嵌入标签语义的元特征再学习和重加权小样本目标检测

李鹏芳 刘 芳 李玲玲 刘 旭 冯志玺 焦李成 熊怡梦

（西安电子科技大学人工智能学院 西安 710071）

（教育部智能感知与图像理解重点实验室 西安 710071）

（国际智能感知与计算研究中心 西安 710071）

（国际智能感知与计算联合研究实验室 西安 710071）

摘 要 小样本目标检测(Few-Shot Object Detection,FSOD)中新类相对基类样本少,且新类和基类目标类别不同,导致 FSOD 方法存在学习到的新类特征判别性不强的问题.为了增强新类元特征的可分性,本文提出了一种嵌入标签语义的元特征再学习和重加权小样本目标检测方法.在小样本训练阶段,本文构建了一个词向量标签语义图产生模块.该产生模块引入标签语义信息生成了词向量标签语义图,用于建模基类和新类间的语义关联.同时,本文构建了一个标签语义嵌入模块.该嵌入模块融入基类和新类间的语义关联,对支持集样本的元特征进行再学习.该再学习过程能够将基类中与新类相关联的特征传递给新类,从而在只有少量新类样本的情况下学习到较好的新类元特征.通过端到端(End-to-End)的训练模型,本文方法增强了新类元特征的可分性,从而提升了新类目标的检测精度.在 PASCAL VOC 和 COCO 数据集上的对比和消融实验表明了本文方法的可行性与有效性.与 FSODFR 方法相比,在 PASCAL VOC 数据集上 2-shot 和 5-shot 下,我们方法的目标检测精度分别提高了 2.2%和 4.3%.

关键词 小样本学习;目标检测;小样本目标检测;元学习;标签语义;特征再学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02561

Meta-Feature Relearning with Embedded Label Semantics and Reweighting for Few-Shot Object Detection

LI Peng-Fang LIU Fang LI Ling-Ling LIU Xu FENG Zhi-Xi JIAO Li-Cheng XIONG Yi-Meng

(*School of Artificial Intelligent, Xidian University, Xi'an* 710071)

(*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xi'an* 710071)

(*International Research Center for Intelligent Perception and Computation, Xi'an* 710071)

(*Joint International Research Laboratory of Intelligent Perception and Computation, Xi'an* 710071)

Abstract Object Detection methods based on Deep Learning (DL) have been able to achieve good detection accuracy. Nevertheless, DL, as a data-driven technique, relies heavily on massive labeled data. Currently, Few-Shot Learning has been extensively studied, as it can alleviate the reliance on a large number of labeled samples. In this paper, we focus on the research on Few-Shot Object Detection (FSOD). In FSOD, the deep model is first learned on the base class data that

本课题得到国家自然科学基金项目(62076192)、陕西省重点研发计划(2019ZDLGY03-06)、国家自然科学基金国家重点项目(61836009)、长江学者及大学创新研究团队计划(IRT_15R53)、高等学校学科创新引智计划(B07048)、教育部重点科技创新研究项目、国家重点研发计划、CAAI 华为 MindSpore 开放基金等资助.**李鹏芳**,博士研究生,主要研究方向为图像处理和机器学习.E-mail:pf33li@163.com.**刘 芳**(通信作者),硕士,二级教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为人工智能和模式识别、机器学习、图像感知和场景理解、进化计算和数据挖掘.E-mail:f63liu@163.com.**李玲玲**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为量子进化优化学习、深度学习方法与应用、复杂遥感影像理解与解译.**刘 旭**,博士,讲师,主要研究方向为机器学习/深度学习理论、图像/视频处理方法.**冯志玺**,博士,副教授,主要研究方向为智能目标信息感知、机器学习.**焦李成**,博士,教授,博士生导师,中国计算机学会(CCF)会士,IEEE Fellow,主要研究方向为信号与图像处理、自然计算和智能信息处理.**熊怡梦**,硕士研究生,主要研究领域为图像处理和深度学习.

must have enough labeled samples. The model then continues to learn new classes with only a few labeled samples. The ultimate goal of model learning is to quickly adapt to the identification and localization of new classes of objects. In FSOD, on the one hand, compared with the base class, the new class has few samples available. On the other hand, the new class and the base class contain different target classes. This results in that FSOD methods are generally able to learn better base class features, but the learned new class features are weakly separable. To enhance the separability of meta-features for new classes, in this paper, we propose a Meta-Feature Relearning with Embedded Label Semantics and Reweighting Few-Shot Object Detection method. The proposed method can transfer the new class-related features from the base class to the new class, thereby enhancing the meta-features of the new class. In detail, in the few-shot training stage, firstly, we construct a word vector label semantic graph generation module. The word vector label semantic graph generation module introduces the label semantic information to generate the word vector label semantic graph. The generated word vector label semantic graph is used to model the semantic association between the base class and the new class. Meanwhile, we construct a label semantic embedding module. The label semantic embedding module first integrates the semantic association between the base class and the new class. Then, the meta-features of the support set samples are relearned based on the semantic association between the base class and the new class. The features associated with the new class in the base class can be passed to the new class when relearning the meta-features of the support set samples. Consequently, the proposed method can learn better meta-features of new classes when the new class has only a few samples. Finally, by training the model end-to-end, the proposed method enhances the separability of the features of the new category and thus improves the detection accuracy of the objects in the new category. Comparison and ablation experiments on the PASCAL VOC and COCO datasets demonstrate the feasibility and effectiveness of our method. Compared with the FSODFR method, the object detection accuracy of our method is improved by 2.2% and 4.3% on the PASCAL VOC dataset under 2-shot and 5-shot, respectively.

Keywords few-shot learning; object detection; few-shot object detection; meta-learning; label semantic; feature relearning

1 引 言

深度学习(Deep Learning,DL)下[1-4]的目标检测[5-7]方法已经能够获得很好的检测精度,但 DL 作为一项数据驱动的技术,严重依赖海量标记数据[8].人类通常能够仅通过少量带标记样本快速学习新概念.受到人类这种小样本学习能力的启发,研究人员开始关注小样本学习(Few-Shot Learning)方法[9-11]来缓解 DL 模型依赖大量标记数据的问题.FSOD 的目的是希望深度模型在学习一定具有足够标记样本的基类数据后,仅利用少量标记的新类样本快速适应新类目标的识别和定位.

现有的 FSOD 方法主要包括基于微调和基于元学习的方法.其中,基于元学习的 FSOD 方法又可以进一步分为基于原型和基于调制的方法.基于微调的 FSOD 方法[12-16]首先在基类数据上通过标准监督学习获得基模型,然后通过在基模型中引入少量参数或加入正则信息等策略,使得利用少量新类样本快速微调基模型能够识别新类目标.基于微调的 FSOD 方法需要研究者设计一定的微调策略来保证模型充分利用基类的知识,通过学习少量的新类样本便能快速适应新类目标的检测而不过拟合.基于原型的 FSOD 方法[17-19]为各类学习判别性的原型,利用类原型与图像特征间的相似度对感兴趣目标进行识别和定位.然而,在小样本场景中,模型推荐的新类目标的感兴趣区域(Region of Interest,ROI)的质量并不高,这给类原型的学习带来一定的挑战.基于调制的 FSOD 方法[20-28]按照 C-way N-shot(C 个类别且每个类别有 N 个样本)学习方式将可供学

习的数据组织为支持集样本和查询集样本，利用带标记的支持样本(或支持目标)的特征调制查询样本的特征来识别查询样本中包含的与支持目标类别相同的目标. 这种 C-way N-shot 学习方式被普遍认为是适应 FSOD 场景的，引起广泛研究. 本文基于调制的 FSOD 框架展开研究.

在 FSOD 任务中，与基类数据相比，新类数据对于 FSOD 模型是不可见或少量可见的. FSOD 模型通常能够学习到较好的基类表示，而学习到的新类表示可判别性不强，导致对新类目标的识别精度较低. 为了缓解该问题，文献[19]方法引入标签语义，通过动态关系图在语义空间进行语义关系推理学习各类的原型实现样本稳定的 FSOD. 文献[19]方法通过引入标签语义学习到样本稳定的类原型，表明标签语义信息用于学习表示不同的类是关键重要的. 但文献[19]方法缺乏对不同类样本的视觉表征的应用，导致可用于区分不同类的重要的视觉信息被损失. 文献[29-30]方法通过支持图像和查询图像互相指导的方式来增强特征的判别性，即通过引入不同的注意力方式将支持图像的特征融合到查询图像特征中或将查询图像的特征融合到支持图像特征中. 我们发现，这种支持图像特征和查询图像特征融合并不是始终是正向的，如将“sofa”类的支持图像特征与包含“chair”的查询图像的特征进行融合，则会增加将查询图像中的“chair”误判为“sofa”的概率，从而导致预测错误. 文献[31]方法通过直接引入支持图像特征的可分性约束来增强其可分性取得不错的 FSOD 效果，表明我们可以通过直接增强支持图像特征的判别性来提高 FSOD 的精度.

基于以上分析和发现，借鉴人类在学习一种新的概念时，尤其是在只有少量的可供学习的样本的情况下，往往需要借助以往经验以及事物之间的关联性进行学习的方式，我们提出了一种嵌入标签语义的元特征再学习和重加权小样本目标检测方法. 本文方法通过引入标签语义并计算它们的词向量之间的相似度构造词向量标签语义图建模基类与新类元特征在语义空间的关联性. 将词向量标签语义图与支持图像的元特征进行融合，基于基类与新类元特征在语义空间的关联性，本文利用图卷积网络(Graph Convolutional Network，GCN)对支持图像的元特征进行再学习，将基类中与新类相关的特征传递给新类，增强新类元特征的判别性.

