

基于视觉-语言预训练模型的开集交通目标检测算法

黄琦强¹, 安国成^{2*}, 熊刚¹

(1. 上海交通大学电子信息与电气工程学院, 上海 200240;

2. 上海华讯网络系统有限公司行业数智事业部, 四川 成都 610074)

摘要: 交通目标检测是智慧交通系统的重要组成部分,但现有的交通目标检测算法只能实现对于预设目标的检测,无法应对开集目标场景。为此,提出一种基于视觉-语言预训练(VLP)模型的开集交通目标检测算法。首先,基于 Faster R-CNN 修改预测网络使其能够适应开集目标的定位问题,并将损失函数改进为交并比(IoU)损失,有效提升定位精度;其次,构建一种新的基于 VLP 的标签匹配网络(VLP-LMN),对预测框进行标签匹配,VLP 模型作为一个强大的知识库,可有效匹配区域图像和标签文本,同时,VLP-LMN 的提示工程和微调网络模块可以更好地发掘出 VLP 模型的性能,有效提高 VLP 模型标签匹配的准确性。实验结果表明,该算法在 PASCAL VOC07+12 数据集上实现了 60.3% 的新类目标检测平均准确率,这证明了其具有良好的开集目标检测性能;同时在交通数据集上的新类目标检测平均准确率达到 58.9%,作为零样本检测,仅比基类目标低 14.5%,证明了该算法在交通目标检测上具有良好的泛化能力。

关键词: 视觉-语言预训练模型;Faster R-CNN;开集目标检测;交通目标检测

源代码链接: <https://github.com/adamhuangqq/VLP-based-OVD>

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0069168

Open-Set Traffic Object Detection Algorithm Based on Vision-Language Pre-training Model

HUANG Qiqiang¹, AN Guocheng^{2*}, XIONG Gang¹

(1. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;

2. Artificial Intelligence Research Institute, Shanghai Huaxun Network System Co., Ltd., Chengdu 610074, Sichuan, China)

【Abstract】 Traffic object detection is a crucial component of intelligent transportation systems. However, existing traffic object detection algorithms can only detect predefined objects and are incapable of handling open-set object scenarios. To address this, a novel open-set traffic object detection algorithm based on a Visual-Language Pre-trained (VLP) model is proposed. First, by leveraging Faster R-CNN as a foundation, the prediction network is modified to adapt to the localization challenges of open-set objects. The loss function is refined to the Intersection over Union (IoU) loss, effectively enhancing the localization accuracy. Second, a new VLP-based Label Matching Network (VLP-LMN) is constructed to perform label matching on the predicted bounding boxes. The VLP model serves as a potent knowledge repository that effectively matches regional images with labelled text. Simultaneously, prompt engineering and fine-tuning of network modules facilitate better exploration of the VLP model's performance, significantly improving the accuracy of label matching. The algorithm achieves an average detection accuracy of 60.3% for new classes on the PASCAL VOC07+12 dataset, demonstrating its commendable performance in open-set object detection. Additionally, the average detection accuracy for new classes on a traffic dataset reaches 58.9%, with only a 14.5% decrease compared with the base classes in zero-shot detection. This underscores the strong generalization capabilities of the algorithm in traffic object detection.

【Key words】 Visual-Language Pre-trained (VLP) model; Faster R-CNN; open-set object detection; traffic object detection

0 引言

随着我国经济的飞速发展,人们的出行方式发生了深刻的变化,汽车已成为最重要的代步工

具,车辆数目的逐年攀升给交通系统带来了巨大的压力,为了缓解这些压力,同时规范交通行为,打击违规行为,智慧交通系统被广泛应用^[1]。交通目标检测作为智慧交通系统的重要一环,是指

收稿日期: 2024-01-04 修回日期: 2024-02-20

基金项目: “十四五”国家重点研发计划(2023YFC3006700);国家自然科学基金(62071293)。

通信作者 E-mail: *anguocheng@eccom.com.cn

对摄像头等传感器捕捉到的图像进行目标对象的检测,对其定位并确定类别信息。随着深度学习的不断发展,目标检测在深度学习领域取得了巨大的成功,各种目标检测算法被引入到了交通目标检测中^[2-4]。然而,尽管这些基于深度学习的目标检测算法已经取得了巨大进步,例如 Faster R-CNN^[5]、YOLO^[6-9]等算法已取得较好的准确性,但训练此类模型需要较大的代价,比如对于每个感兴趣的目标类别需要标注至少数千个边界框,而这都要靠人工标注才能获得。尽管已经有很多机构公开了许多有价值的数据集,例如 KITTI^[10]、COCO^[11]等,但这些数据集都只描述了一些有限的对象类别,而在智慧交通领域中,需要识别的交通目标的类别可能远大于数据集中所出现的这几类。比如在智能驾驶中,不仅需要识别新类别目标,还要适应新场景;此外,许多交通场景中的异常目标也要求模型具有开集检测的能力。按照当前目标检测算法的思路很难解决这些问题,不仅因为标注代价超出了承受范围,而且这么多类别的分类任务也异常复杂,难以实现。除此之外,即使可以收集到相应数据,不同类别的样本数目也必然是差别巨大的,比如异常目标往往不足正常目标数目的千分之一,对于一个分类问题来说,这样的样本差距会带来严重的样本不均衡问题,给准确识别目标带来了极大的困难。以上诸多问题都说明传统的目标检测网络难以满足智慧交通进一步的需求,因此亟须一种开集目标检测方案。

