(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118097289 B
(45)授权公告日 2024.09.20

(21)申请号 202410301193.8
(22)申请日 2024.03.15
(65)同一申请的已公布的文献号
申请公布号 CN 118097289 A
(43)申请公布日 2024.05.28
(73)专利权人 华南理工大学
地址 510640 广东省广州市天河区五山路381号
专利权人 人工智能与数字经济广东省实验室(广州)
(72)发明人 黄阳阳 罗荣华
(74)专利代理机构 广州粤高专利商标代理有限公司 44102
专利代理师 江裕强
(51)Int.Cl.
*G06V 10/764*(2022.01)
*G06N 3/0464*(2023.01)
*G06N 3/08*(2023.01)
*G06V 10/25*(2022.01)
*G06V 10/26*(2022.01)
*G06V 10/44*(2022.01)
*G06V 10/82*(2022.01)
(56)对比文件
CN 114241260 A,2022.03.25
CN 117475148 A,2024.01.30
审查员 邹盼盼

权利要求书4页 说明书11页 附图2页

(54)发明名称

一种基于视觉大模型增强的开放世界目标检测方法

(57)摘要

本发明公开了一种基于视觉大模型增强的开放世界目标检测方法。所述方法利用视觉大模型对输入图像预处理,无监督的方式获取未知对象的原始伪标签,然后利用提出的基于对象重构的韦布尔模型对未知对象进行建模,实现了开放环境下对已知和未知类别的检测,减少了人工标注的成本,提高了开放世界下目标检测精度。


CN 118097289 B

1.一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,包括以下步骤:

S1、利用视觉大模型对输入的图像进行预处理,获取未知对象的原始伪标签;

S2、利用基于对象重构的韦布尔模型,对前景和背景区域建模,并且为伪未知对象计算软标签,估计未知对象的可能性得分,判断未知对象是否是真实未知对象;根据不同图像中前景和背景出现的特征频率不同,从特征频率的角度考量,背景区域和前景区域分别形成两个不同的分布,基于对象重构的韦布尔模型ORW中,基于数据重构的自动编码器学习频率信息,并通过先验概率分布来建模重构误差的两个不同分布,采用韦布尔分布用作基于对象重构的韦布尔模型ORW中的先验模型;基于对象重构的韦布尔模型ORW对前景和背景区域建模,并且为伪未知对象计算软标签,估计未知对象的可能性得分;

对前景和背景区域建模,基于对象重构的韦布尔模型ORW首先使用SAM分割图像中所有物体,根据掩膜生成所有对象的候选框;然后使用骨干网络对输入图像I提取特征图F,$\mathrm{I}\in\mathrm{R}^{H_I\times W_I\times 3}$, $\mathrm{F}\in\mathrm{R}^{H_F\times W_F\times C}$,为了充分表示对象语义信息,将SAM生成的所有对象的候选框映射到特征图中,每个对象的候选框的特征向量表示感受野内的区域;

基于对象重构的韦布尔模型ORW中,利用自动编码器来重构这些区域特征,即每个对象的候选框的特征向量;自动编码器的编码器和解码器分别记为E()和D();编码器首先将特征图F映射到一个具有低维度的潜在空间特征图$\mathrm{F}_{\mathrm{latent}}$,$F_{latent}\in\mathrm{R}^{H_F\times W_F\times C_{latent}}$,解码器将潜在空间特征图$\mathrm{F}_{\mathrm{latent}}$重构为原始维度,得到重构特征$\mathrm{F}_{\mathrm{rec}}$,$\mathrm{F}_{rec}\in\mathrm{R}^{H_F\times W_F\times C}$;使用$1_2$距离来衡量每个对象的重构误差,并将每个对象的重构误差作为自动编码器的训练损失,该过程可以表示如下:

$\mathrm{F}_{\mathrm{rec}}=\mathrm{D}(\mathrm{E}(\mathrm{F}))$; (1)

$$L_{autoencoder}=\frac{1}{H_F\times W_F}\sum_{i=1}^{H_F}\sum_{j=1}^{W_F}L_2(F_{rec}[i,j],F[i,j])\text{;}\quad(2)$$

其中,$\mathrm{L}_{\mathrm{autoencoder}}$表示自动编码器的训练损失,$\mathrm{R}^{H_I\times W_I\times 3}$表示输入图像I属于尺寸维度为$\mathrm{H}_{\mathrm{I}}\times\mathrm{W}_{\mathrm{I}}\times 3$的矩阵,$\mathrm{H}_{\mathrm{F}}$表示输入图像I的高度,$\mathrm{W}_{\mathrm{F}}$表示输入图像I的宽度,C表示输入图像I的通道数,$\mathrm{R}^{H_F\times W_F\times C_{latent}}$表示潜在空间特征图$\mathrm{F}_{\mathrm{latent}}$属于尺寸维度为$\mathrm{H}_{\mathrm{F}}\times\mathrm{W}_{\mathrm{F}}\times\mathrm{C}_{\mathrm{latent}}$的矩阵,$\mathrm{C}_{\mathrm{latent}}$表示潜在空间特征图$\mathrm{F}_{\mathrm{latent}}$的通道数,[i,j]表示特征图的在特征空间中的位置(i,j),$\mathrm{F}_{\mathrm{rec}}$[i,j]和F[i,j]表示位置[i,j]中具有C维的区域特征;$\mathrm{L}_2$表示$1_2$范数损失;

