

文章编号:1007-2780(2022)04-0539-10

引入高斯掩码自注意力模块的 YOLOv3 目标检测方法

孔雅洁^{1,2}, 张 叶^{1*}

(1. 中国科学院 长春光学精密机械与物理研究所 应用光学国家重点实验室, 吉林 长春 130033;
2. 中国科学院大学, 北京 100049)

摘要:基于行车图像的目标检测方法为感知周围的道路环境提供了便宜、有效的解决方案,但同时也对检测效果和检测速度提出了较高要求。本文针对基于深度学习的一阶段目标检测算法 YOLOv3 展开研究,结合自注意力机制,在其网络深层结构中嵌入高斯掩码自注意力模块,缓解卷积操作感受野不足的缺陷,以捕捉更多的全局信息,提高算法的检测效果。实验结果表明,改进后的模型在 MS COCO 2017 数据集上训练结果的 mAP@0.5 达到 56.88%,精度达到 65.31%。与 YOLOv3 相比,mAP@0.5 提高了 2.56%,精度提高了 3.53%。虽然检测速度有一些损失,但有效提高了模型的检测效果,能够更好地支撑辅助驾驶等应用。

关键词:目标检测;YOLOv3 网络;自注意力;辅助驾驶

中图分类号:TP391 文献标识码:A doi:10.37188/CJLCD.2021-0250

YOLOv3 object detection method by introducing Gaussian mask self-attention module

KONG Ya-jie^{1,2}, ZHANG Ye^{1*}

(1. State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics,
Chinese Academy of Sciences, Changchun 130033, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With the video captured in driving, the surroundings can be sensed economically and conveniently by using object detection techniques, but the accuracy and speed of detection requires a lot in such kind of real-time scenes. In this work, a deep learning-based one-stage object detection algorithm called YOLOv3 is studied. Self-attention mechanism is introduced into this method, by embedding Gaussian mask self-attention modules in the high layers of YOLOv3 network. These modules can merge more global information into feature map to improve the accuracy of model. According to the results of experiments, trained on the MS COCO 2017 dataset, the mAP@0.5 and precision of this improved model can reach to

收稿日期:2021-09-30;修订日期:2021-11-27.

基金项目:吉林省与中国科学院科技合作高技术产业化专项基金(No. 2020SYHZ0031)

Supported by Project of the Special Fund of High Technology for Scientific and Technological Cooperation
Between Jilin Province and Chinese Academy of Sciences(No. 2020SYHZ0031)

*通信联系人, E-mail: yolanda@sp. irits. ai

56.88% and 65.31%. Compared with YOLOv3, its mAP@0.5 and precision increase by 2.56% and 3.53%. Although there is a little loss in detection speed, detection accuracy is significantly improved when the method is applied to assisted driving system.

Key words: object detection; yolov3 network; self-attention; assisted driving system

1 引言

随着科技不断进步和社会快速发展,居民生活水平逐渐提高,汽车成为家庭出行的重要交通工具。政府报告资料显示,2016~2020年间,中国汽车年销量均在2 500万辆以上,全国地区汽车保有量也高达2.5亿辆^[1]。汽车的普及给人民生活带来极大便利的同时,也给道路交通运输带来巨大压力,不仅导致交通拥堵,而且增大事故隐患。因此,辅助驾驶、自动驾驶等智能系统的研究应用得到政府、企业、研究机构的广泛关注。驾驶员借助智能驾驶系统,可以及时发现行车危险,有效提升驾驶安全性,降低交通事故发生的几率。在各类智能驾驶系统中,感知环境是第一步,如何快速精准地感知、检测、测量行车过程中出现的各类目标,一直是研究的热点。车载摄像头的拍摄可以获取大量行车图像,而图像数据具有采集成本低、信息丰富等特点,可以提取诸多有用的环境信息以指导驾驶,比如识别前方车辆、行人等。

目标检测作为视觉感知识别的重要任务之一,吸引了许多研究人员的注意,大量的检测算法得到研究和应用。传统的目标检测通过人工提取图像区域特征,并用预先设定的分类器对其进行分类来达到检测的目的。特征描述准确度对于检测结果影响很大,而且结合多个低级图像特征设计的复杂的分类器难以获得很好的性能。随着深度学习的快速发展,算法模型可以自动学习语义,提取图像的深层特征,检测速度和效果都得到很大的提高,已成为目标检测领域的主流方法。基于深度学习的目标检测算法主要分为两类:一类是基于区域选取的二阶段目标检测,算法被分为两步,首先从图像中生成最有可能包含目标的子区域,然后对子区域做图像分类并生成最终的物体边框;另一类是基于回归的一阶段目标检测,可以直接对物体进行分类,生成类别概率和位置坐标。一阶段算法基于卷积特征图

