文章编号：2096-1618(2022)05-0527-06

基于注意力类特定编码的小样本目标检测

林弟忠， 邹书蓉， 符 颖
(成都信息工程大学计算机学院，四川 成都 610225)

摘要：基于 CNN 的小样本目标检测网络在两阶段元训练注入少量新类图像时，不混合基类进行训练已成为一种趋势，这样能高效地向模型注入新类。而在这种增量式训练方式下，由于输入的新类别样本量少，模型由于泛化性能不足，易错检新注入的类别数据为模型训练过的种类。基于此，在 CenterNet 框架上设计了一种新的小样本目标检测器，能快速高效地进行检测。检测器引入了一个重要组件：对图像做有效增强处理后提取类表征信息的注意力类编码器，能有效地提升网络对新类的编码性能，从而增强模型对新类的泛化能力。实验结果表明，方法在一些场景下优于近期比较流行的小样本目标检测框架。

关 键 词：小样本检测；类特定编码；注意力机制；基类；新类
中图分类号：TP756 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.05.007

0 引言

现有的高性能目标检测模型进行训练需要大规模带标签信息的数据集，而大多数实际检测场景中标注数据只占很小一部分。没有足够的标注数据用于训练，导致现有模型检测效果不佳，这阻碍了目标检测的研究及应用于更多的场景中。近几年，小规模标注数据集下的检测任务引起了重视，从而小样本目标检测相关工作得到了迅速发展。当前大多数的研究模型基于传统的 Faster-RCNN[1]、YOLO[2]、SSD[3]等有锚的目标检测框架搭建，并借鉴了小样本学习的元训练策略[4]。但这些基于有锚的小样本目标检测框架依赖于大量的基类数据进行长时间训练，并且为了适应新类样本的检测需要构建基类和新类的平衡小样本集[5]并微调参数。如果还需要额外引入新类，就必须进行二次训练，训练方法十分复杂。为克服这种烦琐的训练方法，让检测器高效地检测新类样本，基于无锚的小样本检测器得到了发展。如 ONCE[6]小样本目标检测器，参考了基于无锚的 CenterNet[7]检测网络并额外引入基于 ResNet 的类编码器对新类进行编码。该网络可以直接注入新类进行检测无需微调及进行二次训练，其中类代码能构建起每个类独有的权重参数用于检测网络进行有效检测。而 ONCE 仅用了单一的 ResNet[8]作为类代码生成器提取类特征，编码性能不佳导致对于困难样本会产生大量错检和漏检。

收稿日期：2022-01-10
基金项目：四川省科技厅重大科技专项资助项目(2019ZDZX0005)

以上研究表明，提升小样本目标检测器对新类检测的泛化性能是十分关键的。近年来，将注意力机制[9-10]融合到特征提取网络有益于让网络专注于学习每个目标最重要的特征信息，提升网络的编码性能。

借鉴注意力机制的思想并参考无锚的 CenterNet 网络框架，本文设计了全新的小样本目标检测器。引入融合注意力模块的类编码器高效地提取新类图像的表征信息，让类编码器专注于学习每个类独有的特征信息，提取类特定的代码用于检测网络，有益于检测新类中的困难样本。实验结果表明，本文的小样本目标检测模型使新类样本的泛化性能得到了有效增强。

1 相关工作

2018 年，Kang 等[11]参考 YOLOv2 框架搭建小样本检测模型并额外引入了权重调整模块生成每个类别的特有的权重向量用于适应对新类别的检测，从而实现在基类和新类混合训练场景下对新类的有效检测。2019 年，Zhang 等[12]提出对比网络，第一阶段利用常规的 Faster-RCNN 进行训练，第二阶段先采用一个孪生网络提取查询图像和目标图像特征并做相似度计算，而注入新类样本到模型中需要重新进行微调和训练，才能进行有效检测。2019 年，Fan 等[13]在候选框区域提取网络中引入 attention-RPN 模块用于融合查询图像与支持集图像的特征，采用双向对比训练策略用于检测新类。多次关联和对比的训练方式充分对目标的相似性特征进行分析，但模型训练方式非常复杂。