每个对象的区域特征表示相应位置的锚框的特征,根据每个对象相应的锚框为每个对象分配前景/背景标签;当自动编码器训练到收敛状态时,通过计算每个对象的$1_2$距离,即E[i,j]=$\mathrm{L}_2$($\mathrm{F}_{\mathrm{rec}}$[i,j],F[i,j]),得到重构误差图E,$\mathrm{E}\in\mathrm{R}^{H_F\times W_F\times 1}$;通过从MS-COCO数据集的训练集中随机抽取已知和背景区域中的对象,收集一组重构误差,分别记为$\varepsilon_{\mathrm{kn}}$和$\varepsilon_{\mathrm{bg}}$;利用从已知对象的样本中抽取的重构误差来估计所有前景区域的分布;

已知区域和背景区域的韦布尔分布分别记为$\mathrm{f}_{\mathrm{kn}}$和$\mathrm{f}_{\mathrm{bg}}$,具体形式如下:

$$f(r_e;a,c)=ac[1-\exp(-r_e^{\,c})]^{a-1}\exp(-r_e^{\,c})r_e^{\,c-1}\quad(3)$$

其中,$\mathrm{r}_{\mathrm{e}}$表示样本对象的重构误差值;f是指数化韦布尔分布的概率密度函数,a和c是概率密度函数形状参数;通过基于前景$\varepsilon_{\mathrm{kn}}$和背景区域$\varepsilon_{\mathrm{bg}}$的采样重构误差,使用最大似然估计

(MLE)计算出最优的a和c;

基于对象重构的韦布尔模型ORW中,计算伪未知对象软标签,并估计未知对象的可能性得分,在对前景和背景区域的分布进行建模后,使用概率函数$f_{kn}$和$f_{bg}$来估计未知对象成为真正未知对象的可能性,具体如下:

给定图像I中的一个伪未知对象$p_{unk}$,使用RoIAlign操作将$p_{unk}$的重构误差池化成一个标量值,如下所示:

$$r_e(p_{unk}) = R_A(E, p_{unk}) \quad (4)$$

其中,$r_e(p_{unk})$是伪未知对象$p_{unk}$的重构误差值;$R_A$表示RoIAlign操作,RoIAlign全称Region of Interest Align,是一种用于目标检测中的特征对齐操作;$E \in R^{H_F \times W_F \times 1}$表示计算得到的重构误差图;然后,使用以下方程计算软标签,该软标签$s(p_{unk})$估计了未知对象成为真实未知对象的可能性得分:

$$s(p_{unk}) = \left( \frac{f_{kn}(r_e(p_{unk}))}{f_{bg}(r_e(p_{unk})) + f_{kn}(r_e(p_{unk}))} \right)^{\gamma} \quad (5)$$

其中,$f_{kn}$和$f_{bg}$分别是输入的图像中已知对象和背景区域的韦布尔概率密度函数,γ是用来缩放可能性得分值的超参数;当γ ———→∞时,所有原始伪标签将被丢弃,当γ ———→0时,所有原始伪标签对应的未知对象将被视为真实未知对象;

S3、在训练阶段,解耦目标检测器的RPN区域建议生成和ROI分类的联合训练,提升区域建议对未知类别的泛化性能,然后利用已知对象的标签和未知对象的伪标签训练目标检测器,得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD;

S4、在推理阶段,输入需要进行开放世界目标检测的图像,采用基于视觉大模型增强的开放世界目标检测模型SAM-OWOD识别已知和未知类别;

S5、根据提供的未知类标签,利用增量学习方法学习新类,进而循环实现开放世界目标检测。

2.根据权利要求1所述的一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,步骤S1中,输入的图像中包括已经标注了的已知对象,以及未标注的未知对象;

使用SAM对输入图像进行预处理,利用SAM分割任何物体的能力,对图像进行分割,根据掩膜生成候选框,候选框的对象包括已知对象、未知对象;为了得到未知对象的伪标签,通过NMS过滤重复候选框,然后计算SAM生成的对象候选框与已标注对象的候选框的IOU,如果IOU小于设定的阈值a,则认为是未知对象,从而得到未知对象的原始伪标签;

其中输入的图像是目标检测常用的MS-COCO标准数据集,所述SAM为分割任何物体的大型视觉模型,NMS是非极大值抑制,IOU是用于评估两个候选框重叠程度的度量。

3.根据权利要求1所述的一种基于视觉大模型增强的开放世界目标检测方法,步骤S3中,将目标检测器的RPN区域建议生成和ROI分类这两个阶段分离并分别训练,然后利用已知对象的标签和未知对象的伪标签训练目标检测器,提升未知对象识别的准确性,具体如下:

首先第一阶段,使用骨干网络训练RPN,生成区域建议,然后冻结RPN训练参数,第二阶段,使用生成区域建议,继续ROI分类训练,此阶段用于预测未见过的类别,接着利用公式(5)中的未知对象可能性得分$s(p_{unk})$加入到Faster-RCNN检测器的分类损失$L_{cls}$中作为一个