同时计算边界框回归和目标分类问题,避免了计算成本很高的子区域生成过程,因此,相较于二阶段算法,一阶段算法具有检测速度快的优势,更适用于辅助驾驶、自动驾驶等对检测速度要求高的应用场景。一阶段算法的典型代表是YOLO系列算法。

卷积神经网络(Convolutional Neural Network, CNN)作为计算机视觉(Computer Vision, CV)领域的主流深度学习模型,取得了很好的效果。在图像问题中,卷积计算具有平移等价性、局部性等天然优势,但也存在感受野固定且有限这一缺陷,通常需要堆叠卷积层来获取更大的感受野。近两年来,Transformer^[2]开始在计算机视觉领域大放异彩,被广泛应用于图像分类、目标检测、语义分割以及图像生成等任务当中,其核心自注意力(Self-attention)机制通过在输入向量之间相互执行注意力操作来提取特征,可以获得全局特征关系,通过全局范围的依赖关系来抑制无关信息、突出有用区域,以提高检测的准确性,有效弥补了卷积感受野固定的缺陷。

因此,针对辅助驾驶系统中车辆、行人的目标检测任务,本文在YOLOv3算法模型的基础上,引入自注意力机制,并对加入的模块进行高斯掩码优化,提出了YOLOv3-GMSA目标检测方法。此外,为了对方法进行评估验证,我们在MS COCO 2017数据集上对YOLOv3-GMSA方法进行了训练和评估,结果表明该方法在保证检测速度的同时有效提高了检测效果。

2 相关工作

2.1 目标检测

目标检测是CV领域的一个重要研究方向,其任务是在图像或视频中找出感兴趣的物体,同时检测出它们的位置和大小。不同于图像分类任务,目标检测不仅要解决分类问题,还需要解决定位问题。传统的目标检测主要关注图像处理、特

征提取算法,但是基于手工提取的特征无法适应复杂多变的现实场景,准确率不高。直到2012年,CNN的兴起将目标检测相关研究推向了新的台阶。

2.1.1 二阶段目标检测算法

二阶段算法将目标检测过程划分为两个步骤,首先根据算法产生一系列候选框,然后根据深度神经网络对候选框进行分类和回归。最具代表性二阶段目标检测算法是R-CNN^[3],通过选择性搜索算法从一组候选框中选择可能出现的对象框,然后将对象框中的图像处理后送入CNN提取特征,最后将提取的特征送入分类器进行预测。R-CNN在VOC-07数据集上取得了质的飞跃,平均精度由传统算法的33.7%提高到58.5%。针对R-CNN训练流程复杂、存储空间占用大等不足,Girshick等提出具有ROI池化层并且利用Softmax和线性回归同时返回分类结果和边界框的Fast R-CNN^[4],该网络在VOC-07数据集上将检测精度提高到70.0%,检测速度也得到了很大的提高。由于选择性搜索算法寻找感兴趣区域运算速度慢,Ren等提出了Faster R-CNN^[5]模型,创造性地采用卷积网络产生候选框,极大提升了候选检测框的生成速度。2017年,Lin等^[6]在Faster R-CNN的基础上进一步提出特征金字塔网络(Feature Pyramid Networks, FPN)结构,有效解决了深层特征图缺乏空间信息影响定位精度的问题。总的来说,二阶段目标检测器可以达到与人类相当甚至更好的精度性能,但是大量的计算阻碍了其在实时系统中的应用。

2.1.2 一阶段目标检测算法

一阶段算法不需要产生候选框,直接将目标框定位问题转化为回归问题处理,仅使用一个CNN网络就可以预测不同目标的类别概率和边界框,因此具有更快的检测速度。YOLO^[7]作为第一个一阶段目标检测算法,将图像分割为 $S \times S$ 个网格,然后为每个网格预测边界框并同时给出类别概率,在VOC-07数据集上实现了155 FPS的实时性能,但同时也存在检测精度相对较低的缺点。针对YOLO小目标检测精度偏低的问题,Liu等^[8]提出了采用先验框和多尺度特征图结合的SSD多框预测算法,不同的分支可以检测不同尺度的目标。随后,YOLOv2^[9]通过引入新的特

征提取网络Darknet-19、批归一化、卷积加先验框预测和多尺度训练等多种改进策略,有效解决YOLO检测精度不足的问题。YOLOv3^[10-12]则借鉴各类研究成果进一步优化网络,采用了先进的Darknet-53残差网络和多分支检测结构,可以有效检测不同尺寸的对象。总的来说,一阶段目标检测器具有很快的运行速度,能够满足实时系统的要求,但预测精度相对二阶段算法较低。