2020 年,Wang 等[5]提出了以 Faster-RCNN 为框架,分两阶段训练,在第二阶段只微调分类和回归子网络,通过重新调整特征的组合权重以适应新类。这种框架也需要在第二阶段进行微调,因此不能轻易地将新类注入模型中。Juan-Manuel 等[6]借鉴 CenterNet[7]框架提出了 ONCE 网络,引入提取图像特征的元网络和用于定位目标的目标定位网络。另外,采用残差网络作为类代码生成器提取与类相关的类代码用于检测网络。这样的网络框架能轻易地引入新类,但单一的残差网络对新类的特征提取性能不足,导致模型的泛化性能不佳。

注意力机制能让网络显著地关注图像比较重要的部分,有效减少周围信息的干扰。近年来,注意力机制[14-16]有效地引入到图像领域 CNN 网络中,取得了不错的效果。Wang 等[17]提出了由多层注意力模块堆叠而成的残差注意力网络,可以进一步提取特征图中的重要信息,对噪声输入也具有一定的鲁棒性,但网络参数量太大,增加了模型训练负担。Hu 等[9]提出由若干个通道注意力模块组合而成的 SeNet(squeeze-and-excitation networks),该注意力模块旨在通过对特征进行全局平均池化学习通道之间的相关性,让网络能够使用动态通道级特征重新进行校准从而提高网络的特征表达能力。Sanghyun Woo 等[10]借鉴了 Shen 等[9]的工作并验证了在注意力模块中只使用平均池化关注单一的通道特征关系并非最优方案,提出对输入注意力模块的特征额外引入最大池化计算。设计出了融合空间特征和通道特征的注意力模块 CBAM(convolutional block attention module),提升了网络对图像重要特征的表达性能。实验证明两种注意力模块的融合优于仅仅只关注通道特征的方式。

2 检测方法

2.1 CenterNet 网络

针对增量式元训练场景,本文选择无锚目标检测算法 CenterNet 网络作为框架,对比 YOLO、SSD 等网络,CenterNet 网络结构见图 1。该网络针对每个类独立进行检测,减少了基类和新类的特征交叉,从而有效减小基类特征对新类的干扰。网络独立学习新类的能力更强,适合用于构建小样本目标检测的框架。首先,利用裁剪、颜色抖动、随机缩放以及随机翻转等操作作为数据增强 T_1 得到增强后的数据。然后,CenterNet 将训练图像传入特征提取器中得到网络的热图,最终根据热图预测目标的中心点并回归得到最终的检测框。

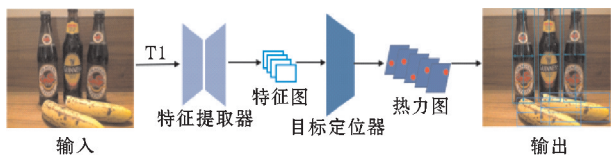

图 1 CenterNet 网络结构

2.2 小样本目标检测网络架构

本文设计了一种新的小样本目标检测模型,其网络结构见图 2。将 CenterNet 网络分解为特征提取器和目标定位器,其中特征提取器为编码解码结构,编码部分为残差网络,解码部分为反卷积网络。并且所有新类和基类共享权重。而目标定位器包含类相关的权重信息,能对特征图进一步做卷积操作生成热图。

另外,引入了注意力类编码器生成具有注意力感知的类代码用于目标定位器,这取代了 CenterNet 利用迭代更新类相关权重参数的操作。注意力类编码器通过全局平均池化输出与类相关的权重信息进一步参数化目标定位器的参数。在该类编码器中融合了空间注意力模块和通道注意力模块,使编码器专注于学习类特定代码。