权重项,从而学习识别未知对象并检测已知对象;修改后的分类损失方程如下:

$$L_{cls}=\frac{1}{N_{cls}}\sum_{r}w_r\,L_{CE}(P_r,P_r^*) \quad (6)$$

其中,r表示区域提议,$w_r$是区域提议r的损失权重,当r属于伪未知对象的区域时,r等于$s(p_{unk})$,否则等于1;$p_r$表示区域提议r的预测概率,而$P_r^*$表示$p_r$的真实值,$L_{CE}$表示交叉熵损失,$N_{cls}$表示区域提议的总数。

4.根据权利要求1所述的一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,步骤S4中,在推理阶段,输入需要进行开放世界目标检测的图像,采用基于视觉大模型增强的开放世界目标检测模型SAM-OWOD,根据已知类别的标签识别已知类别,同时根据训练得到的未知对象的标签识别未知类别,输出检测图像。

5.根据权利要求1所述的一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,步骤S5中,根据提供的未知类标签,输入新的未知类别标签,增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD,进而循环实现开放世界未知类识别;

所述增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD,利用基于样本回放的增量学习方法学习新类,即存储一部分具有代表性的旧数据,并在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整;

将基于视觉大模型增强的开放世界目标检测模型SAM-OWOD除了输出层外其他层参数冻结,只对最后输出层的参数进行调整。

6.根据权利要求5所述的一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,基于样本回放的增量学习是一种机器学习方法,具体包括以下步骤:

S5.1、初始化模型:在增量学习开始之前,初始化SAM-OWOD模型,并将其用于训练一部分数据;

S5.2、训练模型:使用一部分新的数据进行SAM-OWOD模型的训练,得到第一模型;

S5.3、样本回放:将之前训练过的数据集中的设定比例的样本存储在一个缓冲区中,称为回放缓冲区,随后从回放缓冲区中随机抽取设定比例的的样本,将随机抽取的样本与当前训练数据一起用于SAM-OWOD模型的训练,得到第二模型;

S5.4、模型更新:将第一模型与第二模型进行合并,得到训练完成的SAM-OWOD模型;

S5.5、测试模型:使用测试数据集对训练完成的SAM-OWOD模型进行测试;

S5.6、如果还有新的数据需要进行训练,返回步骤S5.2,否则,结束增量学习。

7.根据权利要求6所述的一种基于视觉大模型增强的开放世界目标检测方法,其特征在于,所述在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整是在接收到未知类别的标签时,为避免模型重新训练,使用一部分代表性的历史数据和新数据训练模型;利用预训练模型在大规模数据上学习到的通用特征,只对SAM-OWOD模型的最后几层进行微调,从而使得SAM-OWOD模型在新的任务上能够更好地适应,具体实现的流程如下:

A1、加载预训练SAM-OWOD模型:使用已经在大规模数据上预训练好的SAM-OWOD模型作为初始模型;

A2、冻结模型参数:对于不需要微调的层,将它们的参数冻结,使得它们在训练过程中

不会发生变化;

A3、替换输出层:将SAM-OWOD模型的最后一层输出层替换为新的适应任务的输出层,该输出层包括新任务所需的类别数;

A4、只训练新的输出层:只对新的输出层进行训练,使得SAM-OWOD模型能够更好地适应新的任务;

A5、解冻参数:如果需要调整其他层的参数,则解冻需要调整的网络层的参数,让需要调整的网络层能够在调整中发生变化;

A6、微调模型:对整个SAM-OWOD模型进行调整,直到SAM-OWOD模型在新的任务上收敛。

# 一种基于视觉大模型增强的开放世界目标检测方法

## 技术领域

[0001] 本发明属于信息技术领域,具体涉及一种基于视觉大模型增强的开放世界目标检测方法。

## 背景技术

[0002] 随着深度学习方法的不断发展,加快了目标检测研究的进度,目标检测的任务是识别和定位图像中的目标,传统目标检测方法都是针对在一个封闭的集合下工作,也就是在训练阶段的所有类是已知的,所以它们只能检测已知类别,如果我们的集合是在开放世界时,出现了两个比较有挑战的问题:1)测试过程中图像含未知类别,这些未知类需要检测为未知类,2)当给予未知类相应的标签时,模型需要增量学习新类,我们把这个问题定义为开放世界标检测。

[0003] 开放世界目标检测方法不仅需要识别已知类别,而且需要将所有未知实例识别为未知,然后,人类注释者可以为感兴趣的类添加标签,模型在下一个任务中增量学习这些类;但在某些情况下,识别未知对象是至关重要的。例如,自动驾驶汽车或机器人需要检测未知的障碍物,以避免碰撞并确保安全。

[0004] 目前开放世界目标检测方法,侧重于通过伪标签的方式将那些与已知对象不重叠且具有较高目标性得分的区域标记为潜在的未知类别,它们的性能在很大程度上依赖于已知对象的监督。这些方法可以成功地检测出具有与已知对象相似特征的未知对象。然而,它们存在严重的标签偏差问题,即它们倾向于将与已知不相似的所有区域都视为背景的一部分(一种基于深度神经网络的开集目标检测与识别方法(CN114241260A))。