2.2 Transformer与自注意力机制

2017年,Google团队在自然语言处理(Natural language processing, NLP)领域提出了经典算法Transformer,该模型抛弃了传统的CNN、RNN,网络结构由自注意力机制组成,不仅可以并行化训练,而且能够包含全局信息,基于Transformer的BERT^[13]词向量生成模型在NLP的11项测试中取得了最佳成绩。Transformer在NLP中取得的耀眼成绩引起了AI研究人员的浓厚兴趣,许多相关应用也出现在CV领域中,并取得了很好的效果。

受Transformer中的自注意力机制启发,Wang等^[14]提出了Non-local神经网络,利用全局图像之间的依赖关系为特征进行加权。尽管自注意力机制能够捕捉图像的全局信息,但是需要占用较多的显卡资源,因此Huang等^[15]提出了交叉注意力模块,仅在纵横交错的路径上执行注意力操作,有效限制了对存储和计算资源的占用。Yi等通过对空间特征关系进行建模提出了ASSD^[16]检测器,对全局特征信息进行可靠引导,在精度上实现了很大的提升,但是这些方法都忽略了局部信息的重要影响,在检测精度上仍无法取得满意效果。ViT^[17]是第一个在大规模视觉数据集上完全用Transformer取代标准卷积操作的深度学习模型,其将图像划分成固定大小的块,通过线性变换得到块向量,随后送入Transformer进行特征提取并分类,模型在大规模数据集上做预训练随后在目标数据集进行微调,在ImageNet数据集上可以得到CNN方法相当的成绩,而且随着数据集的扩大,ViT将发挥更大的优势。Facebook AI团队首次将Transformer应用到目标检测任务,提出了DETR^[18]模型,取代了现有模型中需要手工设计的工作,如非极大值抑制和锚框生成等,在准确率和运行时间上能够取

得与 Faster R-CNN 相当的成绩。

总的来说,自注意力机制可以有效捕获图像的全局特征关系,弥补 CNN 感受野受限的不足,设计过程中也需要注意显卡资源占用情况,同时还需要考虑全局特征和局部特征进行融合时的权重。

3 引入 GSMA 模块的模型设计

YOLOv3 是对 YOLO、YOLOv2 进一步优化后提出的一阶段目标检测深度学习模型,只需要一次端到端计算就可以得到目标的类别概率和边界框,具有运算速度快的特点,适用于实时场景下的目标检测任务。考虑到随着特征层的增加,图像的上下文信息和全局关系会逐渐减少,影响了模型的检测精度,因此,本文在 YOLOv3 的网络结构中引入自注意力机制,并在自注意力模块中加入高斯掩码操作,使模型在捕获全局特征关系的同时仍然重视局部特征的影响,以取得更好的检测效果。

3.1 网络模型

YOLOv3 的模型架构主要分为两个部分: Darknet-53 和多分支预测网络。相比于 YOLOv2 的 Darknet-19,模型采用了具有残差单元的 Darknet-53 特征提取网络,构建的网络层次更深。另外,为了使多种尺度下的目标都具有较好的检测效果,模型采用具有 FPN 架构的多分支检测结构,采用 3 个尺度的特征图分别对大、中、小 3 类目标进行检测。

3.1.1 Darknet-53 与多分支预测网络

为了达到更好的分类效果,YOLOv3 的作者借鉴残差网络的思想,设计了 Darknet-53 特征提取网络。Darknet-53 网络是一个全卷积网络,包含 23 层残差单元(Residual Unit),具有非常多的残差跳层连接,其结构见图 1,左上是网络的结构,下方的虚线框展示了各模块的具体结构,模块中的数字表示其重复的次数。为了缓解池化操作带来的梯度负面效果,网络用卷积层替代池化层,通过调整卷积的步长为 2 实现降采样。