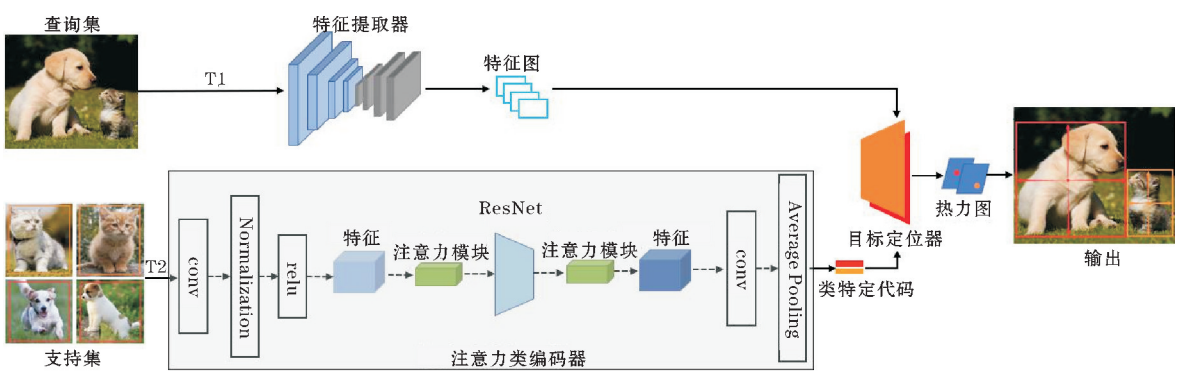

图 2 本文所提网络结构

2.3 注意力类编码器

受注意力机制[15-17]相关工作的启发,在注意力类编码器的结构中,为保持最佳特征编码性能,不改变整个残差网络的结构,在整个残差网络的前面和后面分

别融合注意力模块,见图 2。与 CBAM[17]保持一致,该注意力模块的结构见图 3。图像经过卷积及归一化等操作得到的浅层特征,先通过一个通道注意力模块,得到加权特征之后,再经过一个空间注意力模块,从而得到同时具有通道和空间注意力感知的特征信息。并利用残差网络进一步提取图像深层的特征信息,最终以平均池化的方式输出类特定代码。

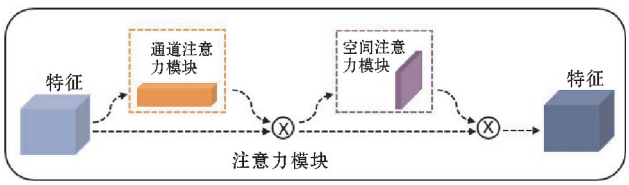

图 3　注意力模块

2.4　元训练策略

借鉴元学习的训练策略[4],为充分利用基础类别,将元训练分为两个串行阶段。第一阶段,利用标注信息丰富的基类数据在标准的 CenterNet 网络上训练出特征提取器的权重参数用于下一个训练阶段。第二阶段的训练分为多个 episode,见图 4。每个 episode 执行多个元任务,每个元任务根据标签信息随机抽取多个样本构成一个支持集用于训练和一个查询集用于测试。这种学习机制有益于网络在不同元任务中学习提取每个类别最重要的特征[4],增强模型的特征提取性能,同时也更有利于学习每个类独有的类代码。

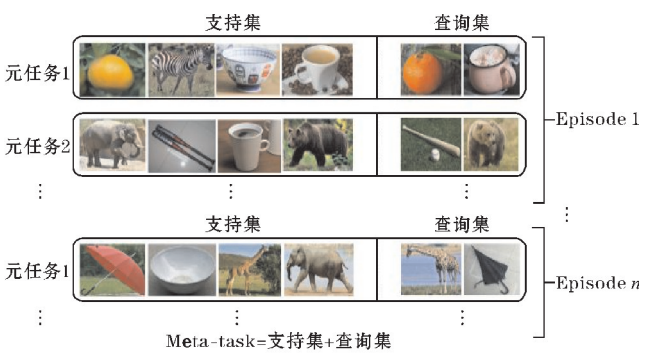

图 4　训练策略

2.4.1　第二阶段训练:注意力类编码器的学习

元训练第一阶段采用标准的 CenterNet 网络进行训练,只是为了学习特征提取器的权重用于下一阶段。在第二阶段固定特征提取器的参数,主要训练一个融合通道注意力模块和空间注意力模块的注意力类编码器,从而具有生成注意力感知的类代码的能力。为了对注意力类编码器进行有效训练,本文采用增量式元训练策略[6]。