## 发明内容

[0005] 本发明旨在解决上述问题,为此,本发明的目的在于提出一种基于视觉大模型增强的开放世界目标检测方法,利用视觉大模型对输入图像预处理,无监督的方式获取未知对象的原始伪标签,然后利用提出的基于对象重构的韦布尔模型对未知对象进行建模,实现了开放环境下对已知和未知类别的检测,减少了人工标注的成本,提高了开放世界下目标检测精度。

[0006] 本发明的目的至少通过如下技术方案之一实现。

[0007] 一种基于视觉大模型增强的开放世界目标检测方法,包括以下步骤:

[0008] S1、利用视觉大模型对输入的图像进行预处理,获取未知对象的原始伪标签;

[0009] S2、利用基于对象重构的韦布尔模型,对前景和背景区域建模,并且为伪未知对象计算软标签,估计未知对象的可能性得分,判断未知对象是否是真实未知对象;

[0010] S3、在训练阶段,解耦目标检测器的RPN区域建议生成和ROI分类的联合训练,提升区域建议对未知类别的泛化性能,然后利用已知对象的标签和未知对象的伪标签训练目标检测器,得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD;

[0011] S4、在推理阶段,输入需要进行开放世界目标检测的图像,采用基于视觉大模型增

强的开放世界目标检测模型SAM-OWOD识别已知和未知类别；

[0012] S5、根据提供的未知类标签，利用增量学习方法学习新类，进而循环实现开放世界目标检测。

[0013] 进一步地，步骤S1中，输入的图像中包括已经标注了的已知对象，以及未标注的未知对象；

[0014] 使用SAM对输入图像进行预处理，利用SAM分割任何物体的能力，对图像进行分割，根据掩膜生成候选框，候选框的对象包括已知对象、未知对象；为了得到未知对象的伪标签，通过NMS过滤重复候选框，然后计算SAM生成的对象候选框与已标注对象的候选框的IOU，如果IOU小于设定的阈值a，则认为是未知对象，从而得到未知对象的原始伪标签；

[0015] 其中输入的图像是目标检测常用的MS-COCO标准数据集，SAM全称Segment Anything Model，是一种分割任何物体的大型视觉模型，NMS是非极大值抑制(Non-Maximum Suppression)的缩写；在目标检测领域，NMS是一种用于筛选具有最高置信度的候选框的技术，IOU全称Intersection over Union，是一种常用于评估两个候选框重叠程度的度量。

[0016] 进一步地，步骤S2中，根据不同图像中前景和背景出现的特征频率不同，从特征频率的角度考量，背景区域和前景区域分别形成两个不同的分布，基于对象重构的韦布尔模型ORW中，基于数据重构的自动编码器学习频率信息，并通过先验概率分布来建模重构误差的两个不同分布，由于韦布尔分布在拟合各种分布形状方面表现出色，因此韦布尔分布被用作基于对象重构的韦布尔模型ORW中的先验模型；基于对象重构的韦布尔模型ORW对前景和背景区域建模，并且为伪未知对象计算软标签，估计未知对象的可能性得分；

[0017] 为此，为了对前景和背景区域建模，基于对象重构的韦布尔模型ORW首先使用SAM分割图像中所有物体，根据掩膜生成所有对象的候选框；然后使用骨干网络对输入图像I提取特征图F，$\mathrm{I}\in\mathrm{R}^{H_I\times W_I\times 3}$，$\mathrm{F}\in\mathrm{R}^{H_F\times W_F\times C}$，为了充分表示对象语义信息，将SAM生成的所有对象的候选框映射到特征图中，每个对象的候选框的特征向量表示感受野内的区域；

[0018] 基于对象重构的韦布尔模型ORW中，利用自动编码器来重构这些区域特征，即每个对象的候选框的特征向量；自动编码器的编码器和解码器分别记为$E()$和$D()$；编码器首先将特征图$F$映射到一个具有低维度的潜在空间特征图$F_{latent}$，$F_{latent}\in\mathrm{R}^{H_F\times W_F\times C_{latent}}$，解码器将潜在空间特征图$F_{latent}$重构为原始维度，得到重构特征$\mathrm{F}_{rec}$，$\mathrm{F}_{rec}\in\mathrm{R}^{H_F\times W_F\times C}$；使用$\ell_2$距离来衡量每个对象的重构误差，并将每个对象的重构误差作为自动编码器的训练损失，该过程可以表示如下：

[0019] $$F_{rec}=D(E(F))\text{；(1)}$$

[0020] $$L_{autoencoder}=\frac{1}{H_F\times W_F}\sum_{i=1}^{H_F}\sum_{j=1}^{W_F}L_2(F_{rec}[i,j],F[i,j])$$

[0021] ；(2)

[0022] 其中，$L_{autoencoder}$表示自动编码器的训练损失，$\mathrm{R}^{H_I\times W_I\times 3}$表示输入图像I属于尺寸维度为$H_I\times W_I\times 3$的矩阵，$H_F$表示输入图像I的高度，$W_F$表示输入图像I的宽度，$C$表示输入图像I的通道数，$\mathrm{R}^{H_F\times W_F\times C_{latent}}$表示潜在空间特征图$F_{latent}$属于尺寸维度为$H_F\times W_F\times C_{latent}$的