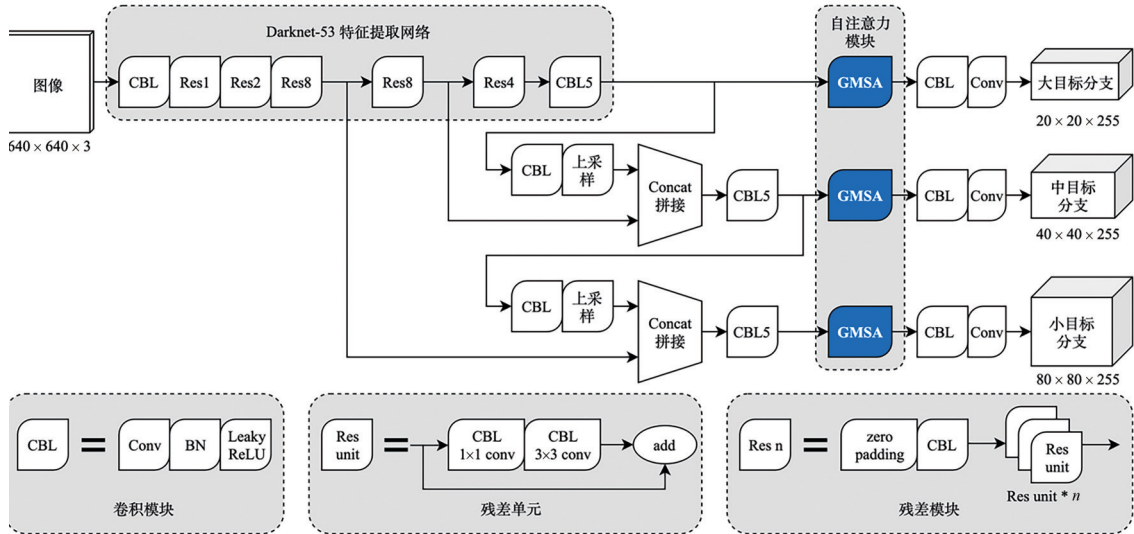


图 1 YOLOv3-GSMA 网络模型架构
Fig. 1 Architecture of YOLOv3-GSMA network

多分支预测网络构建 3 个预测分支,其结构如图 2 所示。当输入图片大小为 640×640 时,3 个尺度的特征图大小见表 1,缩小的倍数分别为 32,16,8。特征图尺寸越小,聚合的信息范围就越大,对大目标检测越有利;反之,特征图尺寸越大,保留的局部细节信息就越多,对小目标检测越有用。为了在大尺寸特征图中保留小尺寸特

征图包含的更高层语义信息,FPN 结构采用了上采样和融合的做法,将较小尺寸特征图上采样至较大尺寸,然后与较大尺寸特征图进行拼接融合。

3.1.2 边界框预测

不同于二阶段算法需要生成候选框,YOLO 系列算法将图像分割为 $S \times S$ 个网格,划分的网格对应特征图的尺寸。在各检测分支中,YO-

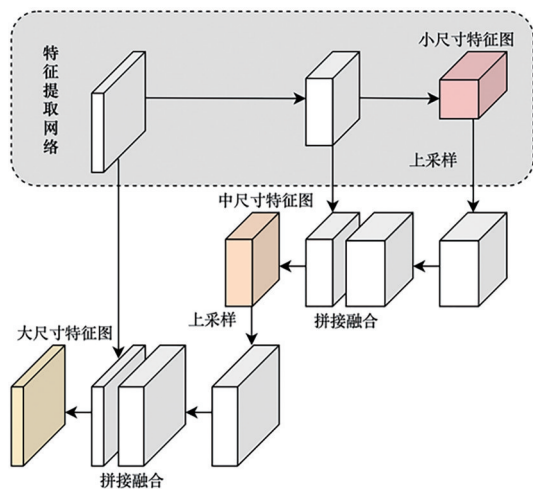


图2 特征金字塔结构的多尺度特征融合

Fig. 2 Multi-scale feature fusion of feature pyramid network

表1 图片大小为640×640时特征图尺寸和先验框尺寸

Tab. 1 Feature map size and prior frame size with picture size of 640×640

特征图尺寸	感受野	先验框尺寸
		(116, 90)
20×20	大	(156, 198)
		(373, 326)
40×40	中	(30, 61)
		(62, 45)
		(59, 119)
80×80	小	(10, 13)
		(16, 30)
		(33, 23)

LOv3对目标采用了K-means聚类,每个分支设置3个先验框,先验框尺寸如表1所示。模型采用全卷积的网络结构,最后一层卷积输出的卷积核个数是 $3 \times (N + 4 + 1)$,对应3个预测边界框,每个预测边界框包含 N 个类别概率、4个边界框位置信息和1个边界框的置信率。针对MS COCO数据集的80个类别,最后一层卷积的卷积核个数是255。

不同于YOLO采用Softmax计算类别概率并用最大概率的标签表示边界框内的目标,YOLOv3可以实现对检测目标的多标签分类,方法是对先验框中的图像采用二元逻辑回归(Logistic Regression),在每一类别上计算是该类别的概率,并设定阈值对多标签进行预测。

3.2 高斯掩码自注意力模块

为了增强模型对全局特征关系的捕捉能力,

本文引入Transformer的自注意力机制,在多分支检测网络的出口附近嵌入高斯掩码自注意力模块,如图1所示。该模块分别对大、中、小3种尺寸的特征图进行自注意力操作,让特征图中包含更多的大范围信息,提高算法的检测精度。