本文的贡献可以概述为：

(1) 本文方法构建了一个词向量标签语义图产生模块，利用标签语义生成词向量标签语义图，建模了基类和新类间的语义关联.

(2) 本文方法提出了一个标签语义嵌入模块，用于再学习支持图像的元特征，将基类中与新类相关的特征传递给新类，增强了新类元特征的判别的表达能力.

(3) 在 PASCAL VOC 和 COCO 数据集上的对比和消融实验表明了本文方法的有效性. 与 FSODFR 方法相比，我们的方法取得了具有竞争性的 FSOD 精度.

2　相关工作

2.1　目标检测

DL 下的目标检测方法包括两大类：两阶段(Two-Stage)和一阶段(One-Stage)的目标检测方法.

Two-Stage 的目标检测方法首先从图像中生成大量的可能存在感兴趣目标的区域提案，然后基于这些区域提案的特征进行目标的分类和定位，以此来实现目标检测. Two-Stage 的目标检测方法的代表性工作是 R-CNN(Region Convolutional Neural Network)[32]、Fast R-CNN[33]、Faster R-CNN[34] 和 Mask R-CNN[35] 这一系列工作. R-CNN[32] 是基于 DL 的目标检测方法的开创之作，该方法包含四个部分：(1) 基于选择式搜索(Selective Search，SS)方法生成目标的 ROI；(2) 将这些 ROI 缩放至固定大小，利用 GCN 提取其特征；(3) 基于这些特征，训练类特定的线性的支持向量机(Support Vector Machine)分类器；(4) 学习边界框回归器用于准确的定位目标. R-CNN[32] 提取不同区域提案的特征时并没有共享卷积神经网络(Convolutional Neural Network，CNN)的参数，导致其计算效率不高. Fast R-CNN[33] 先提取图像的特征，将其输入感兴趣区域池化层(Region of Interest Pooling Layer)提取 ROI 的固定大小的特征，并将其输入到由全连接层(Fully Connected Layers)构成的分类器和回归器中预测目标的类别和位置. Fast R-CNN[33] 虽然通过权值共享显著提高了运算效率，但其仍是需要先通过传统方法来生成区域提案. Faster R-CNN[34] 提出区域提案网络(Region Proposal Network，RPN)端到端地生成不同尺度和纵横比的 ROI，进一步提升了算法的计算效率. Mask R-CNN[35] 通过引入实例掩码预测分支来实现更精准的目标检测.

One-Stage 的目标检测方法将图像作为输入直

接预测图像中感兴趣目标的位置和类别，避免区域提案生成阶段. 其代表性工作是 YOLO(You Only Look Once)[36]及其不同的改进版本和 SSD(Single Shot Detection)[37]模型. YOLO[36]将目标检测作为回归问题，使用一个统一的架构从图像中提取特征并预测目标的边界框位置和类别概率. YOLOv2[38]通过引入各种策略如批归一化(Batch Normalization，BN)、多尺度训练和卷积预测边界框，在确保检测速度的前提下进一步提升了目标检测的精度. YOLOv2[38]还引入 Darknet-19 作为特征提取器降低了模型计算复杂度. 为了达到更好的分类效果，YOLOv3[39]中设计了更深的网络 Darknet-53 作为特征提取器. 为了提升小目标的检测精度，YOLOv3[39]中采用类似特征金字塔(Feature Pyramid)的上采样(Upsample)融合策略在多个尺度的特征上进行目标检测. 与 YOLOv2[38]类似，YOLOv4[40]通过结合一系列前人研究的技术，如加权残差连接(Weighted Residual Connection)、跨阶段部分连接(Cross Stage Partial Connections)和自对抗训练(Self-Adversarial Training)等实现了目标检测精度的进一步提升. YOLOv5① 在 YOLOv4[40]的基础上，提出一些新的改进思路，如自适应锚框(Anchor)计算、用 GIOU(Generalized Intersection over Union)损失替换平滑(Smooth)L1 损失等提升了目标检测的速度和精度. SSD[37]遵循 YOLO[36]将目标检测转换为回归任务的思路，一次完成目标的定位与分类. 同时，SSD[37]借鉴 Faster R-CNN[34]中的 Anchor 提出先验框(Prior Box). 此外，SSD[37]还基于特征金字塔网络(Feature Pyramid Network，FPN)在多个尺度上检测目标. One-Stage 的目标检测方法具有速度较快的优势，由于计算资源的限制，本文的目标检测框架使用具有浅层特征提取器的 YOLOv2[38]目标检测模型.

基于上述经典的目标检测模型，针对目标检测数据中目标尺度不一、目标遮挡和目标姿态多样等数据特性，借助新的技术和手段如注意力机制、GCN 等，许多研究者提出众多改进的目标检测方法. 文献[41]方法引入标签语义信息，提出基于相似度的知识迁移模型用于半监督的目标检测. 该方法在视觉表征和标签语义上分别计算强标注类与弱标注类的样本间的相似度进行不同强度的知识迁移. 该方法表明视觉相似性和语义相关性对任务是互补的，当两者结合时，能够显著提高检测的性能.

2.2 小样本目标检测

基于微调的方法. Chen 等人[12]最先提出基于微调的 FSOD 方法. 为了适应 FSOD 场景，文献[12]方法将标准监督学习中的目标检测模型 SSD[37]中多尺度预测和 Faster R-CNN[34]中由粗到细的预测两种优势结合起来，提出一个新的 FSOD 模型. 随后，针对 FSOD 中目标尺度不一的问题，文献[13]方法提出一个细化分支执行并约束多个尺度的目标检测结果. 文献[14]方法提出元学习框架下 Two-Stage 微调的 FSOD 方法. 为了提升新类目标的检测精度，文献[15]方法在微调阶段冻结基类的 RPN 和基类的检测器，学习微调适用于新类的新的 RPN 和检测器，同时加入新的检测器对基类目标的预测与基类检测器对基类目标的预测的一致性损失. 文献[16]方法通过图结构建模图像中 ROI 之间的关系，通过 GCN 传递不同 ROI 之间的消息，将信息交互后的特征与原特征级联作为检测和分类头的输入来提升 FSOD 的性能. 该方法通过 GCN 建模了图像中不同目标的视觉特征间的关系，但是同类目标的视觉特征因为目标的姿态和纹理等变化而具有不稳定性，这会给模型学习带来一定的困难. 标签语义信息具有类内和类间不变性，我们引入这种不变性，来增强不同类视觉特征的判别性.

基于原型的方法. Karlinsky 等人[17]最先提出基于原型的 FSOD 方法. 该方法将标量 1 作为全连接层的输入，通过端到端的训练，学习获得不同类的原型，最终通过计算类原型与目标嵌入特征间的相似度对感兴趣目标进行识别和定位. 基于原型的方法通过学习更好的类原型来提高 FSOD 精度. 基于文献[17-18]方法进一步挖掘 ROI 的负表示用于 FSOD. 文献[19]方法基于标签语义在语义空间进行语义关系推理学习各类的语义级别原型实现样本稳定的 FSOD. 文献[19]方法缺乏对视觉特征的应用，我们的方法同时利用语义和视觉信息再学习元特征.

基于调制的方法. Kang 等人[20]和 Hsieh 等人[21]分别基于 One-Stage 的目标检测模型 YOLOv2[38]和 Two-Stage 的目标检测模型 Faster R-CNN[34]提出基于调制的 FSOD 方法. 基于这两个基础的方法，许多研究者展开进一步的研究. 文献[22-24]方法通过引入不同的注意力机制从如何更好的调制查询图

① https://github.com/ultralytics/yolov5

像的特征的角度来改进这些工作. 文献[22]方法从支持集图像中学习各类目标的原型，将该原型与查询图像的特征融合得到类特定的查询图像特征用于推荐查询图像中与各支持目标类别相似的区域提案. 文献[23]方法将支持图像的特征经由全局平均池化后作为卷积核与查询图像的特征作相关运算得到注意力特征，从该注意力特征中生成查询图像中存在的与支持目标类别相同的目标的位置. 文献[24]方法学习显著的支持图像的特征，将其与查询图像的特征融合来生成区域提案. 还有一些方法通过各种机制学习更判别的新类特征来提升 FSOD 的精度. 文献[25]方法通过引入可变形卷积和注意力机制学习更判别的特征. 文献[26]方法通过对支持图像的特征和查询图像的特征添加特征点级别的关系蒸馏来学习更判别的特征，并利用多尺度特征融合预测提升 FSOD 精度. 文献[27-28]方法利用对比自监督学习增强模型的判别能力来提高 FSOD 的精度. 不同之处在于文献[27]方法是在图像级别实现对比自监督学习，而文献[28]方法是在 ROI 级别实现对比自监督学习.

2.3　词嵌入模型

词嵌入是自然语言处理(Natural Language Processing，NLP)中语言模型和表征学习技术的统称. 在 NLP 中，词是最小的处理单元. 为了便于计算，需要将词这种抽象的符号表示转换为可计算的数学量的表示，即学习词嵌入. 比较常用的用于生成词的嵌入表示的模型有 Word2Vector[42] 和 GloVe(Global Vectors for Word Representation)[43]. 由于从零开始训练一个词嵌入模型是非常耗时耗力的，因此，通常直接利用从大规模文本库预训练的词嵌入模型提取词的嵌入表示用于下游任务. 本文基于预训练的 GloVe 模型获取标签语义的嵌入表示.