矩阵，$C_{latent}$表示潜在空间特征图$F_{latent}$的通道数，$[i,j]$表示特征图的在特征空间中的位置$(i,j)$，$F_{rec}[i,j]$和$F[i,j]$表示位置$[i,j]$中具有$C$维的区域特征；$L_2$表示$\ell_2$范数损失。

[0023] 进一步地，每个对象的区域特征表示相应位置的锚框的特征，因此根据每个对象相应的锚框为每个对象分配前景/背景标签；如前面讨论的，背景区域通常具有频繁出现的特征，使它们更容易被重构，并且与各种前景对象区域的不常见特征相比，其重构误差较小；当自动编码器训练到收敛状态时，通过计算每个对象的$\ell_2$距离，即$E[i,j]=L_2(F_{rec}[i,j],F[i,j])$，得到重构误差图E，$\mathrm{E}\in\mathrm{R}^{H_F\times W_F\times 1}$；通过从MS-COCO数据集的训练集中随机抽取已知和背景区域中的对象，收集一组重构误差，分别记为$\varepsilon_{kn}$和$\varepsilon_{bg}$；已知对象区域的重构误差通常比背景区域的重构误差要大得多，尽管未知对象可能与已知对象具有不同的外观，但可以假设它们具有类似的低发生频率和高重构误差，因为存在各种类型的未知对象；利用从已知对象的样本中抽取的重构误差来估计所有前景区域的分布；

[0024] 由于韦布尔分布在拟合许多现实世界场景的广泛分布形状方面具有优势，因此它被用作ORW中的先验模型。已知区域和背景区域的韦布尔分布分别记为$f_{kn}$和$f_{bg}$，具体形式如下：

[0025] $$f(r_e;a,c)=ac[1-\exp(-r_e^{\,c})]^{a-1}\exp(-r_e^{\,c})r_e^{\,c-1} \quad (3)$$

[0026] 其中，$r_e$表示样本对象的重构误差值；$f$是指数化韦布尔分布的概率密度函数，$a$和$c$是概率密度函数形状参数；通过基于前景$\varepsilon_{kn}$和背景区域$\varepsilon_{bg}$的采样重构误差，使用最大似然估计(MLE)计算出最优的$a$和$c$。

[0027] 进一步地，基于对象重构的韦布尔模型ORW中，为了计算伪未知对象软标签，并估计未知对象的可能性得分，在对前景和背景区域的分布进行建模后，使用概率函数$f_{kn}$和$f_{bg}$来估计伪未知对象成为真正未知对象的可能性，具体如下：

[0028] 给定图像I中的一个伪未知对象$p_{unk}$，使用RoIAlign操作将$p_{unk}$的重构误差池化成一个标量值，如下所示：

[0029] $$r_e(p_{unk})=R_A(E,p_{unk}) \quad (4)$$

[0030] 其中，$r_e(p_{unk})$是伪未知对象$p_{unk}$的重构误差值；$R_A$表示RoIAlign 操作，RoIAlign全称Region of Interest Align，是一种用于目标检测中的特征对齐操作；$\mathrm{E}\in\mathrm{R}^{H_F\times W_F\times 1}$表示计算得到的重构误差图；然后，使用以下方程计算软标签，该软标签$s(p_{unk})$估计了未知对象成为真实未知对象的可能性得分：

[0031] $$s(p_{unk})=\left(\frac{f_{kn}(r_e(p_{unk}))}{f_{bg}(r_e(p_{unk}))+f_{kn}(r_e(p_{unk}))}\right)^{\gamma}$$

[0032] (5)

[0033] 其中，$f_{kn}$和$f_{bg}$分别是输入的图像中已知对象和背景区域的韦布尔概率密度函数，$\gamma$是用来缩放可能性得分值的超参数；当$\gamma \longrightarrow \infty$时，所有原始伪标签将被丢弃，当$\gamma \longrightarrow 0$时，所有原始伪标签对应的未知对象将被视为真实未知对象。

[0034] 进一步地，步骤S3中，在开放世界目标检测训练过程中，RPN区域提议生成和ROI分类阶段表现不同，提议生成阶段具有泛化能力，因为其类别无关的分类可以轻松扩展到新类别；相比之下，特定类别的ROI分类阶段甚至无法用于新的类别，导致偏向基本类别；这些不同的特性影响它们的联合训练，因为ROI分类阶段对新类别的敏感性会阻碍提议生成阶段的泛化性能；将目标检测器的RPN区域建议生成和ROI分类这两个阶段分离并分别训练，以避免这种冲突，然后利用已知对象的标签和未知对象的伪标签训练目标检测器，提升未知对象识别的准确性；其中RPN的全称是Region Proposal Network，是Faster R-CNN中的一个模块，用于生成目标检测中的候选区域，ROI全称是Region of Interest，在目标检测领域，ROI 是指图像中被认为是具有特殊兴趣或目标的区域，本发明使用的目标检测器Faster R-CNN，是一种两阶段目标检测模型，具体如下：

