单层卷积的计算复杂度为 $O(H_{out} \times W_{out} \times (K^2 \times C_{in} + 1) \times C_{out})$,其中 H_{out} 与 W_{out} 分别为输出特征图的高度和宽度,C_{in} 和 C_{out} 分别为卷积层输入及输出的通道数,K^2 代表了卷积核的大小。对于多层卷积网络而言其时间复杂度是多个卷积层的累加 $O(\sum_{l=1}^{D} H_{out}^{l} \times W_{out}^{l} \times ((K^l)^2 \times C_{in}^{l} + 1) \times C_{out}^{l})$,其中 l 表示当前层的索引,D 为网络的深度。为了减少计算量,本文采用瓶颈层方式,先利用 1×1 卷积将输入通道数降低,对降低之后的特征图进行进一步卷积,最后通过 1×1 卷积将输出通道数重新升高回预设维度。模型的检测速度还是优于 Mask RCNN 但略逊于 Faster RCNN。同时,由于增加了卷积结构,致使本文模型在训练过程中 loss 的下降速度稍微变慢,但下降趋势基本没有变化,如图 8 所示。

实验 3 为了排除数据集中和数据不足带来的干扰,本文进行了第 2 组实验,训练集采用数据集 1 和数据集 2,增加了数据的多样性。实验结果如表 5 所示,从中可以看到 Divide Faster RCNN 模型的 mAP 为 0.94,中心偏移距离为 12.46,Faster RCNN

的 mAP 为 0.88,中心偏移距离为 26.63,实验结果证明了本文模型在类别增加的情况下效果依旧比 Faster RCNN 好。同时,图 9 给出了 DF RCNN 对各种类型吸尘器尘袋的检测结果。从图中可以看到本文模型对于各种类别的吸尘器尘袋都可以实现袋口的有效检测,证明了其在尘袋袋口检测任务上的有

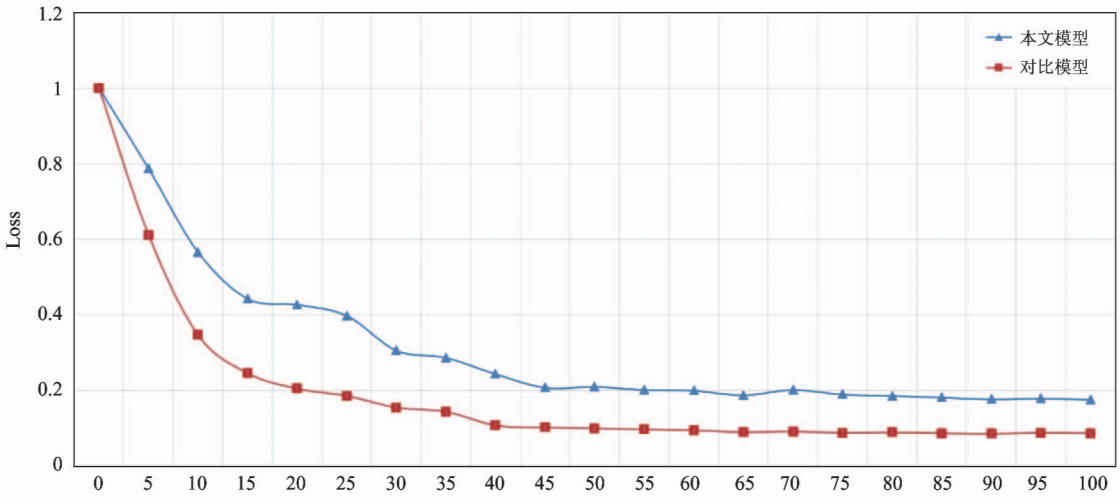
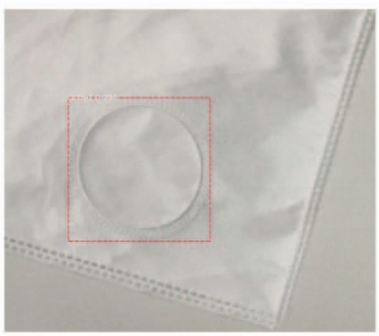