3.2.1 自注意力机制

注意力(Attention)机制^[19]最早由Bengio等在2014年提出,其思想就是将有限的注意力集中到重点信息上。在Transformer中,进出自注意力结构的输入和输出都是一个向量序列,每个向量都表示一个项,将输入序列对自身进行注意力操作,根据完整的输入序列来聚合全局信息以更新序列的每一个项,可以从全局范围对各项之间的相关性进行建模,例如在自然语言处理中可以评估哪几个词更可能组成一个句子。

图3(a)展示了自注意力基本块的运算结构。对于包含 n 个项的序列 $(x_1, x_2, \dots, x_n)$,用 $X \in \mathbb{R}^{n \times d}$ 表示,其中 d 是每个项的向量长度。自注意力的目标是通过全局上下文对每个项进行编码来捕获 n 个项的相互关系,其方式是定义3个可学习的权重矩阵 W^K, W^Q, W^V 将 X 转换为查询 $Q = XW^Q$ 、键 $K = XW^K$ 以及值 $V = XW^V$ 。自注意力模块的输出由公式(1)给出:

$$Z = \text{softmax} \left(\frac{QK^T}{\sqrt{d_q}} \right) V. \quad (1)$$

对于每一个项,自注意力模块将其查询向量

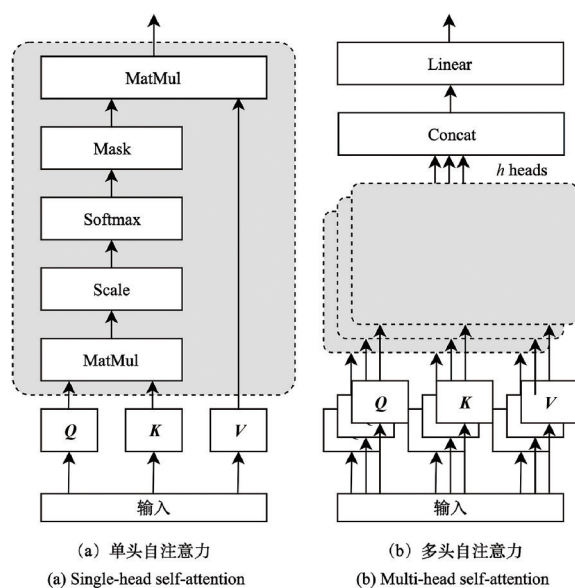


图3 自注意力计算结构示意图

Fig. 3 Structure of self-attention

与全部项的键向量做内积,然后用Softmax激活函数做归一化得到注意力得分,基于注意力得分可以用全部项对每一项进行加权求和得到新的输出。

一个标准的注意力模块会对所有项进行注意力操作,但是对于图像数据而言,局部特征关系的重要性大于全局特征关系,因此,本文对获取的注意力得分进行高斯掩码计算,增加局部信息的比重,削减全局信息的比重。二维高斯分布计算公式如式(2)所示:

$$f(x, y) = (2\pi\sigma_1\sigma_2)^{-1} \exp\left[-\frac{1}{2(1-\rho^2)} \times \left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right], \quad (2)$$

其中参数 ρ, σ_1, σ_2 是可学习参数,根据全部项相对某一项的距离计算得到高斯分布的掩码向量,然后与该项的注意力得分向量计算哈达玛积,归一化后得到新的注意力得分。

为了对序列内各项之间的复杂关系进行建模,Transformer采用了具有多个自注意力基本块的多头自注意力(Multi-head Self-attention)机制,其结构如图3(b)中所示。每一个基本块拥有自己的权重矩阵 $\{W^{Q_i}, W^{K_i}, W^{V_i}\}$,其中 $i=0, 1, \dots, h-1, h$ 是包含基本块的个数。将输出 $[Z_0, Z_1, \dots, Z_{h-1}] \in \mathbb{R}^{n \times h \times d_v}$ 拼接后与权重矩阵 $W \in \mathbb{R}^{h \times d_v \times d}$ 做内积即可得到最终的输出。

3.2.2 位置嵌入

与卷积神经网络不同,自注意力模块需要将

输入的二维特征图转换为一维序列处理,具有排序无关的特性,因此丢失了每一个特征块的位置信息。在缺失位置信息的情况下对特征块之间的语义进行学习将会增大学习成本,所以需要提提供位置嵌入(Position Embedding),即特征块的位置编码,编码公式如式(3)所示:

$$\begin{cases} \text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d}}}\right) \\ \text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d}}}\right) \end{cases}, \quad (3)$$

其中 pos 表示转换后的一维位置序号, $2i$ 和 $2i+1$ 是位置编码的维度, $i \in [0, 1, \dots, \frac{d}{2}]$, 因此每一个位置的编码长度是 d , 位置嵌入与输入序列的大小相同。编码后的位置嵌入向量与输入序列的每一项表示向量相加即可嵌入位置信息。