[0035] 首先第一阶段，使用骨干网络训练RPN，生成区域建议，然后冻结RPN训练参数，第二阶段，使用生成区域建议，继续ROI分类训练，此阶段用于预测未见过的类别，接着利用公式 (5) 中的未知对象可能性得分$s(p_{unk})$加入到 Faster-RCNN 检测器的分类损失$L_{cls}$中作为一个权重项，从而学习识别未知对象并检测已知对象；修改后的分类损失方程如下：

[0036] $$L_{cls}=\frac{1}{N_{cls}}\sum_{r} w_r L_{CE}(P_r, P_r^*) \qquad (6)$$

[0037] 其中，$r$表示区域提议，$w_r$是区域提议$r$的损失权重，当$r$属于伪未知对象的区域时，$r$等于$s(p_{unk})$，否则等于 1；$p_r$表示区域提议$r$的预测概率，而$P_r^*$表示$p_r$的真实值，$L_{CE}$表示交叉熵损失，$N_{cls}$表示区域提议的总数。

[0038] 进一步地，步骤S4中，在推理阶段，输入需要进行开放世界目标检测的图像，采用基于视觉大模型增强的开放世界目标检测模型SAM-OWOD，根据已知类别的标签识别已知类别，同时根据训练得到的未知对象的标签识别未知类别，输出检测图像。

[0039] 进一步地，步骤S5中，根据提供的未知类标签，输入新的未知类别标签，增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD，进而循环实现开放世界未知类识别；

[0040] 所述增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD，利用基于样本回放的增量学习方法学习新类，即存储一部分具有代表性的旧数据，并在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整；

[0041] 将基于视觉大模型增强的开放世界目标检测模型SAM-OWOD除了输出层外其他层参数冻结，只对最后输出层的参数进行调整。

[0042] 进一步地，基于样本回放的增量学习是一种机器学习方法，具体包括以下步骤：

[0043] S5.1、初始化模型：在增量学习开始之前，初始化SAM-OWOD模型，并将其用于训练一部分数据；

[0044] S5.2、训练模型：使用一部分新的数据进行SAM-OWOD模型的训练，得到第一模型；

[0045] S5.3、样本回放:将之前训练过的数据集中的设定比例的样本存储在一个缓冲区中,称为回放缓冲区,随后从回放缓冲区中随机抽取设定比例的的样本,将随机抽取的样本与当前训练数据一起用于SAM-OWOD模型的训练,得到第二模型;

[0046] S5.4、模型更新:将第一模型与第二模型进行合并,得到训练完成的SAM-OWOD模型;

[0047] S5.5、测试模型:使用测试数据集对训练完成的SAM-OWOD模型进行测试;

[0048] S5.6、如果还有新的数据需要进行训练,返回步骤S5.2,否则,结束增量学习。

[0049] 进一步地,所述在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整是在接收到未知类别的标签时,为了避免模型重新训练,使用一部分代表性的历史数据和新数据训练模型;利用预训练模型在大规模数据上学习到的通用特征,只对SAM-OWOD模型的最后几层进行微调,从而使得SAM-OWOD模型在新的任务上能够更好地适应,具体实现的流程如下:

[0050] A1、加载预训练SAM-OWOD模型:使用已经在大规模数据上预训练好的SAM-OWOD模型作为初始模型;

[0051] A2、冻结模型参数:对于不需要微调的层,将它们的参数冻结,使得它们在训练过程中不会发生变化;

[0052] A3、替换输出层:将SAM-OWOD模型的最后一层输出层替换为新的适应任务的输出层,该输出层包括新任务所需的类别数;

[0053] A4、只训练新的输出层:只对新的输出层进行训练,使得SAM-OWOD模型能够更好地适应新的任务;

[0054] A5、解冻参数:如果需要调整其他层的参数,则解冻需要调整的网络层的参数,让需要调整的网络层能够在调整中发生变化;

[0055] A6、微调模型:对整个SAM-OWOD模型进行调整,直到SAM-OWOD模型在新的任务上收敛。

[0056] 相比于现有技术,本发明的优点在于:

[0057] 目前开放世界目标检测方法在实现过程中,侧重于通过伪标签的方式将那些与已知对象不重叠且具有较高目标性得分的区域标记为潜在的未知类别,它们的性能在很大程度上依赖于已知对象的监督。这些方法可以成功地检测出具有与已知对象相似特征的未知对象。然而,它们存在已知类别的严重标签偏差问题,即它们倾向于将与已知不相似的所有区域都视为背景的一部分。本发明通过视觉大模型和无监督的未知识别方法,解决了开放世界目标过程存在的标签偏差问题,提高了开放场景目标检测的精度。

## 附图说明

[0058] 图1为本发明实施例中一种基于视觉大模型增强的开放世界目标检测方法的流程图;

[0059] 图2为本发明实施例中自动编码器区域特征重构示意图;

[0060] 图3为本发明实施例韦布尔区域特征建模示意图;

[0061] 图4为本发明实施例中的效果图。

## 具体实施方式

[0062] 为使本发明地目的、技术方案和优点更加清楚明白,下面结合附图并举实施例,对本发明地具体实施进行详细说明,显然,所描述的实施例是本发明一部分实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得地所有其他实施例,都属于本发明保护的范围。

[0063] 实施例:

[0064] 一种基于视觉大模型增强的开放世界目标检测方法,如图1所示,包括以下步骤:

[0065] S1、利用视觉大模型对输入的图像进行预处理,获取未知对象的原始伪标签;

[0066] 输入的图像中包括已经标注了的已知对象,以及未标注的未知对象;

[0067] 使用SAM对输入图像进行预处理,利用SAM分割任何物体的能力,对图像进行分割,根据掩膜生成候选框,候选框的对象包括已知对象、未知对象;为了得到未知对象的伪标签,通过NMS过滤重复候选框,然后计算SAM生成的对象候选框与已标注对象的候选框的IOU,如果IOU小于设定的阈值a,则认为是未知对象,从而得到未知对象的原始伪标签;

[0068] 其中输入的图像是目标检测常用的MS-COCO标准数据集,SAM全称Segment Anything Model,是一种分割任何物体的大型视觉模型,NMS是非极大值抑制(Non-Maximum Suppression)的缩写;在目标检测领域,NMS是一种用于筛选具有最高置信度的候选框的技术,IOU全称Intersection over Union,是一种常用于评估两个候选框重叠程度的度量。

[0069] 在一个实施例中,设置IOU阈值a为0.3,非极大值抑制NMS设置为0.35。

[0070] S2、利用基于对象重构的韦布尔模型,对前景和背景区域建模,并且为伪未知对象计算软标签,估计未知对象的可能性得分,判断未知对象是否是真实未知对象;

[0071] 如图2所示,根据不同图像中前景和背景出现的特征频率不同,从特征频率的角度考量,背景区域和前景区域分别形成两个不同的分布,基于对象重构的韦布尔模型ORW中,基于数据重构的自动编码器学习频率信息,并通过先验概率分布来建模重构误差的两个不同分布,由于韦布尔分布在拟合各种分布形状方面表现出色,因此韦布尔分布被用作基于对象重构的韦布尔模型ORW中的先验模型;基于对象重构的韦布尔模型ORW对前景和背景区域建模,并且为伪未知对象计算软标签,估计未知对象的可能性得分;

[0072] 为此,为了对前景和背景区域建模,基于对象重构的韦布尔模型ORW首先使用SAM分割图像中所有物体,根据掩膜生成所有对象的候选框;然后使用骨干网络对输入图像I提取特征图F,$\mathrm{I}\in\mathrm{R}^{H_I\times W_I\times 3}$,$\mathrm{F}\in\mathrm{R}^{H_F\times W_F\times C}$,为了充分表示对象语义信息,将SAM生成的所有对象的候选框映射到特征图中,每个对象的候选框的特征向量表示感受野内的区域;

[0073] 基于对象重构的韦布尔模型ORW中,利用自动编码器来重构这些区域特征,即每个对象的候选框的特征向量;自动编码器的编码器和解码器分别记为$E()$和$D()$;编码器首先将特征图$F$映射到一个具有低维度(通道数)的潜在空间特征图$F_{latent}$,在一个实施例中,潜在空间特征图$F_{latent}$的通道数是特征图$F$的1/2,$F_{latent}\in\mathrm{R}^{H_F\times W_F\times C_{latent}}$,解码器将潜在空间特征图$F_{latent}$重构为原始维度,得到重构特征$\mathrm{F}_{rec}$,$\mathrm{F}_{rec}\in\mathrm{R}^{H_F\times W_F\times C}$;使用$\ell_2$距离来衡量每个对象的重构误差,并将每个对象的重构误差作为自动编码器的训练损失,该过程可以表示如下:

[0074] $F_{rec} = D(E(F))$ ; (1)

[0075] $$L_{autoencoder} = \frac{1}{H_F \times W_F} \sum_{i=1}^{H_F} \sum_{j=1}^{W_F} L_2(F_{rec}[i,j], F[i,j])$$

[0076] ; (2)

[0077] 其中，$L_{autoencoder}$表示自动编码器的训练损失，$\mathrm{R}^{H_I \times W_I \times 3}$表示输入图像I属于尺寸维度为$H_I \times W_I \times 3$的矩阵，$H_F$表示输入图像I的高度，$W_F$表示输入图像I的宽度，$C$表示输入图像I的通道数，$\mathrm{R}^{H_F \times W_F \times C_{latent}}$表示潜在空间特征图$F_{latent}$属于尺寸维度为$H_F \times W_F \times C_{latent}$的矩阵，$C_{latent}$表示潜在空间特征图$F_{latent}$的通道数，$[i,j]$表示特征图的在特征空间中的位置$(i,j)$，$F_{rec}[i,j]$和$F[i,j]$表示位置$[i,j]$中具有$C$维的区域特征；$L_2$表示$\ell_2$范数损失。