事实上,针对如何将目标检测拓展到开放世界的问题,计算机视觉领域的学者们做了诸多研究。起初,研究人员采取的研究路线是零样本学习(ZSL),这些方法^[12-14]扩展了检测器,从有标注的类别(基类或可见类)推广到其他没有标注的类别(新类或不可见类)。在训练过程中只使用可见类的标注数据,同时,不使用不可见类的注释。其中大多数方法是采用训练得到词嵌入的投影来构造分类器,从而进行不可见类的分类。然而,上述方法具有很多限制,通常,这些方法缺乏不可见类的实例,并且在训练期间将这些不可见类的目标视为背景目标,因此在推理过程中,该模型仅基于其预定义的词嵌入来识别新的类,这限制了对新类的视觉信息和关系的探索,很难达到好的识别效果。

2021 年,ALIREZA 等^[15]提出的 OVR-CNN 为这一问题给出了另外一种解决思路,OVR-CNN 提出一种全新的开集目标检测范式:开放词汇目标检测(OVD)。在该文中,作者提出两阶段 OVD 的

范式:

1) 基于图像-文本对数据,训练视觉-语言预训练(VLP)模型,得到对齐的图像嵌入(IE)和文本嵌入(TE)空间。

2) 将 VLP 模型作为教师模型,对两阶段检测器的预测网络进行知识蒸馏,使其可以预测出图像嵌入 IE。

3) 使用 VLP 对基类和新类的标签文本进行编码得到 TE,计算其与 IE 的相似度,最后选择相似度得分最高的文本作为建议框的标签。

近期,随着跨模态大模型 CLIP^[16]的成功,研究者们开始将这些性能更优的 VLP 模型引入到 OVD 任务的训练。2022 年,GU 等^[17]利用 CLIP 作为 teacher 模型,对目标检测模型的类别预测头进行蒸馏,得到了优于 OVR-CNN 的结果。同年,ZHONG 等^[18]提出了 RegionCLIP 模型,该模型在 CLIP 的基础上,针对目标检测中的单个目标图片和完整图片的差异问题做了优化,利用大规模的目标文本进行预训练,取得了更好的效果。目前,VLP 模型有了更加长足的进步,ALIGN^[19]、ALBEF^[20]、BLIP^[21]等模型拥有比 CLIP 更强的性能,这为 OVD 任务提供了强大的助力。

尽管已经有了许多针对 OVD 任务的算法,但这些算法依然存在着许多问题。首先是具体评价指标有许多模糊之处,比如现在大多 OVD 模型使用的数据集划分方案会将常见类设为基类,将不常见类设为新类,基类进行训练,新类进行测试,但是这样的划分存在一个明显的矛盾,即在训练过程中,新类目标是被当作背景的,但是在测试中,新类目标是最重要的前景目标。其次现有的 OVD 模型要么需要进行目标-标签预训练,要么使用 VLP 模型对传统网络进行知识蒸馏。前者需要充足的算力和数据支持,成本昂贵,后者由于知识蒸馏这个方法本身的限制,使得 VLP 模型的能力并不能被充分发掘。这些问题都使得 OVD 模型很难直接引入到开集交通目标检测中。

为解决以上问题,本文提出一种基于 VLP 的开集交通目标检测算法。首先,基于 Faster R-CNN 提出了开集目标定位网络,并通过改进损失函数,提高其定位精度;其次,在通过定位网络得到预测框后,基于 VLP 模型提出了标签匹配网络,该网络负责对预测框进行标签匹配。为了提高匹配精度,本文提出提示工程模块和区域图像-标签微调网络,以更有效地利用 VLP 模型的能力,提高其开集识别性能。

1 基于 VLP 模型的开集交通目标检测算法

算法整体框架如图 1 所示,主要由两部分组成,第一部分是基于 Faster R-CNN 的适应开集检测任务

的目标检测网络(OFR-CNN),该网络完成目标定位任务,第二部分是基于 VLP 模型的标签匹配网络(VLP-LMN),完成对 OFR-CNN 所定位目标的标签匹配任务。

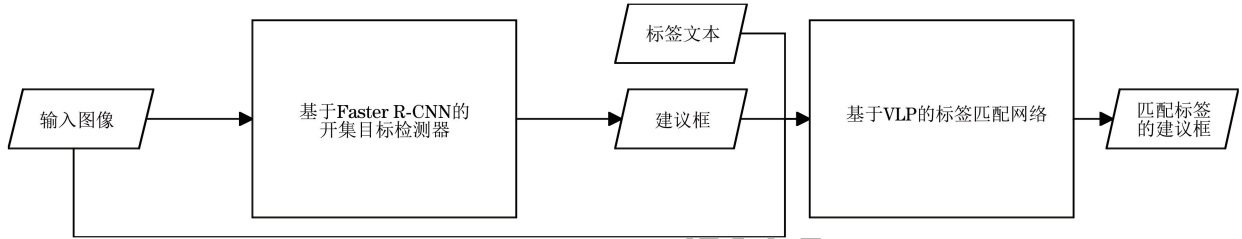


图 1 基于 VLP 模型的开集交通目标检测算法框架

Fig.1 Framework of open set traffic object detection algorithm based on VLP model

1.1 OFR-CNN

Faster R-CNN 是端到端目标检测网络的开山之作,由主干网络(Backbone)、区域建议网络(RPN)和预测网络组成。如图 2 所示,主干网络负责提取图像特征,Faster R-CNN 采用 Resnet50^[22] 的前四层作为主干网络,抽取主干网络的最后一层

作为特征图,以 $640 \times 640 \times 3$ 的图像为例,其每层网络的具体输出参数如表 1 所示。在得到特征图后,先将特征图送入 RPN 网络中,得到感兴趣区域(RoI),再将 RoI 通过 RoI 池化层映射回特征图,这保证了每个 RoI 尺寸一致,最后将其送入预测网络,得到最后的定位和分类结果。

