一种改进的 YOLO V3 目标检测方法

徐 融,邱晓晖
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:目标检测是当今计算机视觉领域较为热门和流行的研究方向,在国防、安全和医疗保障等领域应用广泛。然而小目标的检测准确度一直不高,针对这一问题,提出了一种基于 YOLO V3 网络模型的改进方法,通过增强小目标的检测准确度来提高网络整体的检测成功率。由于小目标在图像中所占像素很少,经过多层卷积之后提取得到的特征不明显。改进方法通过将原网络模型中经 2 倍降采样的特征图进行卷积分别叠加到第二及第三个残差块的输入端,以此增强浅层特征信息。同时,在第一个 8 倍降采样的特征图后连接 RFB 模块,增强特征提取能力。用改进后的网络模型在 PASCAL VOC 数据集上与原网络进行对比实验。结果表明,改进之后的网络模型有效提高了小目标的检测准确率。

关键词:YOLO V3;目标检测;深度学习;RFB;PASCAL VOC
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)07-0030-04
doi:10.3969/j.issn.1673-629X.2020.07.007

An Improved YOLO V3 Object Detection

XU Rong,QIU Xiao-hui
(School of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Object detection is a hot and popular research direction in the field of computer vision,which is widely used in the fields of national defense,security and medical security. However,the detection accuracy of small targets is not high. To solve this problem,we propose an improved method based on YOLO V3 network model,which improves the detection accuracy of the whole network by enhancing the detection accuracy of small targets. Because the small target occupies very few pixels in the image,the features extracted after the multi-layer convolution are not obvious. The improved method enhances the shallow feature information by convolving the feature maps of the original network model by two times down sampling onto the input ends of the second and third residual blocks respectively. In addition,the RFB module is connected after the first 8 times down sampling feature map to enhance the feature extraction ability. The improved network is compared with the original network on the PASCAL VOC data set. The results show that the improved network effectively improves the detection accuracy of small targets.

Key words:YOLO V3;object detection;deep learning;RFB;PASCAL VOC

0 引 言

目标检测(object detection)是计算机视觉领域的基本任务之一,在学术界已有二十多年的研究历史[1]。传统的目标检测算法首先要在给定的图像上进行区域选择(滑窗),然后对这些区域进行特征提取,最后使用训练好的分类器进行分类[2]。这类方法使用手工设计的特征,鲁棒性差,过程复杂。

近年来,随着卷积神经网络的发展,深度学习被广泛应用于目标检测。与传统的目标检测方法相比,使用深度学习进行检测具有很多优势。例如传统方法需要研究人员利用相关知识及经验手动提取特征,基于深度学习的方法可以通过大量数据学习相应数据差异的特征,并且所得到的特征更具代表性。深度学习模型通过模拟人脑的视觉感知系统,直接从原始图像中提取特征,并逐层传递,以获得图像的高维信息。目前优秀的深度学习模型大致可以分为两类:第一类模型将目标检测分为两步(two stage)进行,如 R-CN[3]、SPP-Net[4]、Fast-RCNN[5]、Faster-RCNN[6]等,这类算法首先从目标图像的区域候选框中提取目标信息,然后利用检测网络对候选框中的目标进行位置的预测以

收稿日期:2019-08-19 **修回日期**:2019-12-20
基金项目:江苏省自然科学基金(BK2011789)
作者简介:徐 融(1995-),男,硕士研究生,研究方向为图像处理与模式识别;邱晓晖,博士,教授,通信作者,研究方向为信号与信息处理及模式识别。

及类别的识别;第二类模型则是基于端到端(one stage)进行的,如 SSD[7]、YOLO[8-9]等,这类方法不需要从图像中预先提取候选网络,而是直接对图像中的目标进行位置的预测以及类别的识别。因此,第二类网络比第一类网络具有较快的检测速度。

为了提高网络目标检测的精度,文中以 YOLO V3[10]为基础,在 PASCAL VOC 数据集上进行训练和测试。首先对 YOLO V3 的网络结构进行改进,将经过 2 倍降采样的特征图进行卷积,再分别添加到第二及第三个残差块的输入端,最大化利用浅层特征信息。此外,在 8 倍降采样的特征图后连接 RFB(receptive field block)模块[11]来融合不同尺寸的特征。