3　嵌入标签语义的元特征再学习和重加权小样本目标检测

本节将通过 FSOD 任务定义、方法整体功能模块、词向量标签语义图产生模块的细节、标签语义嵌入模块的详细结构、总的损失函数来介绍本文方法的具体方案和步骤.

3.1　任务定义

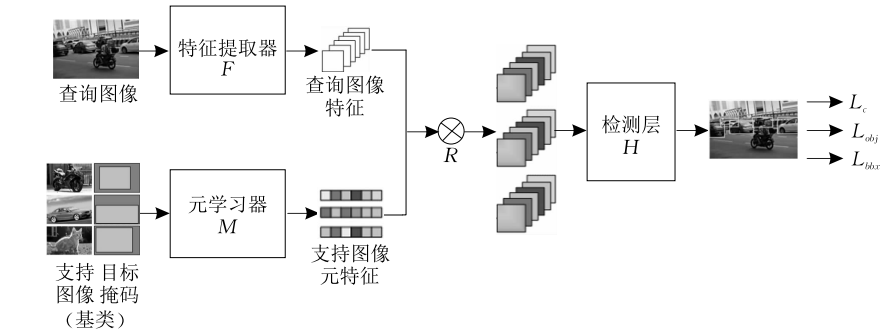
用于学习 FSOD 模型的数据有各个类别都有大量标记样本的基类数据 D_{base} 和各个类别仅有 K 个标记样本的新类数据 D_{new}. 将它们中包含的目标类别分别记为 C_{base} 和 C_{new}，其中，基类和新类的类别无交集，即 $C_{\text{base}} \cap C_{\text{new}} = \varnothing$. 对于单个样本 $(x, y) \in \{D_{\text{base}} \cup D_{\text{new}}\}$，$x = \{o_i \mid i = 1, 2, \cdots, n\}$ 是包含 n 个目标的图像，其中 o_i 表示第 i 个目标. $y = \{(c_i, l_i) \mid i = 1, 2, \cdots, n\}$ 是图像 x 的标签，其中 c_i 和 l_i 分别是图像中第 i 个目标的类别和位置. FSOD 的目标是利用具有大量标记样本的基类数据和少量标记样本的新类数据对模型进行训练，要求训练所得的模型能够对图像中的新类目标进行准确的分类和定位.

基于调制的 FSOD 方法包含两个训练阶段：基类训练阶段和小样本训练阶段. 在基类训练阶段，利用具有大量标记样本的基类数据 D_{base} 学习模型. 在小样本训练阶段，由于新类中每个类别仅有 K 个标记样本，为了避免模型过拟合基类，仅利用基类中每类 K 个样本和新类数据共同训练模型，记为 $D_{\text{balance}} \subset \{D_{\text{base}} \cup D_{\text{new}}\}$. 在两个训练阶段，分别从对应可用的数据集合中构建 FSOD 学习任务 $T_j = S_j \cup Q_j = \{s_1, s_2, \cdots, s_{|C|}\} \cup \{q_1, q_2, \cdots, q_{|Q_j|}\}$ 来学习模型，其中 $|C|$ 表示各个学习阶段所使用的训练数据中包含的目标的类别数，S_j 表示从 $|C|$ 个类别中每类采样一个图像组成的支持集样本，s_i 表示支持集 S_j 中的第 i 个支持图像. Q_j 表示从 $|C|$ 个类中采样的查询集图像，$|Q_j|$ 表示查询集中查询图像的个数，q_i 表示查询集 Q_j 中的第 i 个查询图像. 此外，我们的方法利用了标签语义信息，记为 $B = \{b_1, b_2, \cdots, b_{|C|}\}$，其中 b_i 表示第 i 个类别的标签.

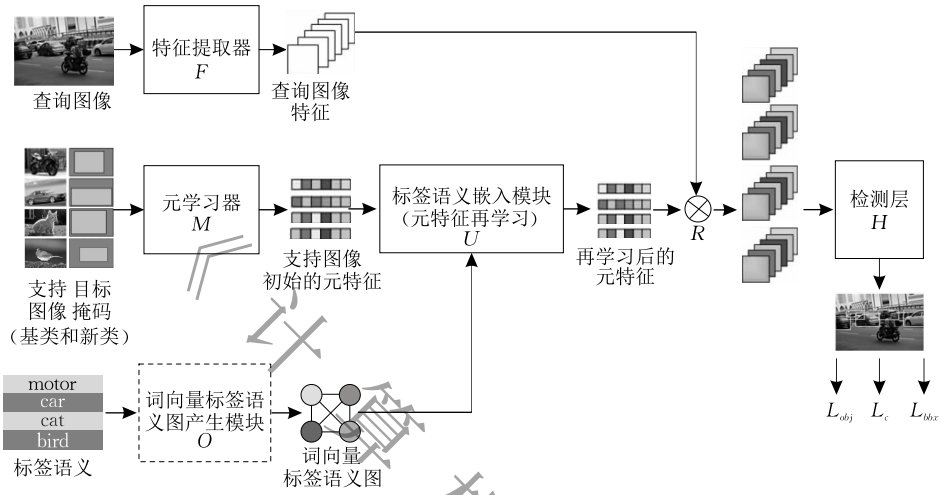
3.2　整体功能模块结构

本文基于调制的 FSOD 框架，以 FSODFR 方法[20]为基础模型，提出了一种嵌入标签语义的元特征再学习和重加权小样本目标检测方法. 所提出的方法在小样本训练阶段引入标签语义信息，构建词向量标签语义图产生模块生成词向量标签语义图，从而建模新类和基类的语义联系. 同时，本文构建标签语义嵌入模块，将该语义关联融入到元特征的再学习过程中. 在再学习过程中，将基类中与新类相关的知识更新到新类，提高新类元特征的判别性，从而提升新类目标的检测效果.

我们方法的整体功能模块结构如图 1 所示. 图 1(a)表示基类训练阶段模型，包括提取查询图像特征的特征提取器 F、提取支持图像元特征的元学习器 M、利用支持图像元特征调制查询图像特征的重权值操作 R、分类和检测头 H，共 4 个功能模块. 图 1



(a) 基类训练阶段：利用具有大量标记样本的基类数据对模型进行训练



(b) 小样本训练阶段：利用基类和新类每类有K个标记样本的数据对模型进行训练

图 1 嵌入标签语义的元特征再学习和重加权小样本目标检测方法的整体功能模块结构

(b)表示小样本训练阶段模型，除了包含与基类训练阶段共享的 4 个模块外，我们额外构建了词向量标签语义图产生模块 O 和用于元特征再学习的标签语义嵌入模块 U，共 6 个功能模块. 词向量标签语义图产生模块利用标签语义生成词向量标签语义图建模不同类别间的语义联系. 标签语义嵌入模块基于新类与基类间的语义关联，将基类中与新类相关的特征传递给新类. 通过端到端的训练，模型最终学习获得判别的再学习后的新类元特征.

小样本训练阶段与基类训练阶段相比，增加元特征再学习过程. 我们以小样本训练阶段为例介绍我们方法的整体功能模块流程. 具体为：首先利用特征提取器 F 获得查询集图像 Q 的特征 $\overline{\boldsymbol{Q}}\in R^{|Q|\times 13\times 13\times 1024}$，其中，$|Q|$ 表示查询图像的个数，13×13 表示特征的宽和高，1024 表示特征的通道数. 除特别说明外，其余特征不同维度的物理含义与此处相同. 此外，本文中统一使用“$*$”表示数值相乘，使用“$\times$”表示分隔不同维度. 本文方法使用的特征提取器 F 的网络结构为 DarkNet19. 利用元学习器 M 获得支持集样本 S 的初始的元特征 $\boldsymbol{X}^0\in R^{|C|\times 1\times 1\times 1024}$. 元学习器 M 的网络结构为 7 层卷积和池化相交替的网络. 根据标签语义 B 利用词向量标签语义图产生模块 O 构建词向量标签语义图 G_{vec}. 该产生模块直接使用预训练好的用于提取词嵌入的 GloVe 模型①，不再参与本文模型学习的过程. 随后，将支持集样本的初始的元特征 $\boldsymbol{X}^0$ 与构建的词向量标签语义图 G_{vec} 作为标签语义嵌入模块 U 的输入利用 GCN 再学习支持集样本的初始的元特征 $\boldsymbol{X}^0$. 最后，再学习后的元特征 $\boldsymbol{X}^*\in R^{|C|\times 1\times 1\times 1024}$ 作为重加权向量对查询集图像的特征 $\overline{\boldsymbol{Q}}$ 进行重权值 R 操作，得到调制后的查询集图像特征 $\hat{\boldsymbol{Q}}\in R^{|C|*|Q|\times 13\times 13\times 1024}$. 将其作为检测和分类头 H 的输入，得到查询集图像中目标的分类和定位结果 $\boldsymbol{y}^*$.