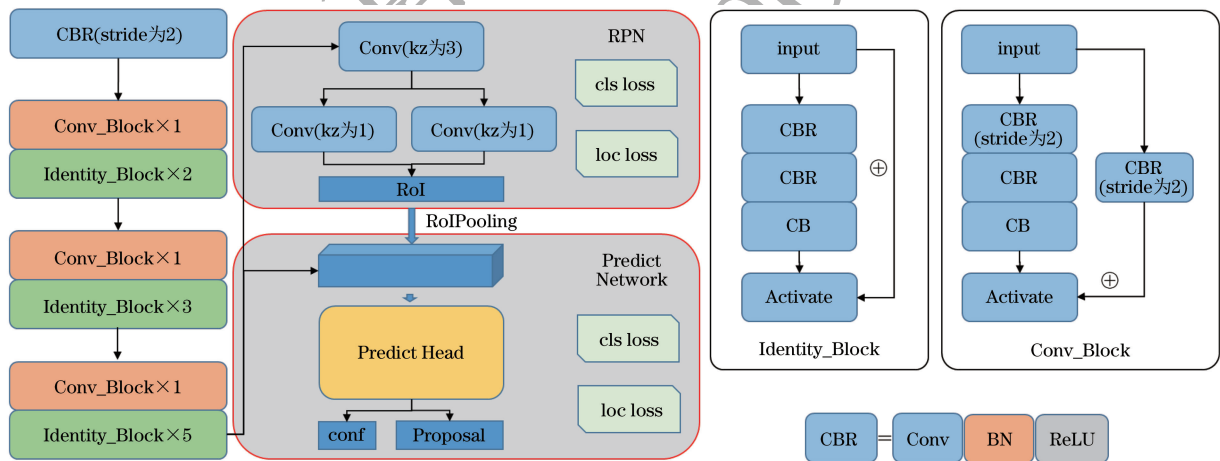


图 2 Faster R-CNN 的网络结构

Fig.2 Network structure of Faster R-CNN

表 1 Faster R-CNN 的主干网络参数

Table 1 Backbone network parameters of Faster R-CNN

层级	网络结构	输出
输入	Conv+Maxpooling	$320 \times 320 \times 64$
Layer1	Conv_Block+Identity_Block $\times 2$	$160 \times 160 \times 256$
Layer2	Conv_Block+Identity_Block $\times 3$	$80 \times 80 \times 512$
Layer3	Conv_Block+Identity_Block $\times 5$	$40 \times 40 \times 1\ 024$

虽然现在已经有了很多诸如 YOLO、RetinaNet^[23] 和 DETR^[24] 等性能更优秀的目标检测网络,但是这些目标检测网络的目标定位和分类 2 个模块耦合度很高,很难将目标定位网络单独解耦,即使解耦之后也很难获得好的定位效果。因此,本文选择 Faster R-CNN 作为基础网络。

如图 3(a) 所示,Faster R-CNN 的种类置信度预测头是固定的,所以在模型定义完成后,其只能检测确定数目的类别,为了适应 OVD 任务的需要,本文对预测头进行如图 3(b) 所示的修改,使其只对 RoI 属于目标还是背景进行判断,即预测部分由一个目标置信度预测头和一个回归参数预测头组成,其中目标置信度预测头只用来预测 RoI 属于目标还是背景,不预测具体类别,这样就剥离了目标的分类和定位 2 个任务,即使只用基类目标进行训练,该模型依然可以定位新类目标,即检测器具备了开集检测能力。

Faster R-CNN 网络的损失函数由 4 个部分组成,即 RPN 网络的定位损失、分类损失、预测头的定位损失和分类损失。2 种定位损失都是使用平滑

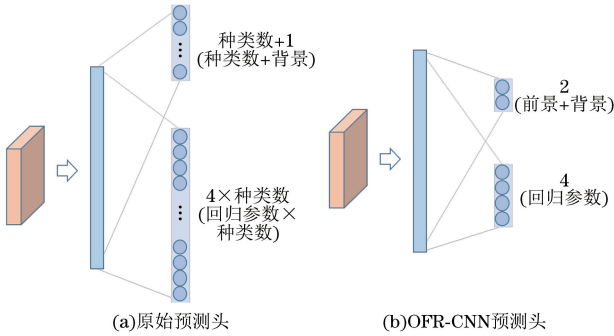


图 3 原始预测头与 OFR-CNN 预测头

Fig.3 Original prediction head and OFR-CNN prediction head

1 范数损失,如式(1)所示:

$$\text{Smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{else} \end{cases} \quad (1)$$

式中: x 代表预测值与真实值之差。

2 种分类损失采用二值交叉熵(BCE)损失,如式(2)所示:

$$B_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log_a(p(y_i)) + (1 - y_i) \cdot \log_a(1 - p(y_i))$$

$$p(y_i) = \text{Softmax}(\hat{y}) \quad (2)$$

式中: y_i 代表预测框对应的真值,若为目标,则 $y_i = 1$,否则为 0。

平滑 1 范数损失是以预测调整量和实际调整量的绝对值为自变量的函数,对于一个预测框,其定位损失由 4 个独立坐标($x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$)的损失之和组成。但在目标检测任务中,实际的评价指标是交并比(IoU),即预测框和真实框的重合度,也就是说,这 4 个坐标的损失之间是存在约束的,平滑 1 范数损失的最优解并不是 IoU 评价指标的最优解。如