具体做法:整个训练由多个 episode 构成,每个 episode 执行一定数量的元任务,并且每个元任务从所有类别中采样一个类标签集 L。比如,$L=\{$香蕉,伞,$\cdots\}$。每个元任务会根据标签集 L 随机抽取一个支持集 S 和一个查询集 Q。每个支持集中的图像 $x(x\in S)$ 做数据增强 T_2 输入类编码器提取特征。类编码器中的残差网络部分用上一阶段训练得到的特征提取器中编码器部分的权重进行初始化。在前向传播过程中,每一个元任务中的查询集图像做数据增强 T_1,并利用一阶段训练好的特征提取器提取查询集特征,见式(1),得到多通道的特征图。同时,注意力类编码器提取支持集图像特征生成类特定代码 c_{lk}:

$$m_Q=f(I),I\in Q \tag{1}$$

$$c_{lk}=g(S_{lk}) \tag{2}$$

其中,m_Q 为查询集图像 I 的特征,S_{lk} 为采样得到的支持集样本。最终,通过全局平均池化输出 768 维的类特征 c_{lk},并以同种类别的类特征 c_{lk} 做平均池化,得到每个类的类特定代码 $\{c_k\}$。将查询集图像特征 m_Q 和类特定代码 $\{c_k\}$ 输入目标定位器中做卷积运算 h 生成每个类对应的热图 Y:

$$Y_Q=h(m_Q,c_k) \tag{3}$$

热图的损失采用 L_1 损失函数进行计算,如式(4)所示。通过更新目标定位器和注意力类编码器的参数使得热图预测偏差最小。

$$L_{Q_{\text{heatmap}}}=\frac{1}{n}\sum_{i=1}^{n}|Y_Q-Z| \tag{4}$$

其中,n 为查询集图像的总数,Z 为真实热图。

该训练阶段总损失 $L_{\text{meta\_det}}$ 由查询集热图 heatmap、回归框尺寸预测 size、中心点偏移量 offset 3 部分组成:

$$L_{\text{meta\_det}}=L_{Q_{\text{heatmap}}}+\lambda_{\text{size}}L_{\text{size}}+\lambda_{\text{off}}L_{\text{off}} \tag{5}$$

2.4.2　元测试:注入新类

经过元训练得到健壮的小样本目标检测器,其中包含注意力特征提取器、注意力类编码以及目标定位器。采用元测试向检测器注入新类进行测试。首先,根据标签集对模型输入随机抽取得到的一组新类支持集样本,注意力类编码器提取新类的类特定代码。同时,特征提取器对输入的测试图像进行特征提取。然后,网络将得到的类特定代码和特征图输入目标定位器中做卷积运算生成热图,并通过回归得到测试样本的检测结果。

3　实验

3.1　实验环境及数据集

实验部署在 8 张英伟达 GT 710 12G 显卡的 Linux

服务器上,配置了符合要求的加速平台和加速库。采用目标检测常用的 COCO[18] 基准数据集进行实验,其中训练集 118287 张,验证集 5000 张,涵盖 80 个目标类别,其中 20 个类别作为新类。该 20 个类别与 PASCAL VOC[19] 数据集所涵盖的类别相同,COCO 数据集中剩余的 60 个类别作为基类。因此,实验可分为两种:第一种为元训练和元测试均在 COCO 上进行的 COCO 同数据集评估;第二种是将 COCO 数据集用于元训练两阶段,在 PASCAL VOC 数据集上进行元测试的跨数据集评估。

3.2 COCO 同数据集评估

首先,调整 COCO 基类训练图像的尺寸到 512×512,以标准的 CenterNet 训练方式进行元训练第一阶段。然后,在第二阶段中,与 ONCE[6] 保持一致,将基类视为伪新类样本进行训练。设置每一个 episode 随机抽取 32 个元任务,每个元任务包含对 3 个类别的检测,并且每个类别含有 5 个标注框,增大元任务的学习量有益于性能的提升。