3.2.3 GMSA 模块结构

加入YOLOv3网络模型的高斯掩码自注意力模块结构如图4所示。该模块输入与输出均为 $\text{batchsize} \times \text{channel} \times \text{width} \times \text{height}$ 的四维张量,在进行注意力操作之前,首先将二维特征图转换为一维序列,然后计算其位置嵌入向量并相加。对嵌入位置信息的序列执行带有高斯掩码的多头自注意力操作,获得带有全局信息的序列,将该序列与嵌入位置信息的序列相加并归一化后得到输出序列,最后进行维度调整即可得到输出特征图。

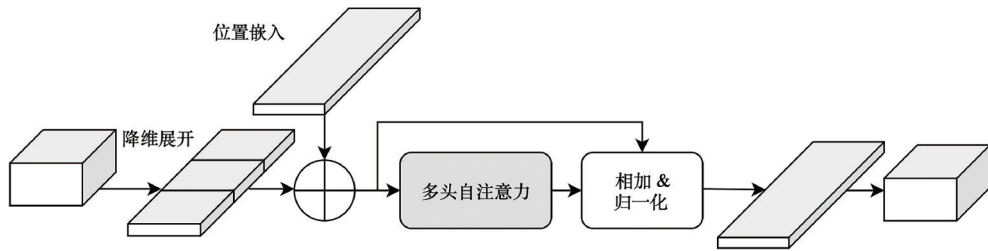


图4 GMSA 模块结构

Fig. 4 Structure of GMSA module

4 实验结果及分析

为了评估改进的YOLOv3-GMSA目标检测模型相关性能,本文在公开数据集上进行了一系

列实验,并对YOLOv3和改进后的YOLOv3-GMSA进行了对比,相关内容介绍如下。

4.1 训练环境及相关参数

训练算法模型使用MS COCO 2017数据集。该数据集由微软公司制作收集,用于目标检测、

分割、图像描述等任务,包含目标检测、关键点检测、实例分割、全景分割和图像描述等5种标注类型,本文采用目标检测标注对模型进行训练。该公开数据集包含80个类别,官方共划分成3个部分,其中训练数据集包含11.8万多张图片,存储空间占用18 GB;训练验证数据集包含5千张图片,存储空间占用1 GB;模型测试数据集包含4万多张图片,存储空间占用6 GB。模型训练过程中,图片尺寸为 640×640 。训练环境的配置如表2所示,算法模型采用PyTorch框架搭建,底层通过CUDA接口使用显卡资源进行加速。

训练相关参数如表3所示,其中批处理大小、

表2 训练环境配置

Tab. 2 Training environment

配置名称	型号、参数
CPU	Intel(R) Core(TM) i9-9900K,8核
固态硬盘	金士顿,512 G
内存	金士顿,16 Gx2,8 Gx2
显卡	NVIDIA TITAN Xp,显存12 G, CUDA 10.2
操作系统	Ubuntu 18.04
程序语言	Python 3.8.11
机器学习框架	PyTorch 1.9.0

迭代次数等参数在官方参考数值之上结合硬件条件及训练情况进行了适当调整,训练的优化算法采用Adam。

表3 训练参数

Tab. 3 Training parameters

参数名称	参数值
批处理大小	10
迭代次数	100
学习率	0.01
动量	0.937
置信度阈值	0.5
NMS 阈值	0.5
类别	80
自注意力头数量	8

4.2 实验结果对比分析

本文在MS COCO 2017数据集上,基于PyTorch框架在Ubuntu 18.04系统和NVIDIA TITAN Xp显卡之上对改进后的YOLOv3-GMSA算法进行了训练。经过100轮训练之后,使用模型测试数据集对算法性能进行评估,模型的性能对比如表4所示。

表4 算法模型性能对比

Tab. 4 Performance evaluation

算法模型	mAP@0.5/%	mAP@0.5:0.95/%	Precision/%	Recall/%	FPS
Faster R-CNN	58.98	38.54	68.67	56.33	21.12
SSD	47.33	29.49	56.19	44.93	35.56
ASSD	52.73	31.81	60.24	49.48	36.71
YOLOv3	54.32	34.17	61.78	51.96	43.26
YOLOv3-GMSA	56.88	36.19	65.31	53.18	39.38

从表中可以看出,改进后模型YOLOv3-GMSA训练的结果mAP@0.5达到56.88%,精度达到65.31%,召回率达到53.18%。与YOLOv3相比,mAP@0.5提高了2.56%,精度提高了3.53%,召回率提高了1.22%,目标检测的效果显著提高。模型的检测效果相比于二阶段检测算法Faster R-CNN略有不足,但在检测速度方面具有较大的优势。人眼自然帧率是24 FPS,车载视觉传感器常见帧率一般不低于该数值,实际场景中显卡等硬件资源受限,处理速度将进一