图 4 所示(彩色效果见《计算机工程》官网 HTML 版,下同),尽管红色预测框和蓝色预测框与真值的平滑 1 范数损失相等,但是两者的 IoU 却并不相同。

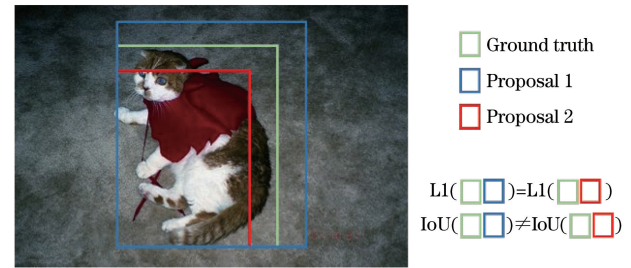


图 4 两种损失对比

Fig.4 Comparison of two types losses

因此,选择以评价指标 IoU 作为自变量的 IoU 损失^[25],如式(3)所示:

$$I_{\text{loss}}^{\text{IoU}} = -\sum_N \ln\left(\frac{\text{Intersection}(\text{gt}, \text{pre})}{\text{Union}(\text{gt}, \text{pre})}\right) \quad (3)$$

式中:gt 代表真实框;pre 代表预测框; N 为预测框的数目。这样 4 个独立坐标的损失就变成了评价指标的损失,相当于给原来的 4 个独立坐标加上了统一的约束,此时损失函数的最优解就是检测效果的最优解,这样可以有效地提高模型的检测性能。

1.2 VLP-LMN

通过 OFR-CNN 可以得到目标的定位结果,即建议框,但此时的建议框不具备类别信息,开集目标检测任务要求为 OFR-CNN 得到的建议框匹配类别信息,因此,本文提出一种基于 VLP 的标签匹配网络(VLP-LMN),如图 5 所示,该网络由提示工程(PE)模块、VLP 特征提取网络和区域图像文本匹配(RTM)微调网络 3 个部分组成。

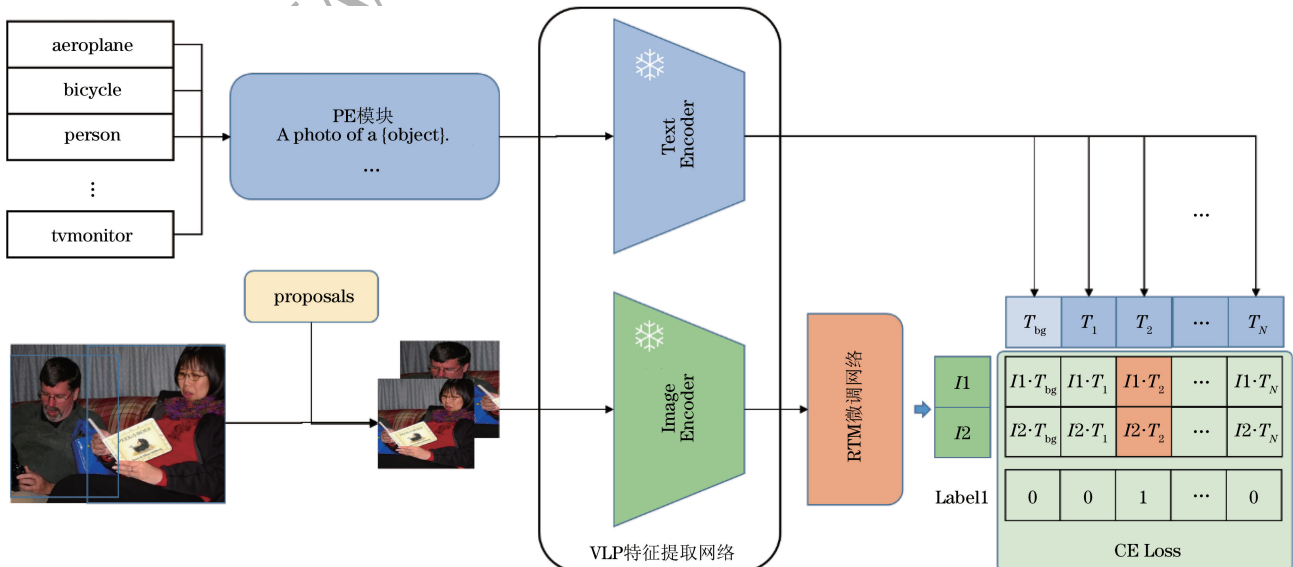


图 5 VLP-LMN 结构

Fig.5 Structure of VLP-LMN

1.2.1 PE 模块

提示工程是自然语言处理(NLP)领域中的一个概念,在提示工程中,任务的描述会被嵌入到输入中用以构成完整语义,从而使模型能够更好地理解自然语言。在 VLP 模型的训练过程中,由于标签文本大多时候并不具备完整语义,甚至可能只是一两个单词,为了得到更加丰富的语义信息,以达到更好的训练效果,其通常采用提示工程的方法扩充文本标签的语义。

本文提出的 PE 模块的主要功能是为文本编码器提供更丰富的语料,使得其与 VLP 模型的训练保持一致。基于文献[17]中的提示模板库,本文采用一种随机提示模板的方法,即对于每个标签 L_i ,在提示模板库中随机选出一个提示模板 T_r 组成文本标签 $T_r(L_i)$ 。

1.2.2 VLP 特征提取网络

VLP 模型在近年来有了飞速的发展,其基本原理是通过大规模的图像-文本对训练图像文本编码器,使得两者编码的文本嵌入(TE)和图像嵌入(IE)处于对齐的特征空间。CLIP 作为 VLP 领域的里程碑,其网络采用双流架构,使用 ViT^[26] 作为视觉编码器,BERT^[27] 作为文本编码器,通过约 4 亿个图像文本对进行对比学习,得到了一个有良好零样本(Zero-Shot)能力的 VLP 模型。但是,CLIP 模型训练使用的数据集是在网上收集的,由于成本原因,并未能进行数据清洗,因此许多图像文本对其并不对应,甚至意思完全相反。针对这个问题,LI 等^[21] 在 2021 年提出了 BLIP 模型,其主要贡献是给出了一个自蒸馏模型,可以实现对训练数据集的进一步清洗,使得训练得到的模型更加准确,在各项任务上的性能都超过了 CLIP。