采用元测试对模型性能进行评估时,网络会从 COCO 训练集中采样多组新类支持集,并且对每组支持集会随机抽取每种新类{shot = 1,5,10}个数据样本用于提取类特定代码。同时,使用 COCO 验证集上的新类样本作为测试图像评估本文小样本目标检测器的性能。

本文的模型与主流的小样本目标检测算法进行了性能对比:标准 Fine-Tuning 检测模型[6];Few-shot object detection via feature reweighting[11];增量式小样本目标检测网络 ONCE[6]。实验结果如表 1 所示。从实验结果可知,对每种新类采样{shot = 1,5,10}个样本用于提取类代码进行检测,本文的方法均最优。证明设计的检测器泛化性能得到了有效增强,能根据少量的新类样本实现有效检测。同时,在{shot = 10}情况下,ONCE 和本文方法检测对比结果见图 5。可以看出,本文的方法有效地减少了对检测目标的错检和漏检情况。

表 1 COCO 同数据集检测对比结果

Shot	方法	新类		基类		新类+基类	
		AP	AR	AP	AR	AP	AR
1	Fine-Tuning	0.0	0.0	1.1	1.8	0.8	1.4
	Feature-Reweight	0.1	0.3	2.5	4.3	1.9	3.3
	ONCE	0.7	6.3	17.9	19.5	13.6	16.2
	本文	0.7	6.4	18.0	19.6	14.0	16.4
5	Fine-Tuning	0.2	3.5	2.6	7.4	2.0	6.4
	Feature-Reweight	0.8	5.1	3.3	8.2	2.6	7.4
	ONCE	1.0	7.4	17.9	19.5	13.7	16.4
	本文	1.2	7.6	18.2	20.1	14.2	16.6
10	Fine-Tuning	0.6	4.2	2.8	8.0	2.3	7.0
	Feature-Reweight	1.5	8.3	3.7	8.9	3.1	8.7
	ONCE	1.2	7.6	17.9	19.5	13.7	16.5
	本文	1.6	8.1	18.5	21.4	14.2	17.9

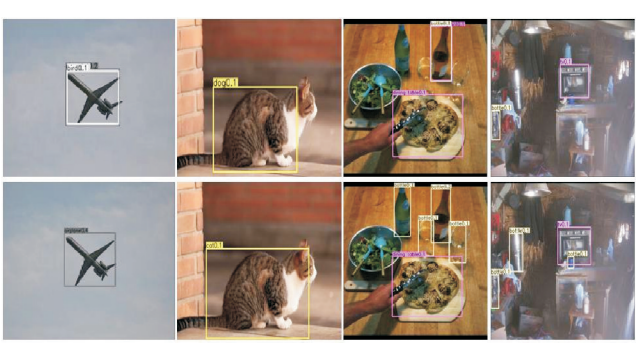

图 5 COCO 数据集上 ONCE(上)和本文的方法检测效果对比(下)

3.3 PASCAL VOC 跨数据集评估

对于从 COCO 到 PASCAL VOC 的跨数据集评估,同样采用 COCO 数据集的基类数据进行两个阶段的元训练,而元测试采用 PASCAL VOC 测试集作为测试图像评估本文的小样本目标检测器的性能,实验对比结果如表 2 所示。从实验结果可知,对新类采样{shot = 5,10}个标注样本提取类特定代码进行元训练,并对 PASCAL VOC 测试集进行元测试,在得到的测试结果中,AP 和 AR 值均优于其他主流算法。说明本文的检测器可以有效地迁移到新数据集上进行检测,这对于实际检测场景具有重要意义。