3.3 词向量标签语义图产生模块

本文利用标签语义 B 通过词向量标签语义图产生模块 O 构建词向量标签语义图 G_{vec}，建模类别间的语义关联. 词向量标签语义图产生模块构建词向量标签语义图 G_{vec} 的细节如图 2 所示. 首先将标签语义 B 中的各个标签 b_i（如“cat”）通过预训练的 GloVe

① https://github.com/stanfordnlp/GloVe

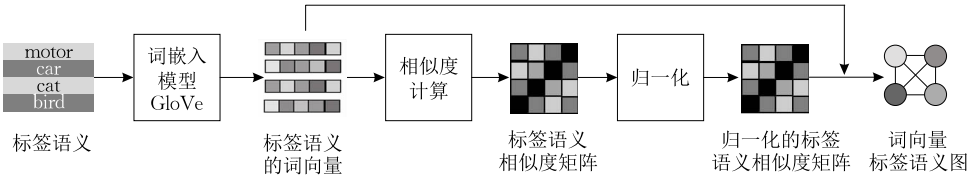


图 2　词向量标签语义图产生模块细节

模型映射为各标签对应的词向量 $\boldsymbol{vec}_{b_i}\in R^{300}$，得到全部标签语义的词向量集合 $V_{vec}=\{\boldsymbol{vec}_{b_1},\boldsymbol{vec}_{b_2},\cdots,\boldsymbol{vec}_{b_{|C|}}\}$. 然后通过计算各标签对应词向量之间的余弦距离得到标签语义相似度矩阵 $\boldsymbol{A}_{vec}\in R^{|C|\times|C|}$，度量类别之间的语义关联程度. 标签语义相似度矩阵 $\boldsymbol{A}_{vec}$ 中各个元素 $A_{vec}^{i,j}$ 的值的具体计算方式为

$$A_{vec}^{i,j}=\frac{\sum_{e=1}^{E}vec_{b_i}^{e}\times vec_{b_j}^{e}}{\sqrt{\sum_{e=1}^{E}(vec_{b_i}^{e})^2}\times\sqrt{\sum_{e=1}^{E}(vec_{b_j}^{e})^2}} \tag{1}$$

其中 $vec_{b_i}^{e}$ 表示第 i 个标签对应的词向量 $\boldsymbol{vec}_{b_i}$ 的第 e 个维度的值，E 表示词向量的维度，$E=300$. 当前各标签之间的语义相似度通过其对应的词向量间的余弦距离来度量，即 $A_{vec}^{i,j}$ 取值范围为 $[-1,1]$. 为了避免后续图卷积计算过程导致特征的尺度改变，我们将该取值范围缩放到 $[0,1]$ 范围，具体的归一化计算方式为

$$\widetilde{A}_{vec}^{i,j}=\frac{\sum_{i=1}^{|C|}A_{vec}^{i,j}-\mathrm{Min}(\boldsymbol{A}_{vec}^{i,*})}{\mathrm{Max}(\boldsymbol{A}_{vec}^{i,*})-\mathrm{Min}(\boldsymbol{A}_{vec}^{i,*})} \tag{2}$$

其中，$\mathrm{Min}(\boldsymbol{A}_{vec}^{i,*})$ 和 $\mathrm{Max}(\boldsymbol{A}_{vec}^{i,*})$ 分别指第 i 个标签与各个标签语义相似度中的最大值和最小值，最终得到归一化的标签语义相似度矩阵 $\widetilde{\boldsymbol{A}}_{vec}$.

在得到归一化的标签语义相似度矩阵 $\widetilde{\boldsymbol{A}}_{vec}$ 之后，构建词向量标签语义图 G_{vec}. G_{vec} 是一个全连接的无向图，G_{vec} 的结点集合为全部标签语义的词向量集合 V_{vec}，G_{vec} 的邻接矩阵为归一化的标签语义相似度矩阵 $\widetilde{\boldsymbol{A}}_{vec}$. 使用词向量标签语义图能够建模基类和新类之间的语义关联.

3.4　标签语义嵌入模块

标签语义嵌入模块的结构如图 3 所示，通过包含一个图卷积层的 GCN 来进行消息传递，将基类中与新类相关的特征传递给新类，学习较好的新类元特征. 在 GCN 之后，添加 BN 和跳跃连接. 将图卷积操作之前的输入特征传递到图卷积操作之后，降低在消息传播过程中引入的噪声信息的影响. 标签语义嵌入模块的输入是词向量标签语义图 G_{vec} 和支持集样本的初始的元特征 $\boldsymbol{X}^0$. 首先，将词向量标签语义图 G_{vec} 和支持集样本的初始的元特征 $\boldsymbol{X}^0$ 融合生成初始的标签语义图 G. 具体是用各类支持集样本的初始的元特征替换词向量标签语义图 G_{vec} 中对应的标签的词向量（也就是对应的结点）. 即此时，初始的标签语义图的结点为支持样本初始的元特征，边为结点特征对应的标签之间的语义相似度. 在更新标签语义图时，我们仅更新图结点对应的支持样本的初始的元特征的信息，而不对图的边进行更新. 本文只利用最终的图结点特征进行后续操作.

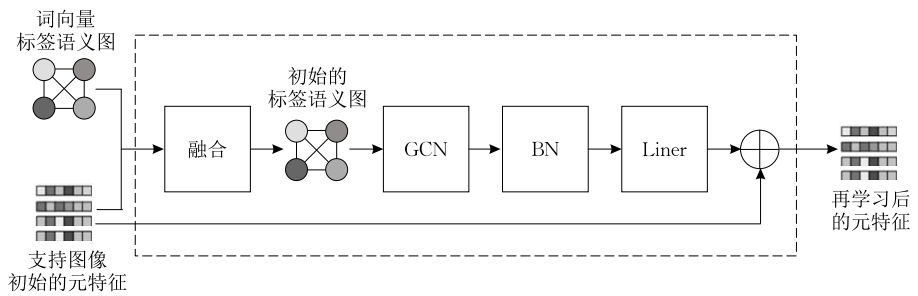


图 3　标签语义嵌入模块细节

通过在图上进行卷积操作实现图结点集即支持集初始的元特征的再学习，具体计算方式为

$$\boldsymbol{X}^{*}=\sigma(BN(\widetilde{\boldsymbol{A}}_{vec}\boldsymbol{X}^{0}\boldsymbol{\Phi}))+\boldsymbol{X}^{0} \tag{3}$$

其中，$\boldsymbol{X}^0$ 表示支持集样本的初始的元特征，$\widetilde{\boldsymbol{A}}_{vec}$ 表示归一化的标签语义相似度矩阵，$\boldsymbol{\Phi}\in R^{1024\times1024}$ 是标签语义嵌入模块的参数，$BN(\cdot)$ 表示 BN 函数，$\sigma(\cdot)$ 表示激活函数. 本文方法中默认使用线性激活函数，表示为 Liner-GCN. 我们初始化 $\boldsymbol{\Phi}$ 中每个元素均为

0,即在小样本训练阶段模型初次前向传播时 $\boldsymbol{X}^*=\boldsymbol{X}^0$.随后通过反向传播更新 $\boldsymbol{\Phi}$,自适应的计算不同类别间支持样本的元特征的融合强度.

在小样本训练阶段,本文方法使用 GCN 构建标签语义嵌入模块.利用 GCN 消息传递机制,将基类元特征中相关的信息融合到新类支持样本的元特征中,从而提高模型对新类目标的检测效果.

3.5 总的损失函数

用于优化模型的目标函数 L 为

$$L=L_c+L_{bbx}+L_{obj} \tag{4}$$

其中,L_c 为预测的查询集图像中目标类别与真实目标类别的交叉熵损失、L_{bbx} 为预测的查询集图像中目标的位置与真实目标位置的均方差损失、L_{obj} 为预测出的查询集图像中目标框与预先设定好的 Anchor 的 IoU 和真实坐标与预设的 Anchor 的 IoU 的均方差损失.在基类训练阶段和小样本训练阶段,使用相同的目标函数端到端的联合优化模型全部可学习的模块的参数.如图 1 所示,用实线边框标记的模块是需要学习并更新其参数的模块,用虚线边框标记的词向量标签语义图产生模块 O 直接使用 NLP 中预训练的参数,本文方法中不再对其进行学习.以基类训练阶段学习得到的模型为基础,小样本训练阶段仅需少量的迭代模型便能快速收敛.

4 实 验

4.1 数据集

PASCAL VOC. 本文采用 PASCAL VOC 2007+2012 数据集[44]进行实验.该数据集共有 20 个目标类别,整个训练集包括 PASCAL VOC 2007 和 2012 的训练集和验证集,共 16551 幅图像.测试集为 PASCAL VOC2007 测试集,共 4952 幅测试图像.按照 FSODFR 方法[20]中新类和基类类别划分方式,本文将“bird,bus,cow,motor,sofa”5 个目标类别作为新类,将剩余 15 个目标类别“bicycle,train,boat,car,aeroplane,horse,cat,dog,sheep,person,bottle,chair,dining table,tv monitor,potted plant”作为基类进行实验.

COCO. COCO 2014 数据集[45]共包含 123287 幅图像.该数据集共有 80 个目标类别.受计算资源的限制,本文参照文献[46]方法,抽取 COCO 数据集的子集进行实验.与文献[46]方法相同,本文定义类别号为{1,2,4,5,7,8,9,19,20,21,23,24,25,62,63,64,65,70}的 18 个类作为基类,定义类别号为{3,6,16,17,18,22,67}的 7 个类作为新类.删除不包含这 25 个类的目标的图像后,所抽取的 COCO 数据集的子集共包含 106287 幅图像.从中随机抽取 5000 幅图像作为测试集,其余图像则作为训练集.对于 COCO 数据集中和 PASCAL VOC 数据集标签命名不同但实际属于相同类别的类,本文统一使用 VOC 数据集中的标签命名来生成词向量.