因此,本文基于 BLIP 构建了多模态特征提取网络,使用 BLIP 的图像和文本编码器组成多模态特征提取网络,分别对 OFR-CNN 得到的建议框区域图像和标签文本进行特征提取,最后分别得到 IE 和 TE。

1.2.3 RTM 微调网络

虽然 VLP 模型具有强大的特征提取能力,但是其训练使用的数据是图像-文本对,对于按照 OFR-CNN 生成的建议框裁剪下来的区域图像,其标签匹配效果并不理想,主要面临以下 2 个问题:

1) VLP 模型对于区域图像的识别完全和区域图像的主体目标一致,如图 6(a)所示,虽然该区域的主体目标是 Person,但是标签 Horse 的得分更高。

2) VLP 模型的标签匹配分数和区域图像对应建议框的质量并不一致,如图 6(b)所示,定位质量差的框对应的区域图像反而获得了更高的标签匹配分数。

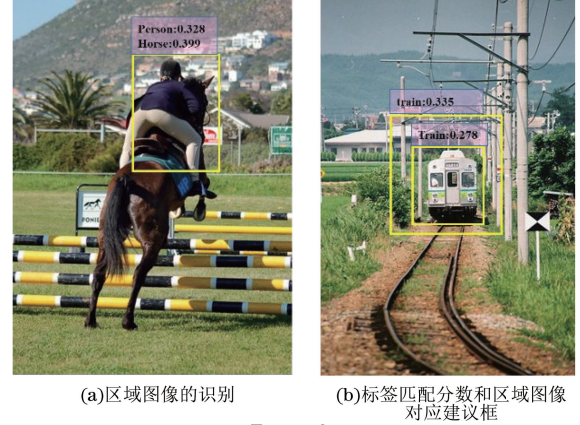


图 6 VLP 模型示意图

Fig.6 Schematic diagram of the VLP model

为解决以上问题,本文设计了 RTM 微调网络来实现 VLP 模型在区域图像与文本标签匹配任务上性能的提升。由于 VLP 模型编码器的特征提取能力已经很强,为防止区域图像标签文本对的加入对其产生破坏,冻结 VLP 特征提取网络,只将图像编码器得到的 IE 送入一个由多层感知机(MLP)构成的 RTM 微调网络,得到微调后的 IE,将其与 VLP 文本编码器编码得到的文本标签 TE 进行匹配,组成 RTM 微调模型,从而实现 VLP 模型的进一步微调。

RTM 微调网络的损失函数采用对比损失,先计算经过 IE 与所有标签 TE 的内积,此时,期望与 IE 正确标签 TE 的内积最大,而与其他标签的 TE 内积最小,因此,可以用交叉熵(CE)损失作为目标函数。具体流程如式(4)所示:

$$\begin{aligned} I_r &= \text{RTM}(\text{IEncoder}(\text{image}), r) \\ z(r) &= [\text{Sim}(I_r, T_{bg}), \text{Sim}(I_r, T_1), \dots, \text{Sim}(I_r, T_{|C_B|})] \\ L &= \frac{1}{N} \sum_{r \in P} \text{CE}(\text{Softmax}(z(r)), y_r) \\ \text{Sim}(a, b) &= a^T b / (\|a\| \cdot \|b\|) \end{aligned} \quad (4)$$

式中: I_r 表示通过 VLP 图像编码器和 RTM 微调网络得到的区域图像嵌入; T_r 表示经过 PE 模块和 VLP 文本编码器得到的标签文本嵌入; C_B 表示基类; N 表示一张图 P 中包含的建议框数目。由于背景的标签很难直接用一个文本来准确描述,因此选取一部分背景区域图像,通过 BLIP 模型的文本生成器得到其文本标签,再使用 VLP 的文本编码器对其进行编码,最后取平均嵌入得到背景标签。

2 实验结果与分析

2.1 数据集

由于现有的交通目标数据集类别太少,如 KITTI 只有 5 类目标,很难充分验证算法的开集检测性能。因此,本文采用 PASCAL VOC07+12 数据集(后文称 VOC 数据集),该数据集包含 21 503 张图片,标注类别为 20 类,总计目标实例为 52 091 例。为了验证算法的开集目标检测性能,将 20 类目标分为 10 类基类目标和 10 类新类目标,将新类目标出现过的图像作为测试数据,其余作为训练数据,具体的数据集划分方案如表 2 所示。

表 2 VOC 数据集划分

Table 2 Division of VOC dataset

类别	样本数	训练样本	测试样本	数据集划分
aeroplane	1 456	1 441	15	训练图片 12 232 张
bicycle	1 401	1 312	89	
bird	2 064	2 018	46	
bottle	2 233	1 655	578	
car	4 468	3 477	991	
chair	3 908	1 844	2 064	
dog	2 514	2 291	223	
person	17 784	11 072	6 712	
pottedplant	1 967	1 445	522	
tvmonitor	1 416	1 201	215	
bus	1 035	0	1 035	测试图片 9 271 张
motorbike	1 377	0	1 377	
train	1 207	0	1 207	
diningtable	1 031	0	1 031	
sofa	1 053	0	1 053	
cow	1 091	0	1 091	
horse	1 420	0	1 420	
sheep	1 312	0	1 312	
boat	1 403	0	1 403	
cat	1 951	0	1 951	
总计	52 091	27 756	24 355	21 503