1　YOLO V3 网络模型

YOLO,即 You Only Look Once 的缩写,是一个基于卷积神经网络(CNN)的目标检测算法。YOLO V3 使用维度聚类得到的锚框来预测边界框,每个边界框预测 4 个坐标:边界框的中心坐标以及边界框的宽和高。其使用逻辑回归预测每个边界框的类别得分,并使用均方和误差作为损失函数。通过置信度来表示边界框含有目标的可能性大小。如果某个先验边界框与真实对象重叠超过任何其他边界框,则该值置为 1。如果边界框的优先级不是最高但是与真实对象重叠超过某个阈值,那么该值置为 0。YOLO V3 使用 DarkNet53 网络进行特征提取,其网络结构如图 1 所示。

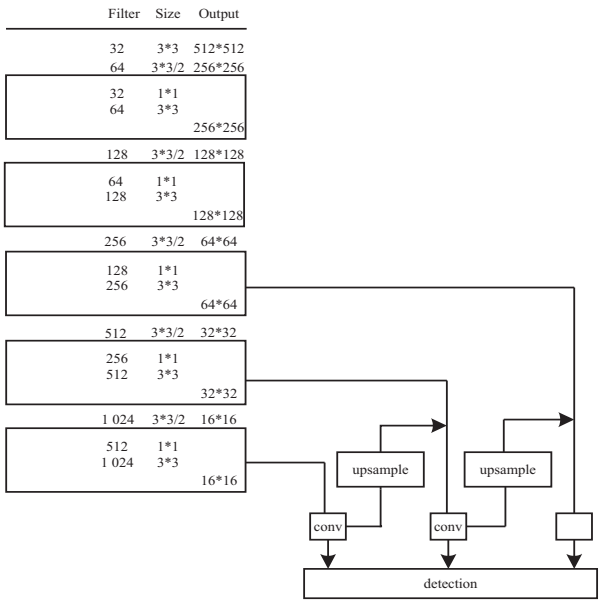
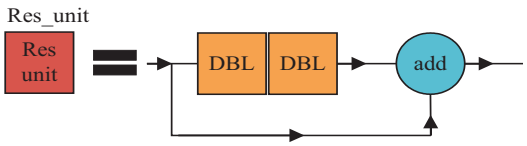
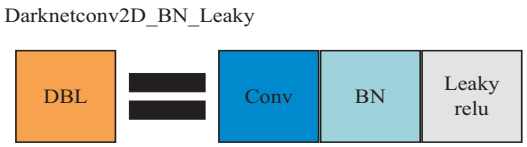


图 1　YOLO V3 网络结构

DarkNet53 融合了 ResNet[12],共包含 5 个残差块,每个残差块由数量不等的残差单元组成,每个残差单元又由两个 DBL(Darknetconv2d_BN_Leaky)单元及残差操作构成[13],如图 2(a)所示。其中,每个 DBL 单元又是由卷积层、归一化(batch normalization)[14]和激活函数(leaky relu)组成,如图 2(b)所示。残差块的使用既可以防止有效信息的丢失,也能够防止深层网络训练时出现梯度消失[15]。除此之外,该网络中没有池化层,它使用步长为 2 的卷积做下采样来代替池化操作,进一步防止有效信息的丢失,这对小目标来说是十分有利的。



(a)残差单元



(b)DBL 单元

图 2　残差单元构成

YOLO V3 网络使用均方和误差作为损失函数,其由三部分组成,分别预测框定位误差、有无目标的 IOU 误差以及分类误差。损失函数 loss 如下所示:

$$\begin{aligned}\text{loss} = {} & \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{J=0}^{B} 1_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] + \\ & \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{J=0}^{B} 1_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + \right. \\ & \left. (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] + \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2 + \\ & \lambda_{\text{noord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2 + \\ & \sum_{i=0}^{s^2} 1_{ij}^{\text{obj}} \sum_{c \in \text{classes}} \left[p_i(c) - \hat{p}_i(c) \right]^2\end{aligned}$$

其中,第一项和第二项表示预测框的定位误差,λ_{coord} 表示中心坐标误差的权重,设为 5;S 表示图像被划分成的网格数;B 表示每个网格所预测的框的个数;1_{ij}^{obj} 表示第 i 个网格的第 j 个预测框是否检测到了目标;x_i,y_i,w_i,h_i 分别表示真实框的中心坐标以及宽和高(带宝盖帽子的表示预测框对应的值)。第三项和第四项表示 IOU 误差,c 表示置信度得分;最后一项表示分类误差,$p_i(c)$ 表示检测到的目标属于 C 的条件概率。