4.2 实验设置

在基类训练阶段和小样本训练阶段,我们均使用随机梯度下降(Stochastic Gradient Descent)优化器进行参数的更新与优化.基类训练阶段的学习率为 0.00033,共迭代模型 170 代.小样本训练阶段的学习率为 0.00067,共训练模型 10 代.本文所使用的超参数与 FSODFR 方法提供的公开代码中的超参数相同.但由于计算资源的限制,在两个训练阶段,我们都没有采用多尺度训练策略,而是将输入图像的尺寸设置为 416×416 的固定大小.构建 FSOD 学习任务 T_j 时,针对 PASCAL VOC 和 COCO 数据集,我们分别设置查询集中查询图像的个数 $|Q_j|$ 为 16 和 32.为了增大批大小,在经过 4 次 FSOD 学习任务后我们才更新模型,采用梯度累积变相增大批大小.

程序运行的操作系统版本为 Ubuntu16.04,DL 计算框架为 Pytorch,代码全部用 Python 语言实现.本文实验所用的硬件条件为:2 个 8 核的 Intel 至强 E5-2650CPU,主频为 2.4 GHz,内存 64 GB,GPU 为 Nvidia TiTan X,GPU 显存为 12GB.

4.3 实验结果及分析

4.3.1 对比实验

本文方法与其他方法在 PASCAL VOC 数据集上的对比实验结果如表 1 所示.为了更直观的比较不同的 FSOD 方法,我们在表 1 中标注了不同方法的类别、采用的检测框架、所使用的特征提取器的网络结构(N/S 表示支持图像和查询图像不共享特征提取器).受益于 Two-stage 的目标检测框架 Faster R-CNN 和更深的特征提取器 ResNet101,TFA 方法[14]比 LSTD 方法[12]获得显著的性能增益.我们的方法以 FSODFR 方法[20]为基线(Baseline),但由于计算资源的限制,我们的方法并未使用能带来性能增益的多尺度的图像训练策略.并且,我们使用了会损害少量精度的梯度累积策略变相扩大了批大小,虽然会损失一定的精度,但使用梯度累积策略比直接降低批大小性能会更好.我们在同样的实验设置下,复现了 FSODFR 方法[20]和 CME 方法[31],分别表示为 FSODFR* 和 CME*.对比我们复现的结

果可以看出，Liner-GCN 方法在所有 shot 下，检测精度均优于 FSODFR*. 与 CME* 相比，Liner-GCN 在 5-shot 和 10-shot 设置下分别提升了 5%和 3.6%的准确率. 与 FSODFR 方法[20]直接相比，虽然 Liner-GCN 在 1-shot、2-shot 和 3-shot 设置下性能降低，但在 5-shot 和 10-shot 设置下，Liner-GCN 取得了具有竞争性的检测精度. 此外，基于 CME 方法[31]开源的在基类训练阶段获得的基模型①，我们实现了 FSODFR 方法[20]和本文方法 Liner-GCN，分别表示为 FSODFR-CME 和 Liner-GCN-CME. 本文方法、FSODFR 方法[20]和 CME 方法[31]在基类训练阶段具有相同的模型结构. 与 FSODFR* 和 Liner-GCN 相比，虽然 FSODFR-CME 和 Liner-GCN-CME 使用了更好的基模型，但在小样本训练阶段我们仍采用了会损害少量精度的梯度累积策略，并且未使用能带来性能增益的多尺度训练策略，这仍然会限制 FSODFR-CME 和 Liner-GCN-CME 的性能. 将 FSODFR-CME 和 Liner-GCN-CME 分别与 FSODFR* 和 Liner-GCN 相比，除了在 1-shot 条件下，FSOD 性能均有提升，表明更好的训练设置确实能够带来性能增益，这也间接说明我们复现的有效性. 实验结果表明，基于不同支持类别间的语义联系，我们的方法在不同 shot 设置下，通过标签语义嵌入模块，不同支持类的元特征能够继承其它类别的有用信息. 利用再学习后的元特征对查询图像特征进行调整，能有效提高 FSOD 的精度.

表 1　本文方法与其他方法在 PASCAL VOC 上不同 shot 下新类的平均检测精度比较结果　（单位：%）

类别	方法	检测框架	特征提取器	1-shot	2-shot	3-shot	5-shot	10-shot
基于微调	YOLOv2-joint[20]	ICCV2019	YOLOv2	Q:DarkNet19 S:N/S	0.0	0.0	1.8	1.8	1.8
	YOLOv2-ft[20]	ICCV2019	YOLOv2	Q:DarkNet19 S:N/S	3.2	6.5	6.4	7.5	12.3
	LSTD(YOLOv2)[20]	ICCV2019	YOLOv2	Q:DarkNet19 S:N/S	6.9	9.2	7.4	12.2	11.6
	TFA[14]	ICML2020	Faster R-CNN	ResNet101	39.8	36.1	44.7	55.7	56.0
	MPSR[13]	ECCV2020	Faster R-CNN	ResNet101	41.7	—	51.4	55.2	61.8
	Retentive R-CNN[15]	CVPR2021	Faster R-CNN	ResNet101	42.4	45.8	45.9	53.7	56.1
	FSCE[28]	CVPR2021	Faster R-CNN	ResNet101	44.2	43.8	51.4	61.9	63.4
	FSOD-SR[16]	PR2021	Faster R-CNN	ResNet-50	50.1	54.4	56.2	60.0	62.4
基于原型	RepMet[17]	CVPR2019	Faster R-CNN	Inception	26.1	32.9	34.4	38.6	41.3
	NP-RepMet[18]	NeurIPS2020	Faster R-CNN	ResNet101	37.8	40.3	41.7	47.3	49.4
	SRR-FSD[19]	CVPR2021	Faster R-CNN	Q:ResNet101 S:Knowledge Graph	47.8	50.5	51.3	55.2	56.8
基于调制	FSODFR[20]	ICCV2019	YOLOv2	Q:DarkNet19 S:N/S	14.8	15.5	26.7	33.9	47.2
	CME[31]	CVPR2021	YOLOv2	Q:DarkNet19 S:N/S	17.8	26.1	31.5	44.8	47.5
	DCNet[26]	CVPR2021	Faster R-CNN	ResNet101	33.9	37.4	43.7	51.1	59.6
	FSODFR*[20]	ICCV2019	YOLOv2	Q:DarkNet19 S:N/S	13.7	16.6	21.8	30.0	42.8
	CME*[31]	CVPR2021	YOLOv2	Q:DarkNet19 S:N/S	17.2	21.2	26.7	33.2	43.9
	FSODFR-CME		YOLOv2	Q:DarkNet19 S:N/S	11.4	16.9	22.9	36.3	44.7
	本文方法(Liner-GCN)		YOLOv2	Q:DarkNet19 S:N/S	14.4	17.7	22.3	38.2	47.5
	本文方法(Liner-GCN-CME)		YOLOv2	Q:DarkNet19 S:N/S	13.2	21.8	25.3	41.9	48.0

本文方法与其他方法在 COCO 数据集上的性能比较如表 2 所示. 与 FSODFR[20]和 CME[31]相比，本文方法虽然能够获得较好的 FSOD 性能，但获得的检测精度值很低. 这可能是因为与原始 COCO 数据集相比，本文利用的子数据集中的图片包含未被抽取的类的目标的干扰，这给 FSOD 带来巨大挑战.

本文方法与其他方法在 PASCAL VOC 数据集上不同 K 值下各类的平均检测精度如表 3 所示. 从表 3 可以看到，与 FSODFR* 方法相比，1/2/3-shot 时，本文方法 Liner-GCN 在几乎不损失其他新类的检测精度的情况下能够显著增加某个新类的检测精

① https://github.com/Bohao-Lee/CME

表 2 本文方法与其他方法在 COCO 数据集上不同 shot 下新类的平均检测精度比较结果 (单位:%)

K	方法	Average Precision						Average Recall					
		0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
10	FSODFR[20]	0.1	0.4	0.1	0.0	0.2	0.2	2.8	3.4	3.4	0.1	1.8	4.6
	CME[31]	0.3	0.7	0.2	0.0	0.2	0.4	3.8	4.5	4.5	0.0	2.3	5.8
	本文方法 (Liner-GCN)	0.4	0.8	0.3	0.0	0.1	0.6	8.3	13.6	14.4	0.6	9.2	20.0
30	FSODFR[20]	0.2	0.5	0.2	0.0	0.1	0.4	4.3	5.4	5.4	0.0	1.8	7.7
	CME[31]	0.4	0.9	0.3	0.0	0.3	0.7	4.9	6.2	6.3	0.0	1.8	9.2
	本文方法 (Liner-GCN)	0.3	0.8	0.3	0.0	0.1	0.6	8.7	14.0	15.2	0.6	9.7	21.0

表 3 本文方法与其他方法在 PASCAL VOC 上不同 shot 下各类的平均检测精度比较结果 (单位:%)