此外,为了验证该算法在交通目标上的泛化性能,本文收集了某高速公路服务区视频监控数据(下文称交通数据集),该数据集包含不同天气、不同拍摄角度以及昼夜的图像数据,并标注了 4 类车型目标用于车辆目标检测,该数据集共有 5 219 张图像,将 car 和 truck 两类作为基类, bus 和 van 两类作为新类,采用和 VOC 数据集相同的划分规则,最后将 2 478 张图像用于训练,2 741 张图像用于测试。

2.2 训练参数

为了使模型泛化性更强,在 OFR-CNN 训练时

随机采用几何变换、色域调整、Mosaic 等方式进行数据增强。OFR-CNN 的训练均采用相同的训练参数,其中,训练世代为 100 次,使用 Adam 优化器,动量为 0.9,初始学习率为 1.0×10^{-4} ,学习率按照余弦方式更新。VLP-LMN 的训练采用使用 SGD 优化器,初始学习率为 5.0×10^{-3} ,学习率按照余弦方式更新,训练世代设为 300 次,在硬件方面,CPU 使用 13th Gen Intel® Core™ i7-13700K, GPU 使用 NVIDIA GeForce RTX 4090 24 GB。

2.3 评价指标

在目标检测中,常用的评价指标有平均精度 AP@0.5 和 AP@0.5:0.95,对于多类别的目标检测问题,还有 mAP@0.5 和 mAP@0.5:0.95。本文为更好地表征 OFR-CNN 对目标的定位能力,还采用了平均召回率 $AR_{\text{threshold}}@k$ 和召回率 $R(\text{conf} > \alpha)$ 两个指标, $AR_{\text{threshold}}@k$ 中下标 threshold 表示判定为正样本的 IoU 阈值, k 表示每张图片取建议框的个数; $R(\text{conf} > \alpha)$ 中 α 表示所选预测框的目标置信度阈值,选取为正样本的预测框与真实框的 IoU 阈值为 0.5。

2.4 实验结果

2.4.1 OFR-CNN 性能实验

对于 OFR-CNN,需要验证 2 项性能:1)改进后模型定位性能的提升;2)仅在基类数据上训练的模型对于新类目标的定位能力。

在表 3 中,定位损失第 1 项表示 RPN 网络的定位损失函数,第 2 项表示预测网络的定位损失函数, L1 代表平滑 L1 范数 L1 损失, IoU 代表 IoU 损失。可以发现,在 RPN 网络和预测网络中的定位损失均使用 IoU 损失时达到最佳效果,相较于原始的模型,其 AP@0.5 提升了 2.1 百分点,因此可以证明 IoU 损失比原有损失函数的定位性能更好。

表 3 OFR-CNN 在 VOC 数据集上的性能

Table 3 Performance of OFR-CNN on VOC dataset %

定位损失	AP@0.5	AP@0.5:0.95	AR@0.5:0.95@10	AR@0.5:0.95@100
L1+L1	65.9	35.2	45.4	48.7
IoU+L1	66.6	35.7	45.9	48.7
L1+IoU	67.0	36.5	45.9	47.8
IoU+IoU	68.0	37.0	47.1	49.6

由于 OFR-CNN 只在基类上进行训练,为了验证对于新类的定位能力,本文设计了如表 4 所示的实验。可以看到, OFR-CNN 即使仅使用基类数据进行训练,其在新类上的召回率在 3 个置信度阈值下也仅分别下降了 7.1、7.3 和 6.8 百分点,这说明了 OFR-CNN 对新类目标具有良好的定位效果。

表 4 OFR-CNN 在 VOC 数据集上的定位性能

Table 4 Localization performance of OFR-CNN on VOC dataset

类别	$R(\text{conf} > 0.5)$	$R(\text{conf} > 0.75)$	$R(\text{conf} > 0.9)$
基类	73.8	68.9	62.7
新类	66.7	61.6	55.9

2.4.2 VLP-LMN 性能实验

如第 1.2 节所述,基于 VLP 的标签匹配网络由 PE 模块、VLP 网络和微调网络 3 个部分组成,在本文中,VLP 特征提取网络选择 BLIP 模型的图像和

表 5 OFR-CNN 和 VLP-LMN 在 VOC 数据集上的性能

Table 5 Performance of OFR-CNN and VLP-LMN on VOC dataset

网络	VLP-LMN		mAP@0.5				
	PE	RTM	基类			新类	总体
			样本数小于 100		总体		
			样本数小于 100	样本数大于 100			
Faster R-CNN	—	—	29.9	46.5	42.5	—	—
OFR-CNN+VLP-LMN	×	×	30.8	42.3	38.8	52.1	45.5
	✓	×	43.3	42.1	42.4	50.0	46.2
	×	✓	39.9	59.0	53.3	60.5	56.9
	✓	✓	53.1	62.1	59.4	60.3	59.8

对比实例较少类别的检测效果,发现 RTM 微调网络可能并没有表现出较好的性能,这主要是因为对于数目较少的类别,其检测结果受个别实例的影响大,因此偶然性较大。此外,一些类别的检测效果不好是因为这些类别目标的尺度与目标的平均尺度相差较大,OFR-CNN 在其上的定位效果比较差,这对基类目标整体的检测性能产生了很大的影响。

在 PE 模块和 RTM 微调网络组合时,VLP-LMN 对于基类的检测能力相较于单独使用 RTM 微调网络有明显提高,推测这主要是因为 PE 模块为 VLP 的文本编码器提供了更加丰富的文字信息,从而可以训练得到更优的图像嵌入,进而提高基类的检测性能。

2.4.3 对比实验

为了进一步验证本文所提算法的性能,对现有的 2 种主流开集目标检测算法进行对比实验。

第 1 种是混合区域嵌入算法^[29](HRE),属于零样本学习的方法,其中 LE、CC 和 H 分别代表其所提出的 3 种不同的原型向量。

第 2 种算法是 ViLD,属于 OVD 算法。由于原算法采用了包括区域剪裁、区域扩张等一系列的数据增强方法,使得数据量非常大,导致训练异常困难,而本文所提算法在训练时并未进行特殊的数据增强,为保持一致并减少数据量,本文去掉了其数据增强的操作;同时,为了验证本文所改进检测器的性