2　改进的 YOLO V3 网络

2.1　数据集聚类分析

原 YOLO V3 网络是通过对 COCO 数据集的聚类来生成 9 个锚框,每个尺寸的特征图分别对应 3 个锚框。网络在训练阶段,需要计算真实框与哪个锚框的 IOU 最大,标记该锚框对应的置信度为 1。在计算 loss

时,这个锚框对应的预测有回归、置信度和分类的误差,大于某个阈值但不是最优的锚框对应的预测值则没有置信度和定位损失,小于阈值的则有置信度损失。需要说明的是,训练时预测的值为高和宽相对于锚框高和宽的值。在测试阶段,则根据置信度与阈值的关系来判断预测的边框是否有效,这时锚框的作用就是还原预测边框在输入图像中的大小。而文中采用的是 PASCAL VOC 数据集,所以需要重新进行聚类。

2.2 改进的 YOLO V3 模型

YOLO V3 网络中采用特征金字塔来增强检测效果,输出的特征图分别经过了 8 倍、16 倍、32 倍的降采样,也就是说当被检测目标不足 8 pixel×8 pixel 时,最后在输出的特征图上将很难检测到它。为了使更多的小目标信息得以更充分地利用,文中将经过一次降采样的特征图叠加到第二及第三个残差块的输入端。此外,在 52×52 的特征图后连接 RFB(RF Block)模块。RFB 模块通过模拟人类视觉的感受野结构(receptive fields,RFs)来加强网络的特征提取功能。特征图首先通过由不同尺寸卷积核构成的多分支结构,然后再经过空洞卷积层增加感受野,最后将不同尺寸的卷积层输出进行 concat 操作,从而达到融合不同特征的目的。连接结构如图 3 所示。

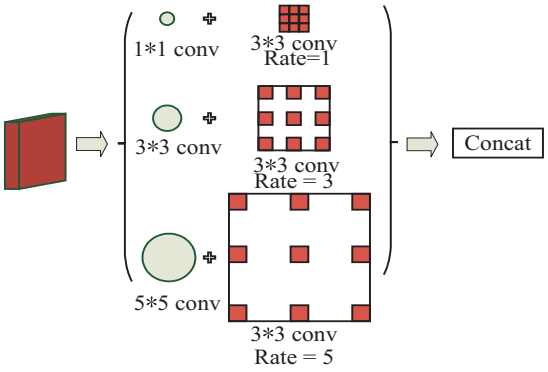


图 3 RFB 连接结构

3 实验结果

文中利用 PASCAL VOC 2007+2012 数据集中的训练图片对改进后的网络进行训练,使用 PASCAL VOC 2012 中的测试图片进行测试。采用平均精确值(mean average precision,mAP)作为评价指标,与原网络进行比较实验。

为了验证提出的改进网络的有效性,在 PC 机上进行了实验,机器配置如下:Window10 操作系统,Intel Core i7-8750H 处理器,NVIDIA GTX1060 独立显卡,6 G 显存,8 G 内存。采用 Tensorflow 框架进行训练,时长为 24 小时。

3.1 实验步骤

(1)训练算法:每次训练随机选取 8 张图片,初始学习率为 1e-4,且逐步递减,但不小于 1e-6,IOU 置为 0.5;

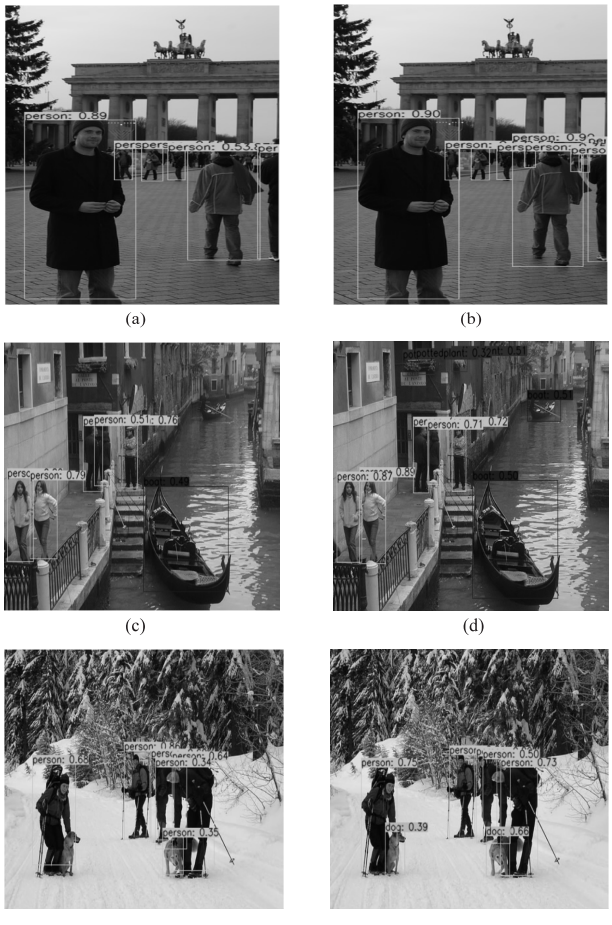
(2)网络参数微调:采用反向传播对网络参数进行微调。前 20 个 epoch 先对最后一层网络参数进行优化,后 30 个 epoch 对整个网络的网络参数进行调整;