步下降,二阶段算法难以满足实时性要求。相比于SSD、ASSD等目标检测算法,改进后模型的检测效果更好,检测速度更快。由于嵌入了高斯掩码自注意力模块,模型复杂度增加,参数增多,检测速度有所下降。改进后的模型检测处理速度达到39.38 FPS,相比于YOLOv3下降3.88,但仍满足实时检测的要求。通过以上数据分析可以得出,在YOLOv3网络深层插入高斯掩码自注意力模块以加强网络模型对全局特征关系的捕捉能力,使特征图包含更多的全局信息,更有

意义,可以有效提高目标检测的效果。

为了更好地观察 YOLOv3-GMSA 改进算法的训练情况,图 5 记录了其训练过程中 mAP@0.5、mAP@0.5:0.95、精度以及召回率的趋势曲线。可以看到 mAP@0.5、mAP@0.5:0.95 始终呈上升趋势,在 40 轮左右逐渐收敛,精度和召回率则在上下波动中整体呈上升趋势,且精度和召

回率的波动趋势相反,这是精度和召回率本身存在一定矛盾所致。

为了更直观地感受改进后算法的检测效果,图 6 通过对 3 张实际行车环境中拍摄得到的照片进行检测,对比了 YOLOv3 和改进后的 YOLOv3-GMSA 算法检测效果,检测时置信度阈值设置为 0.25。图中第一列是原图,第二列是 YO-

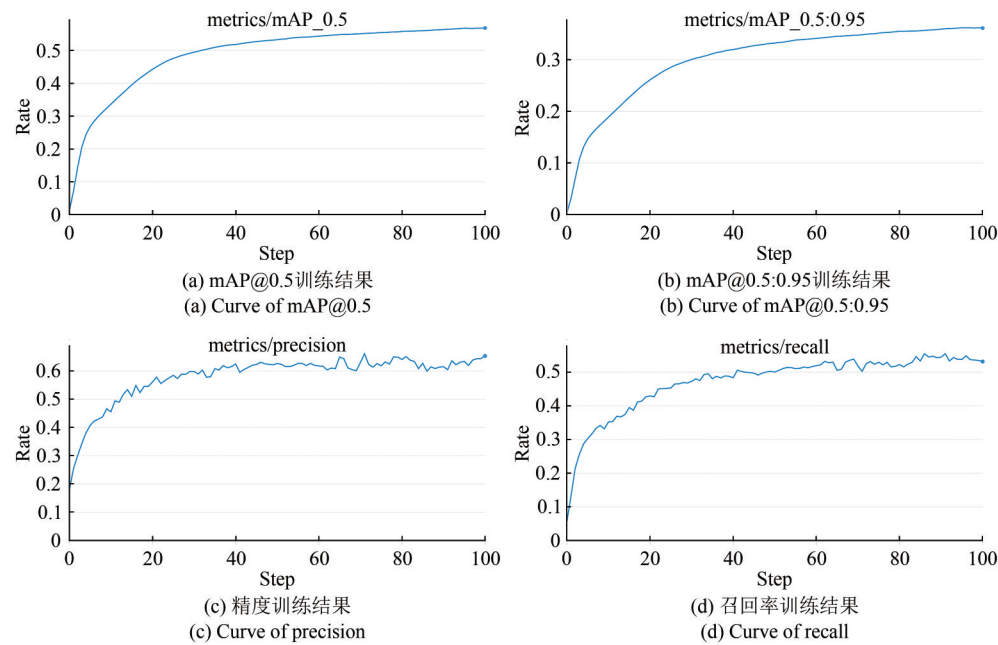


图5 YOLOv3-GMSA 模型训练关键参数曲线

Fig. 5 Training curve of pivotal parameters in YOLOv3-GMSA



图6 检测结果对比

Fig. 6 Comparison of detection results

LOv3的检测结果,第三列是YOLOv3-GMSA的检测结果。从第一行图中可以看出,YOLOv3和改进算法均具有良好的检测效果,能够对马路上的行人进行检测定位,相比于YOLOv3,改进后算法具有更高的置信度;第二行图中,YOLOv3出现明显的识别错误,将巴士的后视镜识别为人,置信度达到0.36,而改进后的算法能够有效捕捉全局信息,融入全局特征关系,没有出现类似的误识别情况;第三行图中,YOLOv3没有识别出汽车驾驶员以及车后骑自行车的行人,改进后算法则有效标定了这两块地方,置信度达到0.26。