能,还针对不同的检测器做了对比实验。文本编码器,对 PE 模块和微调网络设计了消融实验。如表 5 所示,总共做了 5 组实验,第 1 组是 Faster R-CNN 的实验结果,可以发现,Faster R-CNN 没有检测新类别的能力。第 2~5 组实验是 VLP-LMN 的消融实验结果,通过对比可以发现,PE 模块和微调网络对于 VLP-LMN 的匹配性能均有提升,而组合之后则能达到最佳的效果,相比于初始的 VLP 模型,基类上的 mAP@0.5 提升了 20.6 百分点,新类上的 mAP@0.5 提升了 8.2 百分点。

能,还针对不同的检测器做了对比实验。

实验结果如表 6 所示,首先从检测的准确性来看,零样本学习的方法 HRE 在基类和新类上的 mAP@0.5 有着明显的差距,虽然 HRE-H 算法在基类上的 mAP@0.5 可以达到 57.80%,但是在新类上锐减到 25.5%,这反映了零样本方法的最大问题,由于其并不依赖额外的知识库,因此其新类的检测率和基类检测准确率之间存在着巨大鸿沟。相对而言,ViLD 虽然在基类上的 mAP@0.5 和 HRE 相差不大,但是由于 VLP 模型为其提供了一个强大的知识库,其在新类上的检测能力明显优于 HRE 算法。同时也可以发现,由于是采用蒸馏的方式来进行目标检测,ViLD 对于新类的检测性能还是弱于基类的检测性能,并且在一阶段的目标检测框架上尤为明显,如 SSD 检测器差了 10.5 百分点,而两阶段目标检测框架 Faster R-CNN 则只差了 4.1 百分点,这说明了两阶段目标检测框架更适应开集检测任务。而本文所提出的算法则明显优于 HRE 和 ViLD,其在基类和新类上的效果都达到了最优,新类的 mAP@0.5 相对上述算法提高了 6.1 百分点。

尽管可以达到比较好的开集检测准确率,但是由于两阶段目标检测框架本身的局限性,该算法在模型的参数量和 FPS 上均弱于以一阶段检测器为基础的算法,因此,如何在兼顾检测准确率的同时提升其推理速度是未来研究的重点。

表 6 与其他的主流开集检测算法的结果对比

Table 6 Comparison of results with other mainstream open set detection algorithms

算法	检测器	VLP 模型	检测器参数量/ 10^6	mAP@0.5/%		FPS
				基类	新类	
HRE-LE	YOLOv2	—	50.54	29.6	19.5	92.3
HRE-CC	YOLOv2	—	50.54	58.7	24.5	93.2
HRE-H	YOLOv2	—	50.54	57.8	25.5	93.5
ViLD	SSD ^[28]	CLIP(ViT-B)	26.29	55.2	44.5	84.2
	YOLOv4-tiny	CLIP(ViT-B)	6.06	56.3	45.2	211.8
	YOLOv3	CLIP(ViT-B)	61.95	55.8	46.3	100.3
	YOLOv5s	CLIP(ViT-B)	47.08	57.4	49.6	110.9
	YOLOv8s	CLIP(ViT-B)	11.18	59.2	50.4	150.5
	Faster R-CNN	CLIP(ViT-B)	137.10	58.3	54.2	43.6
本文算法	OVR-CNN	VLP-LMN	126.30	59.4	60.3	48.2

2.4.4 泛化性能实验

为了验证 OFR-CNN 的泛化能力,使用服务区场景中的车辆的数据集进行目标定位实验,结果如表 7 所示,在服务区数据集上,即使只用基类数据进行训练,OFR-CNN 依然可以在新类上展示出良好的定位能力。

表 7 OFR-CNN 在交通数据集上的定位性能

Table 7 Localization performance of OFR-CNN on traffic dataset

网络	AP@0.5	AP@0.5:	AR@0.5:	AR@0.5:
		0.95	0.95@10	0.95@100
OFR-CNN	81.6	44.3	37.5	52.8

为了更直观地验证 VLP-LMN 强大的泛化能力,使用验证集的真实框图像进行标签匹配,最后得到的混淆矩阵如图 7 所示。可以看到,VLP-LMN 对于真实框的标签匹配准确度很高,即使在新类 bus 上的识别效果稍差,但也达到了 92.3%,总体的准确率达到了 99.8%。

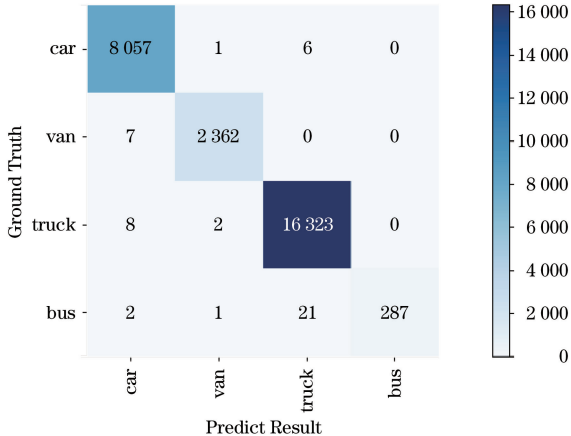


图 7 VLP-LMN 在验证集真实框上的混淆矩阵

Fig.7 Confusion matrix of VLP-LMN on the truth box of validation set

使用 VLP-LMN 对 OFR-CNN 产生的建议框区域图像匹配标签,得到最终的目标检测结果如表 8 和图 8 所示。可以看到,OFR-CNN 和 VLP-LMN 组合的目标检测框架具有良好的泛化能力。