(3)训练数据:采用 PASCAL VOC 数据集进行训练和测试,其中将 VOC2007 与 VOC2012 的 trainval 文件夹下的图片(共 12 000 多张)作为训练集;测试集选用 VOC2007 的测试集。

3.2 实验结果分析

为了验证改进之后网络模型检测的准确率,分别选取原网络第 60 000 次训练得到的权重和改进网络第 50 000 次训练得到的权重进行对比实验。改进之前原网络的 mAP 为 80.26%,改进之后网络的 mAP 为 81.35%,提升了 1.09%。

图 4 给出了三组分别基于原 YOLO V3 网络及改进后网络的实验结果。



((a)、(c)、(e)为原网络的检测结果,(b)、(d)、(f)为改进后网络的检测结果)

图 4 YOLO V3 网络改进前后的实验结果对比

从前两组对比图中可以看到:图 4(a)、图 4(c)中出现了漏检的情况,图 4(b)、图 4(d)检测出了更多的

小目标;图 4(e)中不仅出现了漏检的情况,而且将 dog 类检测为 person 类,图 4(f)则检测正确。由此可以看出,改进后的网络对小目标有更好的检测效果,漏检率更低。

4　结束语

提出了一种改进型 YOLO 网络,通过将经过 2 倍降采样的特征图进行卷积,再分别叠加到第二及第三个残差块的输入端的方法来增强浅层特征的利用;通过在 8 倍降采样的特征图后增加一个 RFB 模块,通过模拟人类视觉的感受野加强网络的特征提取能力。实验结果表明,改进后的网络具有更好的检测效果。

参考文献:

[1] 陈　聪,杨　忠,宋佳蓉,等. 一种改进的卷积神经网络行人识别方法[J]. 应用科技,2019,46(3):51-57.

[2] SABZMEYDANI P, MORI G. Detecting pedestrians by learning shapelet features[C]//IEEE conference on computer vision and pattern recognition. Minneapolis: IEEE, 2007: 1-8.

[3] 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. USA: IEEE, 2014: 580-587.

[4] 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, 2014, 37(9): 1904-1916.

[5] GIRSHICK R. Fast R-CNN[C]//International conference on computer vision (ICCV). Santiago: IEEE, 2015: 1440-1448.

[6] 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 & Machine Intelligence, 2015, 39(6): 1137-1149.

[7] LIU W, AUGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European conference on computer vision (ECCV). San Francisco: IEEE, 2016: 6517-6525.

[8] REDMON J, DIVVALS S, GRISHICK R, et al. You only look once: unified, real time object detection[C]//IEEE conference on computer vision and pattern recognition. USA: IEEE, 2016: 779-788.

[9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//IEEE conference on computer vision and pattern recognition. USA: IEEE, 2017: 6517-6525.

[10] REDMON J, FARHADI A. YOLOv3: an incremental improvement[C]//IEEE conference on computer vision and pattern recognition. USA: IEEE, 2018: 2311-2314.

[11] LIU S, HUANG D, WANG Y. Receptive field block net for accurate and fast object detection[C]//European conference on computer vision. [s. l.]: [s. n.], 2018: 404-419.

[12] HE K M, ZHANG X Y, REN S Q. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. San Francisco: IEEE, 2015: 770-778.

[13] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 4510-4520.

[14] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shife[C]//International conference on machine learning. USA: ICML, 2015.

[15] 冯　帅,张　龙,贺小慧. 基于 Jetson TK1 和深度卷积神经网络的行人检测[J]. 信息技术,2017(10):62-64.