K	方法	新类						基类															
		bird	bus	cow	motor	sofa	mean	aero	bike	boat	bottle	car	cat	chair	table	dog	horse	person	plant	sheep	train	tv	mean
1	YOLOv2-joint[20]	0.0	0.0	0.0	0.0	0.0	0.0	78.4	76.9	61.5	48.7	79.8	84.5	51.0	72.7	79.0	77.6	74.9	48.2	62.8	84.8	73.1	70.2
	YOLOv2-ft[20]	6.8	0.0	9.1	0.0	0.0	3.2	77.1	78.2	61.7	46.7	79.4	82.7	51.0	69.0	78.3	79.5	74.2	42.7	68.3	84.1	72.9	69.7
	LSTD(YOLOv2)[20]	12.0	17.8	4.6	0.0	0.1	6.9	75.5	76.9	63.2	46.2	78.9	84.1	52.5	66.8	79.2	79.4	74.1	44.7	66.4	84.6	73.6	69.7
	FSODFR[20]	13.5	10.6	31.5	13.8	4.3	14.8	75.1	70.7	57.0	41.6	76.6	81.7	46.6	72.4	73.8	76.9	68.8	43.1	63.0	78.8	69.9	66.4
	FSODFR*	14.6	1.8	23.8	28.1	0.3	13.7	76.6	74.6	54.8	43.3	63.6	76.2	48.9	55.0	74.3	76.0	70.6	39.2	59.5	79.4	71.6	64.2
	Liner-GCN	14.1	1.9	28.2	27.6	0.3	14.4	65.9	72.3	52.6	38.3	12.4	77.4	37.9	7.0	73.7	71.4	64.2	33.8	45.2	78.2	54.2	52.3
	Liner-GCN-New	11.1	2.7	20.0	25.6	0.2	11.9	70.9	66.9	54.9	38.9	48.6	79.1	42.7	33.6	73.1	74.9	70.1	33.2	50.5	79.1	57.4	58.3
2	YOLOv2-joint[20]	0.0	0.0	0.0	0.0	0.0	0.0	77.6	77.6	60.4	48.1	81.5	82.6	51.5	72.0	79.2	78.8	75.2	47.0	65.2	86.0	72.7	70.4
	YOLOv2-ft[20]	11.5	5.8	7.6	0.1	7.5	6.5	77.9	75.0	58.5	45.7	77.6	84.0	50.4	68.5	79.2	79.7	73.8	44.0	66.0	77.5	72.9	68.7
	LSTD(YOLOv2)[20]	12.3	10.1	14.6	0.1	8.9	9.2	77.4	77.1	59.4	46.4	77.8	84.5	50.9	67.1	79.1	80.6	73.8	43.3	64.9	79.4	72.4	68.9
	FSODFR[20]	21.2	12.0	16.8	17.9	9.6	15.5	74.6	74.9	56.3	38.5	75.5	68.0	43.2	69.3	66.2	42.4	68.1	41.8	59.4	76.4	70.3	61.7
	FSODFR*	18.1	0.5	24.8	28.0	7.0	16.6	73.4	75.8	49.5	39.5	75.7	78.2	46.0	58.3	73.2	61.8	67.5	42.1	56.4	66.7	68.4	62.2
	Liner-GCN	17.5	5.6	25.8	33.5	6.1	17.7	69.6	71.5	42.2	35.8	74.4	77.7	41.0	56.3	68.9	70.7	65.5	35.2	43.9	71.5	63.9	59.2
	Liner-GCN-New	18.5	6.7	31.3	29.7	13.6	18.0	75.8	71.6	42.9	35.2	75.8	79.8	39.9	64.8	73.6	70.2	66.2	34.4	51.8	67.0	59.4	60.5
3	YOLOv2-joint[20]	0.0	0.0	0.0	0.0	9.1	1.8	78.0	77.2	61.2	45.6	81.6	83.7	51.7	73.4	80.7	79.6	75.0	45.5	65.6	83.1	72.7	70.3
	YOLOv2-ft[20]	10.9	5.5	15.3	0.2	0.1	6.4	76.7	77.0	60.4	46.9	78.8	84.9	51.0	68.3	79.6	78.7	73.1	44.5	67.6	73.6	72.4	69.6
	LSTD(YOLOv2)[20]	12.3	7.1	17.7	0.1	0.0	7.5	75.9	76.2	59.7	46.6	78.3	84.4	49.4	64.5	78.7	79.7	72.6	42.5	63.8	80.5	73.9	68.4
	FSODFR[20]	26.1	19.1	40.7	20.4	27.1	26.7	73.6	73.1	56.7	41.6	76.1	78.7	42.6	66.8	72.0	77.7	68.5	42.0	57.1	74.7	70.7	64.8
	FSODFR*	22.9	8.6	24.1	22.8	30.6	21.8	70.6	69.7	53.0	40.0	72.5	81.0	44.5	56.5	73.0	75.3	67.2	30.7	52.3	64.5	69.7	62.0
	Liner-GCN	25.1	8.8	22.6	32.5	22.7	22.3	66.0	69.7	54.4	38.0	75.0	80.6	42.2	43.2	70.7	73.7	65.3	34.1	41.2	69.8	60.4	59.0
	Liner-GCN-New	24.2	14.7	26.0	18.5	15.6	21.8	68.6	71.2	53.7	39.1	77.0	80.5	42.6	63.4	70.8	73.7	67.3	34.4	34.9	65.1	64.0	60.4
5	YOLOv2-joint[20]	0.0	0.0	0.0	0.0	9.1	1.8	77.8	76.4	65.7	45.9	79.5	82.3	50.4	72.5	79.1	79.0	75.5	47.9	67.2	83.0	72.5	70.3
	YOLOv2-ft[20]	11.6	7.1	10.7	2.1	6.0	7.5	76.5	76.4	61.0	45.5	78.7	84.5	49.2	68.7	78.3	78.1	73.7	45.4	66.8	85.3	70.0	69.2
	LSTD(YOLOv2)[20]	12.9	8.1	13.6	16.1	10.2	12.2	77.4	75.0	61.1	45.2	78.4	85.0	50.6	68.0	78.1	79.3	73.1	44.6	65.5	84.5	71.1	69.1
	FSODFR[20]	31.5	21.1	39.8	40.0	37.0	33.9	69.3	57.5	56.8	37.8	74.8	82.8	41.2	67.3	74.0	77.4	70.9	40.9	57.3	73.5	69.3	63.4
	FSODFR*	21.8	20.8	32.6	38.2	36.8	30.0	67.7	47.0	57.2	36.4	72.7	77.7	38.1	56.6	67.6	75.4	66.6	33.3	56.7	67.9	64.2	59.0
	Liner-GCN	29.0	19.5	36.2	57.3	48.9	38.2	67.0	64.9	50.5	34.4	73.1	80.0	33.4	51.4	69.7	74.0	66.4	32.7	44.0	67.9	64.3	58.2
	Liner-GCN-New	28.6	14.9	41.5	55.4	49.1	37.9	67.5	64.1	48.1	35.4	73.5	78.2	36.1	61.4	70.7	75.5	66.4	34.1	48.0	67.7	54.4	58.7
10	YOLOv2-joint[20]	0.0	0.0	0.0	0.0	9.1	1.8	76.9	77.1	62.2	47.3	79.4	85.1	51.3	70.1	78.6	78.0	75.2	47.4	63.9	85.0	72.3	70.0
	YOLOv2-ft[20]	11.4	28.4	8.9	4.8	7.8	12.2	77.4	76.9	60.9	44.8	78.3	83.2	48.5	68.9	78.5	78.9	72.6	44.8	67.3	82.7	69.3	68.9
	LSTD(YOLOv2)[20]	11.3	32.3	5.6	1.3	7.7	11.6	77.1	75.2	62.0	44.5	78.2	84.2	49.9	68.6	78.8	78.8	72.6	45.0	66.9	82.6	69.5	68.9
	FSODFR[20]	30.0	62.7	43.2	60.6	40.6	47.2	65.3	73.5	54.7	39.5	75.7	81.1	35.3	62.5	72.8	78.8	68.6	41.5	59.2	76.2	69.2	63.6
	FSODFR*	21.1	60.2	36.0	54.2	42.5	42.8	67.0	68.2	51.4	32.0	73.9	77.3	33.5	52.8	65.7	75.3	65.5	35.4	55.5	72.4	66.2	59.5
	Liner-GCN	32.1	59.3	47.2	56.8	42.3	47.5	66.6	67.7	48.2	35.4	74.3	80.5	36.8	48.2	70.4	77.0	66.3	30.4	50.2	72.5	63.0	59.2
	Liner-GCN-New	34.3	59.9	51.5	58.1	43.1	49.4	68.3	67.3	47.4	33.9	74.1	79.1	34.0	57.4	70.8	75.0	65.4	32.8	50.0	73.5	63.8	59.5

度.具体的,1-shot 时“cow”类别的平均检测精度提升了 4.4%,2-shot 时“bus”类别和“motor”类别的平均检测精度分别提升了 5.1%和 5.5%,3-shot 时“motor”类别的平均检测精度提升了 9.7%.在 5/10-shot 下,本文方法能够有效提升多个类别的检测精度,表现更好.但是,与 FSODFR 方法[20]相比,本文方法 Liner-GCN 对于新类的检测性能在所有 shot 下都是部分类有显著提升,而在其他类性能有所下降.Liner-GCN-New 方法表示在小样本训练阶段,基类和新类分别取元学习器 M 输出的初始的元特征和标签语义嵌入模块 U 输出的再学习后的元特征调制查询图像.与 Liner-GCN 方法相比,Liner-GCN-New 方法能够更好的保留基类元特征的可分性.同时,Liner-GCN-New 方法在检测新类目标时也能够获得

较好的性能. 这表明 Liner-GCN-New 方法可以限制基类元特征的更新，保留基类元特征的可分性.