表 8 OFR-CNN 和 VLP-LMN 在交通数据集上的性能

Table 8 Performance of OFR-CNN and VLP-LMN

on traffic dataset							%
基类 AP@0.5		新类 AP@0.5		mAP@0.5			
car	truck	van	bus	基类	新类	全部	
67.8	81.1	59.3	58.4	73.4	58.9	66.6	



图 8 开集交通目标检测算法效果

Fig.8 Effect of open set traffic object detection algorithm

3 结束语

本文提出一种基于 VLP 和 Faster R-CNN 的开集目标检测算法,该算法由 OFR-CNN 和 VLP-

LMN 两部分组成, OFR-CNN 可以解决开集目标的定位问题, 并改进了原始 Faster R-CNN 的损失函数, 使其 AP@0.5 提高了 2.1 百分点。VLP-LMN 基于 BLIP, 可以实现开集的标签匹配。实验结果表明, 所提算法不仅对于新类有检测能力, 并且和主流算法相比具有更好的开集检测性能。与零样本学习的方法相比, 其在新类上的 mAP@0.5 提高了 34.8%; 与 ViLD 相比, 其在新类上的 mAP@0.5 也提高了 6.1 百分点。除此之外, 该算法还具备良好的泛化能力, 在交通数据集上也展现出良好的开集检测性能, 其在新类上的 mAP@0.5 达到了 58.9%。上述实验结果说明, 该算法在具有开集目标的交通场景, 比如异常车型检测、智能驾驶等开放场景中有着巨大的应用潜力。但是, 由于 Faster R-CNN 自身的限制, 该算法对于小尺度、高重叠目标的检测性能一般, 运行速度也有待提升, 这些问题需要在未来进行更多的研究。

参考文献

- [1] 柳长源, 曹园园, 罗一鸣. 基于视频图像的车辆实时检测系统[J]. 计算机工程, 2019, 45(2): 265-269, 277.
LIU C Y, CAO Y Y, LUO Y M. Real-time vehicle detection system based on video image[J]. Computer Engineering, 2019, 45(2): 265-269, 277. (in Chinese)
- [2] 李明熹, 林正奎, 曲毅. 计算机视觉下的车辆目标检测算法综述[J]. 计算机工程与应用, 2019, 55(24): 20-28.
LI M X, LIN Z K, QU Y. Survey of vehicle object detection algorithm in computer vision[J]. Computer Engineering and Applications, 2019, 55(24): 20-28. (in Chinese)
- [3] 李松江, 耿兰兰, 王鹏. 基于改进 YOLOv4 的车辆目标检测[J]. 计算机工程, 2023, 49(4): 272-280.
LI S J, GENG L L, WANG P. Vehicle target detection based on improved YOLOv4 [J]. Computer Engineering, 2023, 49(4): 272-280. (in Chinese)
- [4] 杨秀璋, 武帅, 李娜, 等. 复杂环境下自适应去雾的 YOLOv3 汽车识别算法[J]. 计算机科学, 2023, 50(S2): 208-215.
YANG X Z, WU S, LI N, et al. YOLOv3 vehicle recognition algorithm for adaptive dehazing in complex environments[J]. Computer Science, 2023, 50(S2): 208-215. (in Chinese)
- [5] 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, 39(6): 1137-1149.
- [6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 779-788.
- [7] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 7263-7271.
- [8] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. [2023-12-01]. <https://arxiv.org/abs/1804.02767>.
- [9] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2023-12-01]. <https://arxiv.org/abs/2004.10934v1>.
- [10] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the KITTI dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [11] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 324-333.
- [12] ROMERA-PAREDES B, TORR P H S, ROMERA-PAREDES B, et al. An embarrassingly simple approach to zero-shot learning [C] // Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York, USA: ACM Press, 2015: 2152-2161.
- [13] BANSAL A, SIKKA K, SHARMA G, et al. Zero-shot object detection[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 397-414.
- [14] XIAN Y Q, AKATA Z, SHARMA G, et al. Latent embeddings for zero-shot classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 69-77.
- [15] ZAERIAN A, ROSA K D, HU D H, et al. Open-vocabulary object detection using captions[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 14393-14402.
- [16] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2021: 8748-8763.
- [17] GU X Y, LIN T Y, KUO W C, et al. Open-vocabulary object detection via vision and language knowledge distillation [EB/OL]. [2023-12-01]. <https://arxiv.org/abs/2104.13921v3>.
- [18] ZHONG Y W, YANG J W, ZHANG P C, et al. RegionCLIP: region-based language-image pretraining[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 16793-16803.
- [19] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of IEEE International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2021: 4904-4916.
- [20] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: vision and language representation learning with momentum distillation [C] // Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2021: 9694-9705.
- [21] LI J, LI D, XIONG C, et al. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2022: 12888-12900.
- [22] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 770-778.
- [23] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 356-366.
- [24] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]. Berlin, Germany:

- Springer, 2020.
- [25] YU J H, JIANG Y N, WANG Z Y, et al. UnitBox: an advanced object detection network [C]//Proceedings of the 24th ACM International Conference on Multimedia. New York, USA: ACM Press, 2016: 516-520.
- [26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. [2023-12-01]. <https://arxiv.org/abs/2010.11929>.
- [27] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-12-01]. <https://arxiv.org/abs/1810.04805v2>.
- [28] LIU W. SSD: single shot MultiBox detector[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2015:5675-567.
- [29] BERKAN D. Zero-shot object detection by hybrid region embedding [C] // Proceedings of British Machine Vision Conference. London, UK: [s. n.], 2018:431-442.
- 编辑 索书志

计算机工程
www.ecice06.com