总而言之,YOLOv3-GMSA改进算法通过引入自注意力机制,增强了模型对全局特征关系的捕捉能力,在牺牲一定检测速度的前提下,可以有效提高目标检测的效果。

5 结 论

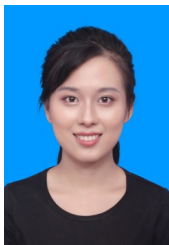
针对辅助驾驶的目标检测场景,受计算机视觉领域的Transformer应用启发,本文在一阶段目标检测算法YOLOv3的基础之上,嵌入了高斯掩码自注意力模块,加强网络模型对全局特征关系的捕捉能力,以提高检测精度。在公开数据集MS COCO 2017上训练后的实验结果表明,改进后的模型将mAP@0.5提高了2.56%,虽然在检测速度上略有损失,但仍能满足实时性需求。该方法为辅助驾驶系统中的目标检测提供了一种新的优化方案,而且通过调整自注意力模块的设计,可以在检测精度和检测速度之间进行有效权衡,以适应不同场景下的检测要求。

参 考 文 献:

- [1] 中国汽车工业协会. 2020年中国汽车工业经济运行报告[EB/OL]. [2021]. <http://lwzb.stats.gov.cn/pub/lwzb/tzgg/202107/W020210723348607396983.pdf>.
China Association of Automobile Manufacturers. The economic operation report of China's automobile industry in 2020 [EB/OL]. [2021]. <http://lwzb.stats.gov.cn/pub/lwzb/tzgg/202107/W020210723348607396983.pdf>. (in Chinese)
- [2] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need [C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates Inc., 2017: 6000-6010.
- [3] GIRSHICK R, DONAHUE J, DARRELL T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014: 580-587.
- [4] GIRSHICK R. Fast R-CNN [C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015: 1440-1448.
- [5] REN S Q, HE K M, 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, 39(6): 1137-1149.
- [6] LIN T Y, DOLLÁR P, GIRSHICK R, *et al.* Feature pyramid networks for object detection [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 936-944.
- [7] REDMON J, DIVVALA S, GIRSHICK R, *et al.* You only look once: unified, real-time object detection [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 779-788.
- [8] LIU W, ANGUELOV D, ERHAN D, *et al.* SSD: single shot MultiBox detector [C]//*Proceedings of the European Conference on Computer Vision*. Amsterdam: Springer, 2016: 21-37.
- [9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 6517-6525.
- [10] REDMON J, FARHADI A. YOLOv3: an incremental improvement [J]. *arXiv*:1804.02767, 2018.
- [11] 唐悦,吴戈,朴燕. 改进的GDT-YOLOV3目标检测算法[J]. *液晶与显示*, 2020, 35(8): 852-860.
TANG Y, WU G, PIAO Y. Improved algorithm of GDT-YOLOV3 image target detection [J]. *Chinese Journal of Liquid Crystals and Displays*, 2020, 35(8): 852-860. (in Chinese)

- [12] 李天宇, 李栋, 陈明举, 等. 一种高精度的卷积神经网络安全帽检测方法[J]. 液晶与显示, 2021, 36(7): 1018-1026.
- LI T Y, LI D, CHEN M J, *et al.* High precision detection method of safety helmet based on convolution neural network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2021, 36(7): 1018-1026. (in Chinese)
- [13] DEVLIN J, CHANG M W, LEE K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding [C]//*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019.
- [14] WANG X L, GIRSHICK R, GUPTA A, *et al.* Non-local neural networks [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 7794-7803.
- [15] HUANG Z L, WANG X G, HUANG L C, *et al.* CCNet: criss-cross attention for semantic segmentation [C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul, Korea: IEEE, 2019: 603-612.
- [16] YI J R, WU P X, METAXAS D N. ASSD: attentive single shot multibox detector [J]. *Computer Vision and Image Understanding*, 2019, 189: 102827.
- [17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16×16 words: transformers for image recognition at scale [C]//*Proceedings of the 9th International Conference on Learning Representations*. Virtual Event, Austria: OpenReview.net, 2021.
- [18] CARION N, MASSA F, SYNNAEVE G, *et al.* End-to-end object detection with transformers [C]//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer, 2020: 213-229.
- [19] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C]//*Proceedings of the 3rd International Conference on Learning Representations*. San Diego, 2015.

作者简介:



孔雅洁(1995—),女,河南洛阳人,硕士研究生,2017年于东北大学获得学士学位,主要从事计算机视觉方面的研究。E-mail: 704778463@qq.com



张叶(1981—),女,吉林长春人,博士,研究员,2008年于中国科学院长春光学精密机械与物理研究所获得博士学位,主要从事计算机视觉、模式识别算法方面的研究。E-mail: yolanda@spirits.ai