4.3.2　消融实验

本文对标签语义嵌入模块采用不同的结构进行了消融实验研究，实验结果如表 4 所示. Liner-GCN 方法表示标签语义嵌入模块的激活函数 $\sigma(\cdot)$ 选择线性激活函数. LeakyReLU-GCN 方法表示标签语义嵌入模块的激活函数 $\sigma(\cdot)$ 选择非线性激活函数. 从实验结果可以看出，支持样本数更少(1/2/3/5-shot)时，选择线性激活函数较好，支持样本数较多(10-shot)时，选择非线性激活函数更好. Liner-GCN-New 方法和 LeakyReLU-GCN-New 方法表示分别选择对应的激活函数，在小样本训练阶段，基类和新类分别取元学习器 M 输出的初始的元特征和标签语义嵌入模块 U 输出的再学习后的元特征调制查询图像. 我们将其称之为仅再学习新类的元特征. 与 Liner-GCN 方法和 LeakyReLU-GCN 方法的实验结果对比表明，在两种激活函数设置下，当支持样本数很少(1-shot)时，基类和新类均取标签语义嵌入模块后的元特征，新类目标检测效果更好. 当支持样本数较多时，仅再学习新类的元特征能够得到更好的新类目标检测效果. 这表明对新类元特征的再学习始终是必要的，能够提升对新类目标检测的准确度. 本文还尝试使用不同的初始化方式初始化标签语义嵌入模块中 GCN 的参数，Liner-GCN-Xavier 方法表示使用 Xavier 初始化方法初始化 GCN 网络的参数，与默认使用全 0 初始化的 Liner-GCN 方法相比，使用 Xavier 初始化方法在 1-shot 时性能明显下降，表明在样本数很少时，让模型自适应的学习不同类元特征之间融合多少知识会更好. 此外，我们在表 4 中也给出了基于 CME 方法[31]开源的基模型实现的不同标签语义嵌入模块结构下本文方法的检测精度. 我们通过在表示对应标签语义嵌入模块结构的本文方法后添加“-CME”对它们进行表示. 从实验结果可以看出，基于 CME 方法[31]开源的更好的基模型实现的本文方法通常比基于我们在受限设备上训练的较弱的基模型实现的本文方法取得更好的检测性能. 尤其是本文方法 Liner-GCN-CME 获得了与 FSODFR 方法[20]相比具有竞争性的检测精度.

表 4　标签语义嵌入模块不同设置在 PASCAL VOC 数据集上不同 shot 下对新类的平均检测精度的影响　(单位：%)

方法	1-shot	2-shot	3-shot	5-shot	10-shot
FSODFR[20]	14.8	15.5	26.7	33.9	47.2
FSODFR*	13.7	16.6	21.8	30.0	42.8
FSODFR-CME	11.4	16.9	22.9	36.3	44.7
本文方法(Liner-GCN)	14.4	17.7	22.3	38.2	47.5
本文方法(Liner-GCN-CME)	13.2	21.8	25.3	41.9	48.0
本文方法(LeakyReLU-GCN)	12.9	17.4	22.3	37.2	48.7
本文方法(LeakyReLU-GCN-CME)	12.0	17.0	26.9	40.6	47.5
本文方法(Liner-GCN-New)	11.9	18.0	21.8	37.9	49.4
本文方法(Liner-GCN-New-CME)	11.8	21.4	24.0	40.7	46.3
本文方法(LeakyReLU-GCN-New)	11.5	18.4	23.8	38.5	48.7
本文方法(LeakyReLU-GCN-New-CME)	10.3	17.6	24.0	40.0	47.0
本文方法(Liner-GCN-Xavier)	10.7	18.0	21.5	38.6	48.4
本文方法(Liner-GCN-Xavier-CME)	12.1	19.7	25.4	41.3	48.2

4.3.3　可视化

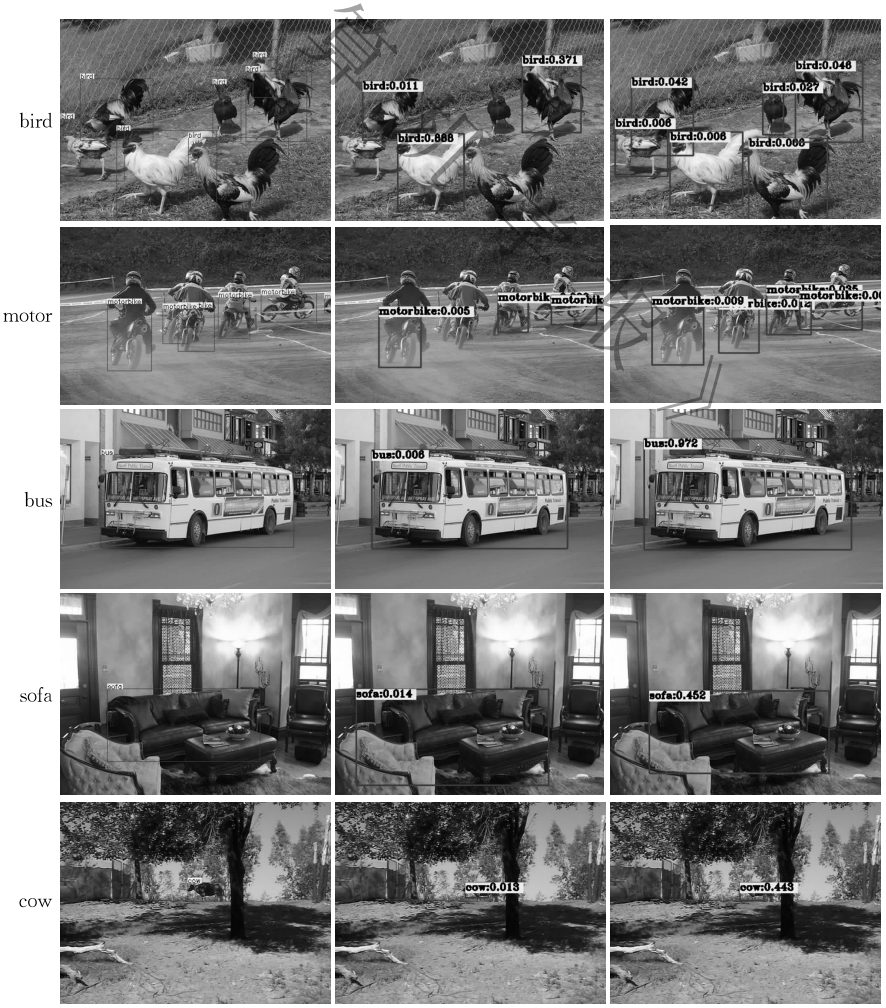
为了进一步分析本文方法的有效性，我们展示了 PASCAL VOC 数据集上新类与其它类别的标签语义相似度值，如表 5 所示. 为了表达清晰，我们将类别按照共有属性和不同用途分为四大类：交通工具类、动物类、人和家庭用品类. 从表 5 可以看出，相同大类中的基类和新类间的语义相似度通常较高，这些基类除了在语义上与新类相似，在视觉特征上

表 5　PASCAL VOC 数据集上新类与其它类别的标签语义相似度

类别	sofa	cow	bird	bus	motor
人					
person	0.39	0.46	0.51	0.47	0.37
家庭用品					
sofa	1.00	0.38	0.33	0.38	0.30
chair	0.71	0.40	0.44	0.42	0.28
bottle	0.39	0.48	0.50	0.39	0.30
tv monitor	0.17	0.14	0.16	0.21	0.14
dining table	0.26	0.10	0.10	0.12	0.13
potted plant	0.26	0.24	0.35	0.17	0.2
动物					
cat	0.41	0.60	0.67	0.45	0.38
dog	0.41	0.63	0.61	0.44	0.38
cow	0.38	1.00	0.56	0.41	0.34
bird	0.33	0.56	1.00	0.41	0.32
horse	0.38	0.68	0.56	0.45	0.47
sheep	0.32	0.77	0.53	0.37	0.35
交通工具					
car	0.44	0.42	0.45	0.65	0.66
bus	0.38	0.41	0.41	1.00	0.48
boat	0.40	0.38	0.51	0.58	0.53
train	0.36	0.45	0.47	0.73	0.42
motor	0.30	0.34	0.32	0.48	1.00
bicycle	0.37	0.41	0.43	0.57	0.67
aeroplane	0.26	0.31	0.40	0.41	0.43

也与新类更相似.例如交通工具大类中的新类“bus”与基类中“train”和“car”都有轮胎、玻璃、方形外观等特点.因此,可以充分利用基类“train”和“car”的元特征来增强新类“bus”的相似性属性的元特征.新类与其所属的大类中不同的基类间的语义相似度有所不同.例如家庭用品大类中新类“sofa”与基类“chair”的语义相似度很高,但新类“sofa”与基类“dining table”、“tv monitor”、“potted plant”虽都出现在家庭用品大类中,但它们之间的相似度很低,因为它们并不严格属于家庭用品,而且它们的视觉表观相似度也不高.因此,利用不同支持类别间标签语义相似度建模的不同类间的语义关联不仅在语义上有效,也能够辅助补充不同类别间视觉表观的联系.基于该语义关联利用标签语义嵌入模块对元特征进行再学习是可行并有效的.

图4是本文方法与FSODFR* 方法在PASCAL VOC数据集上10-shot下对每个新类的FSOD可视化结果.图4中每行表示一个新类类别,且不同的行展示了具有不同特点的场景.图4中第一列为标签的可视化结果,第二列为对比方法FSODFR* 的可视化结果,最后一列为本文方法的可视化结果.从图4中第一行和第二行可以看出,本文方法在密集场景中仍然能够检测到全部的目标.即使是在图像简单的场景中,如图4第三行所示,本文方法相比FSODFR* 方法虽然定位不精确,但本文方法的分类置信度得分(置信度得分97.2%)显著高于FSODFR* 方法(置信度得分0.6%).图4第四行和最后一行分别表示了在背景混淆和目标尺度较小的场景中,本文方法仍然能够获得较高的目标分类置信度得分.可视化结果表明,本文方法能够学习到新类更判别性的表征.图5是本文方法在PASCAL VOC数据集上10-shot下对每个新类的FSOD效果差的示例.图5中边线较细的边框表示标签,边线较宽的边框表示本文方法预测的边框.从图5中可以看出,当图像场景中目标姿态变化太大或目标被严重遮挡时,本文方法存在漏检、分类置信度不高或边界定位不准确的问题.