图 8　Divide Faster RCNN 与对比算法的 loss 曲线

表 5　在数据集 1 +2 上的检测结果

	Backbone	Anchor boxes	mAP/%	中心偏移距离	fps
Faster RCNN	Resnet50	12	88	26.63	1.04
Mask RCNN	Resnet50	12	92	16.46	0.93
本文算法	Resnet50	12	94	12.46	0.97

(a) 跑道形椭圆



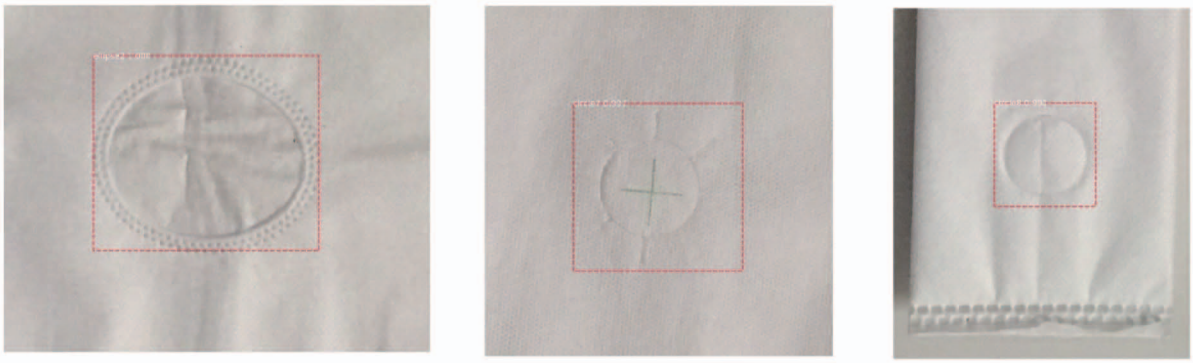
(b) 圆形大开口

(c) 带十字花纹开口

(d) 带十字花纹开口

(e) 带字花纹开口

(f) 带花纹开口



(g) 圆形中等开口　　(h) 花纹椭圆　　(i) 带十字开口

图 9　DF RCNN 在吸尘器尘袋数据集 1 +2 的检测效果

效性。为了进一步测试其检测能力,将来自两个数据集的吸尘器尘袋原始对象拍摄在同一张图片中,并采用训练好的 Divide Faster RCNN 进行检测,结果如图 10 所示。以上结果共同证明了本文模型可以有效地检测出不同类别的吸尘器尘袋并且对其开口进行精准的定位,同时对类别的增加具有良好的鲁棒性。

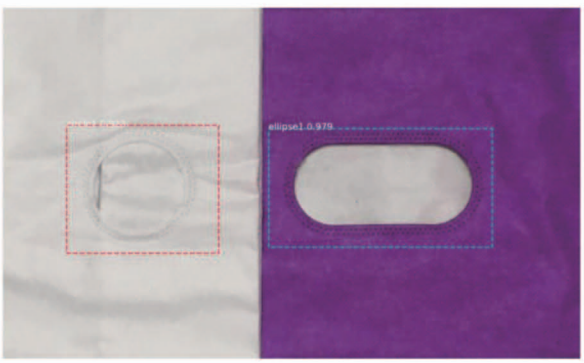


图 10　在同一图片中两种不同尘袋的检测效果

4　结　论

为了有效提升目标检测的定位精度,本文提出了一种采用分裂机制的基于 Faster RCNN 的目标检测模型。不同于一般基于 Faster RCNN 模型将 ROI 特征经过全连接网络同时输入给分类与边界框回归网络,本文利用卷积结构与全连接结构在分类与定位方面不同性能差异,将 ROI 特征分别输入给使用全连接结构的分类网络以及使用卷积结构的边界框回归网络。引入卷积结构,使得 ROI 特征中的空间信息被更充分地利用。采用 ROI Align 替代 ROI Pooling 取消了 ROI Pooling 中的量化,消除了 Faster RCNN 模型本身的边界框偏差。实验结果表明,本文提出的模型在目标检测效果与精度两方面都优于现有算法,不仅实现了目标的有效识别,而且定位精度也比 Faster RCNN 高。在今后工作中,将寻求更好的优化策略,进一步提升检测速度以及泛化能力。