表 2　PASCAL VOC 跨数据集检测对比结果

Shot	方法	AP	AP_S	AP_M	AP_L	AR	AP_S	AP_M	AP_L
5	Fine-Tuning	0.1	0.1	0.8	0.3	1.9	0.7	2.9	7.6
	ONCE	2.4	1.2	2.4	3.4	12.2	5.9	16.4	33.6
	本文	2.6	1.3	2.4	3.8	12.8	5.9	16.3	34.6
10	Fine-Tuning	0.3	0.1	0.8	1.0	2.8	0.9	3.3	10.2
	ONCE	2.6	5.7	2.2	4.9	11.6	8.3	19.4	32.6
	本文	2.9	5.8	2.2	4.6	11.8	8.5	19.9	32.5

3.4　消融实验

本文方法的消融实验结果如表 3 所示。实验结果表明,本文的检测框架采用融合了通道注意力和空间注意力的类编码器,在性能上能够达到最优。

表 3　消融实验结果

实验次数	注意力模块		新类	shot = 10
	通道注意力模块	空间注意力模块	AP	AR
第 1 次			1.0	7.2
第 2 次	√		1.3	7.5
第 3 次		√	1.4	7.6
第 4 次	√	√	1.6	8.0

4　结论和讨论

引入的注意力类编码器能对输入的少量新类样本高效编码出类特定代码用于目标定位器,从而提高目标检测的准确性,减少了错检和漏检。同时,采用的增量式元训练策略并没有在元训练中构建基类和新类的平衡小样本集,而在元测试阶段直接注入新类样本进行检测。采取这种增量式元训练策略,在实际应用场景当中更易于引入新类。而这也是极具挑战性的工作, 因为注入的新类样本很容易被模型误判为经过训练的基类。在实验部分通过目标检测主要评价指标 AP 和 AR 对本文的方法进行了评估并取得了不错的检测效果。此外,检测结果受益于更大的元任务学习量,如果有更多的 GPU 内存,方法可以在每一个元任务增大训练样本量,模型检测性能还能够得到进一步提升。

参考文献:

[1] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.

[3] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. European conference on computer vision. Springer, Cham, 2016: 21-37.

[4] Ravi S, Larochelle H. Optimization as a model for few-shot learning[J]. 2016.

[5] Wang Xin, Thomas E Huang, Trevor Darrell, et al. Frustratingly simple few-shot object detection[C]. In International Conference on Machine Learning (ICML), 2020.

[6] Juan-Manuel, Perez-Rua, Xiatian zhu, et al. Incremental few-shot object detection [C]. CVPR, 2020.

[7] Zhou X, Wang D, Krähenbühl P. Objects as points [J]. arXiv preprint arXiv, 2019.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7132-7141.

[10] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV), 2018: 3-19.

[11] Kang B, Liu Z, Wang X, et al. Few-shot object detection via feature reweighting[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8420-8429.

[12] Zhang T, Zhang Y, Sun X, et al. Comparison network for one-shot conditional object detection [J]. arXiv preprint arXiv: 2019.

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

[14] Wang X, Girshick R, Gupta A, et al. Non-local neural networks [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7794-7803.

[15] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer [J]. arXiv preprint arXiv: 2016.

[16] Chen L, Zhang H, Xiao J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 5659-5667.

[17] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 3156-3164.

[18] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context [C]. European conference on computer vision. Springer, Cham, 2014: 740-755.

[19] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge [J]. International journal of computer vision, 2010, 88(2): 303-338.

Attention-based Class Encoding for Few-Shot Object Detection

LIN Dizhong, ZOU Shurong, FU Ying

(College of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Recent research shows that when two-stage few-shot object detection networks enroll a small number of new classes of images, the second stage of meta-training without base classes can efficiently enroll new classes into the network. However, under this incremental meta-training method, due to the small amount of input novel class samples, the model is prone to incorrectly detect the newly injected category data as base classes trained by the model due to insufficient generalization performance. In this paper, we design a new few-shot object detector within the CenterNet framework, which can detect objects quickly and efficiently. Our detector introduces an important component: an attention class encoder that extracts class representation from augmented images, which plays an important role in improving the generalization performance of the detection model for new classes. Experimental results show that our proposed method is better than the recently state-of-art few-shot object detection framework in some scenarios.

Keywords: few-shot object detection; class-specific encoding; attention mechanism; base class; new class