(a) GT (b) FSODFR* (c) 本文方法

图4 本文方法与FSODFR* 方法在PASCAL VOC数据集上10-shot下对每个新类的FSOD可视化结果

图 5　本文方法与 FSODFR* 方法在 PASCAL VOC 数据集上 10-shot 下对每个新类的 FSOD 效果差的示例

5　结　论

基于调制的 FSOD 框架，本文提出了一种嵌入标签语义的元特征再学习和重加权小样本目标检测方法. 该方法在小样本训练阶段，引入了标签语义信息构建词向量标签语义图建模基类和新类间的语义联系. 同时，本文提出了一个标签语义嵌入模块将该关联性通过 GCN 融入到支持集样本元特征的再学习过程中，将基类中与新类相关的信息传递给新类元特征，使在只有少量新类数据的情况下学习到较好的新类元特征，从而提高新类目标的检测精度. 在 PASCAL VOC 和 COCO 数据集上的实验及分析表明了与 FSODFR 方法相比，我们的方法能有效提高在只有少量样本的情况下新类目标的检测效果.

参　考　文　献

[1] Jiao Li-Cheng, Zhao Jin, Yang Shu-Yuan, et al. Deep Neural Network Learning, Optimization and Recognition. Beijing: Tsinghua University Press, 2017(in Chinese)
(焦李成，赵进，杨淑媛等. 深度神经网络学习、优化与识别. 北京：清华大学出版社，2017)

[2] Jiao Li-Cheng. Neural Network Computing. Xi'an: Xidian University Press, 1993(in Chinese)
(焦李成. 神经网络计算. 西安：西安电子科技大学出版社，1993)

[3] Jiao Li-Cheng. The Application and Realization of Neural Network. Xi'an: Xidian University Press, 1993(in Chinese)
(焦李成. 神经网络的应用与实现. 西安：西安电子科技大学出版社，1993)

[4] Jiao Licheng, Shang Ronghua, Liu Fang, et al. Brain and nature-inspired learning, computation and recognition. Amsterdam: Elsevier, 2020

[5] Wang H, Jiao L, Liu F, et al. IPGN: Interactiveness proposal graph network for human-object interaction detection. IEEE Transactions on Image Processing, 2021, 30: 6583-6593

[6] Zhang W, Jiao L, Li Y, et al. Laplacian feature pyramid network for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14

[7] Jiao L, Zhang R, Liu F, et al. New generation deep learning for video object detection: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 3195-3215

[8] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR), 2020, 53(3): 1-34

[9] Li Fan-Zhang, Liu Yang, Wu Peng-Xiang, et al. A survey on recent advances in meta-learning. Chinese Journal of Computers, 2020, 32(2): 349-369(in Chinese)
(李凡长，刘洋，吴鹏翔等. 元学习研究综述. 计算机学报，2021，44(2)：422-446)

[10] Zhao Kai-Lin, Jin Xiao-Long, Wang Yuan-Zhuo. Survey on few-shot learning. Journal of Software, 2021, 32(2): 349-369(in Chinese)
(赵凯琳，靳小龙，王元卓. 小样本学习研究综述. 软件学报，2021，32(2)：349-369)

[11] Liu Ying, Lei Yan-Bo, Fan Jiu-Lun, et al. Survey on image classification technology based on small sample learning. Acta Automatica Sinica, 2021, 47(2): 297-315(in Chinese)
(刘颖，雷研博，范九伦等. 基于小样本学习的图像分类技术综述. 自动化学报，2021，47(2)：297-315)

[12] Chen H, Wang Y, Wang G, et al. LSTD: A low-shot transfer detector for object detection//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2836-2843

[13] Wu J, Liu S, Huang D, et al. Multi-scale positive sample refinement for few-shot object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 456-472

[14] Wang X, Huang T E, Darrell T, et al. Frustratingly simple few-shot object detection//Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 9919-9928

[15] Fan Z, Ma Y, Li Z, et al. Generalized few-shot object detection without forgetting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 4527-4536

[16] Kim G, Jung H G, Lee S W. Spatial reasoning for few-shot object detection. Pattern Recognition, 2021, 120: 108118

[17] Karlinsky L, Shtok J, Harary S, et al. RepMet: Representative-based metric learning for classification and few-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5197-5206

[18] Yang Y, Wei F, Shi M, et al. Restoring negative information in few-shot object detection//Advances in Neural Information Processing Systems. Virtual, 2020: 3521-3532

[19] Zhu C, Chen F, Ahmed U, et al. Semantic relation reasoning for shot-stable few-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 8782-8791

[20] Kang B, Liu Z, Wang X, et al. Few-shot object detection via feature reweighting//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea(South), 2019: 8420-8429

[21] Hsieh T I, Lo Y C, Chen H T, et al. One-shot object detection with co-attention and co-excitation. arXiv preprint arXiv: 1911.12529, 2019

[22] Wu X, Sahoo D, Hoi S. Meta-RCNN: Meta learning for few-shot object detection//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 1679-1687

[23] Fan Q, Zhuo W, Tang C K, et al. Few-shot object detection with attention-RPN and multi-relation detector//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4013-4022

[24] Fu K, Zhang T, Zhang Y, et al. OSCD: A one-shot conditional object detection framework. Neurocomputing, 2021, 425: 243-55

[25] Yang H, Lin Y, Zhang H, et al. Towards improving classification power for one-shot object detection. Neurocomputing, 2021, 455: 390-400

[26] Hu H, Bai S, Li A, et al. Dense relation distillation with context-aware aggregation for few-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 10185-10194

[27] Li A, Li Z. Transformation invariant few-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 3094-3102

[28] Sun B, Li B, Cai S, et al. FSCE: Few-shot object detection via contrastive proposal encoding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 7352-7362

[29] Zhang L, Zhou S, Guan J, et al. Accurate few-shot object detection with support-query mutual guidance and hybrid loss//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 14424-14432

[30] Chen D J, Hsieh H Y, Liu T L. Adaptive image transformer for one-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 12247-12256

[31] Li B, Yang B, Liu C, et al. Beyond max-margin: Class margin equilibrium for few-shot object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 7363-7372

[32] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580-587

[33] Girshick R. Fast R-CNN//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1440-1448

[34] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks//Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[35] He K, Gkioxari G, Dollár P, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969

[36] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[37] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21-37

[38] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7263-7271

[39] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018

[40] Bochkovskiy A, Wang C Y, Liao H Y. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934, 2020

[41] Tang Y, Wang J, Wang X, et al. Visual and semantic knowledge transfer for large scale semi-supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 3045-3058

[42] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the Workshop at the 1st International Conference on Learning Representations. Scottsdale, USA, 2013: 1-12

[43] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[44] Everingham M, Van Gool L, Williams C K, et al. The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338

[45] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: Common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[46] Tian P, Wu Z, Qi L, et al. Differentiable meta-learning model for few-shot semantic segmentation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12087-12094

LI Peng-Fang, Ph. D. candidate. Her main research areas are image processing and machine learning.

LIU Fang, M. S., second-level professor, Ph. D. supervisor. Her current research interests include artificial intelligence and pattern recognition, machine learning, image perception and scene interpretation, evolutionary computation and data mining.

LI Ling-Ling, Ph. D., associate professor. Her main research fields are quantum evolutionary optimization learning, deep learning methods and applications, and complex remote sensing image understanding and interpretation.

LIU Xu, Ph. D., lecturer. His main research areas are machine learning/deep learning theory, image/video processing methods.

FENG Zhi-Xi, Ph. D., associate professor. His main research areas are intelligent target information perception and machine learning.

JIAO Li-Cheng, Ph. D., professor, Ph. D. supervisor. His current research interests include signal and image processing, natural computation, and intelligent information processing.

XIONG Yi-Meng, M. S. candidate. Her main research fields are image processing and deep learning.

Background

Compared with Object Detection, the purpose of Few-Shot Object Detection is to hope that the deep model will use only a small number of new samples to quickly adapt to the identification and positioning of new objects after learning base data with sufficient samples. It can reduce the dependence of the depth model on a large number of labeled samples.

With the extensive research on Few-Shot Learning and combining it with Object Detection models, many Few-Shot Object Detection methods have been proposed. However, because there are few samples available for learning in the new class, the features of the new class learned by the model are not strongly separable, which limits the detection accuracy for the new class of objects.

In response to the above problem, existing works introduced attention mechanism or context information into Few-Shot Object Detection model to enhance the separability of new class features. Our work imitates the ability of human association learning. By introducing label semantics, we design a word vector label semantic graph generation module, and propose a label semantic embedding module to model the semantic association of new classes and base classes. The features related to the new class in the base class are transferred to the new class, and the separability of the new class features is enhanced by End-to-End learning. The experimental results are given also prove the feasibility and effectiveness of our method.

This work was supported in part by the National Natural Science Foundation of China (No. 62076192), the Key Research and Development Program in Shaanxi Province of China (No. 2019ZDLGY03-06), the State Key Program of National Natural Science of China (No. 61836009), in part by the Program for Cheung Kong Scholars and Innovative Research Team in University (No. IRT_15R53), in part by the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No. B07048), in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Key Research and Development Program of China, and the CAAI Huawei MindSpore Open Fund. This work is part of Few-Shot Learning method research in these projects, and other related work has been published in international journals and conferences.