参考文献

[1] 张泽苗,霍欢,赵逢禹. 深层卷积神经网络的目标检测算法综述[J]. 小型微型计算机系统,2019, 40(9): 1825-1831

[2] Schmidhuber J. Deep learning in neural networks: an overview[J]. *Neural Network*, 2015, 61(8): 85-117

[3] 许必宵,宫婧,孙知信. 基于卷积神经网络的目标检测模型综述[J]. 计算机技术与发展,2019, 29(12): 87-92

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

[5] Lowe D. Distinctive image features from scale-invariant key points[J]. *International Journal of Computer Vision*, 2003, 20:91-110

[6] Felzenszwalb P, Mcallester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008:1-8

[7] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]// Proceedings of 13th International Conference on International Conference on Machine Learning, Bari, Italy, 1996:148-156

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

[9] Girshick R. Fast R-CNN[C]//IEEE International Con-

ference on Computer Vision, New York, USA, 2015: 1440-1448

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 36(6): 1137-1149

[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 39(4): 640-651

[12] Uijlings J R R, K E A van de Sande. Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, 104(2): 154-171

[13] Zitnick C L, Dollar P. Edge Boxes: locating object proposals from edges[C]//Proceedings of European Conference on Computer Vision, Zurich, Switzerland, 2014: 391-405

[14] Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Conference on Neural Information Processing Systems, Las Vegas, USA, 2016:379-387

[15] Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Patter Recognition, Salt Lake, USA, 2018: 798-802

[16] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Computer Vision and Pattern Recognition, New York, USA, 2016: 779-788

[17] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016:21-37

[18] Zhang S F, Wen L Y, Lei Z, et al. RefineDet++: single-shot refinement neural network for object detection [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 31(2): 674-687

[19] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 7263-7271

[20] Redmon J, Farhadi A. YOLOv3: an incremental improvement[J]. *arXiv*:1804.02767v1, 2018

[21] 蒋弘毅,王永娟,康锦煜. 目标检测模型及其优化方法综述[J]. 自动化学报, 2020, doi:10.16383/j.aas.c190756

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv*:1409.1556, 2014

[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 and Machine Intelligence*, 2015, 37(9): 1904-1916

[24] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 386-397

[25] Lin T Y, Dollar P, R. Girshick, He K et al. Feature pyramid networks for object detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dhaka, Bangladesh, 2017:1-10

[26] Lin M, Chen Q, Yan S. Network in network[J]. *arXiv*: 1312.4400v3, 2013

Target detection method based on improved Faster RCNN

Wang Xianbao, Zhu Xiaoyong, Yao Minghai

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310000)

Abstract

For the region-based target detection algorithm, there is a general problem that the boundary position is far from the real value. This paper proposes an improved Faster RCNN algorithm using the splitting mechanism. Firstly, the algorithm selects the convolutional neural network(CNN) with strong feature extraction ability as the backbone network to extract features, and then generates candidate target regions through 12 different anchors to further improve the accuracy of detection. Finally, the obtained features are transmitted to two different sub-networks: the classification network is based on the fully-connected structure, and the targets are classified; the positioning network is based on the convolutional neural network structure to achieve the target positioning. The experiments verify the effectiveness of the algorithm on the Pascal VOC2007 dataset, Pascal VOC2012 dataset and vacuum bag dataset. The results show that the proposed algorithm is more accurate than the Faster RCNN in the effective detection of the target, and achieves the accurate regression of the bounding box.

Key words: target detection, convolutional neural network(CNN), positioning accuracy, improved Faster RCNN, dividing mechanism