[0078] 进一步地，每个对象的区域特征表示相应位置的锚框的特征，因此根据每个对象相应的锚框为每个对象分配前景/背景标签；如前面讨论的，背景区域通常具有频繁出现的特征，使它们更容易被重构，并且与各种前景对象区域的不常见特征相比，其重构误差较小；当自动编码器训练到收敛状态时，通过计算每个对象的$\ell_2$距离，即$E[i,j] = L_2(F_{rec}[i,j], F[i,j])$，得到重构误差图E，$\mathrm{E} \in \mathrm{R}^{H_F \times W_F \times 1}$；通过从MS-COCO数据集的训练集中随机抽取已知和背景区域中的对象，收集一组重构误差，分别记为$\varepsilon_{kn}$和$\varepsilon_{bg}$；已知对象区域的重构误差通常比背景区域的重构误差要大得多，尽管未知对象可能与已知对象具有不同的外观，但可以假设它们具有类似的低发生频率和高重构误差，因为存在各种类型的未知对象；利用从已知对象的样本中抽取的重构误差来估计所有前景区域的分布；

[0079] 如图3所示，由于韦布尔分布在拟合许多现实世界场景的广泛分布形状方面具有优势，因此它被用作ORW中的先验模型。已知区域和背景区域的韦布尔分布分别记为$f_{kn}$和$f_{bg}$，具体形式如下：

[0080] $$f(r_e; a, c) = ac[1 - \exp(-r_e^c)]^{a-1} \exp(-r_e^c) r_e^{c-1} \quad (3)$$

[0081] 其中，$r_e$表示样本对象的重构误差值；$f$是指数化韦布尔分布的概率密度函数，$a$和$c$是概率密度函数形状参数；通过基于前景$\varepsilon_{kn}$和背景区域$\varepsilon_{bg}$的采样重构误差，使用最大似然估计(MLE)计算出最优的$a$和$c$。

[0082] 基于对象重构的韦布尔模型ORW中，为了计算伪未知对象软标签，并估计伪未知对象的可能性得分，在对前景和背景区域的分布进行建模后，使用概率函数$f_{kn}$和$f_{bg}$来估计伪未知对象成为真正未知对象的可能性，具体如下：

[0083] 给定图像I中的一个伪未知对象$p_{unk}$，使用RoIAlign操作将$p_{unk}$的重构误差池化成一个标量值，如下所示：

[0084] $r_e(p_{unk}) = R_A(E, p_{unk})$ (4)

[0085] 其中，$r_e(p_{unk})$是伪未知对象$p_{unk}$的重构误差值；$R_A$表示RoIAlign 操作，RoIAlign全称Region of Interest Align，是一种用于目标检测中的特征对齐操作；$E \in R^{H_F \times W_F \times 1}$表示计算得到的重构误差图；然后，使用以下方程计算软标签，该软标签$s(p_{unk})$估计了伪未知对象成为真实未知对象的可能性得分：

[0086] $$s(p_{unk}) = \left( \frac{f_{kn}(r_e(p_{unk}))}{f_{bg}(r_e(p_{unk})) + f_{kn}(r_e(p_{unk}))} \right)^{\gamma}$$

[0087] (5)

[0088] 其中，$f_{kn}$和$f_{bg}$分别是输入的图像中已知对象和背景区域的韦布尔概率密度函数，$\gamma$是用来缩放可能性得分值的超参数；当$\gamma \longrightarrow \infty$时，所有原始伪标签将被丢弃，当$\gamma \longrightarrow 0$时，所有原始伪标签对应的未知对象将被视为真实未知对象。

[0089] S3、在训练阶段，解耦目标检测器的RPN区域建议生成和ROI分类的联合训练，提升区域建议对未知类别的泛化性能，然后利用已知对象的标签和未知对象的伪标签训练目标检测器，得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD；

[0090] 在开放世界目标检测训练过程中，RPN区域提议生成和ROI分类阶段表现不同，提议生成阶段具有泛化能力，因为其类别无关的分类可以轻松扩展到新类别；相比之下，特定类别的ROI分类阶段甚至无法用于新的类别，导致偏向基本类别；这些不同的特性影响它们的联合训练，因为ROI分类阶段对新类别的敏感性会阻碍提议生成阶段的泛化性能；将目标检测器的RPN区域建议生成和ROI分类这两个阶段分离并分别训练，以避免这种冲突，然后利用已知对象的标签和未知对象的伪标签训练目标检测器，提升未知对象识别的准确性；其中RPN的全称是Region Proposal Network，是Faster R-CNN中的一个模块，用于生成目标检测中的候选区域，ROI全称是Region of Interest，在目标检测领域，ROI 是指图像中被认为是具有特殊兴趣或目标的区域，本发明使用的目标检测器Faster R-CNN，是一种两阶段目标检测模型，具体如下：

[0091] 首先第一阶段，使用骨干网络训练RPN，生成区域建议，然后冻结RPN训练参数，第二阶段，使用生成区域建议，继续ROI分类训练，此阶段用于预测未见过的类别，接着利用公式 (5) 中的未知对象可能性得分$s(p_{unk})$加入到 Faster-RCNN 检测器的分类损失$L_{cls}$中作为一个权重项，从而学习识别未知对象并检测已知对象；修改后的分类损失方程如下：

[0092] $$L_{cls} = \frac{1}{N_{cls}} \sum_r w_r L_{CE}(P_r, P_r^*) \quad (6)$$

[0093] 其中，$r$表示区域提议，$w_r$是区域提议$r$的损失权重，当$r$属于伪未知对象的区域时，$r$等于$s(p_{unk})$，否则等于 1；$p_r$表示区域提议$r$的预测概率，而$P_r^*$表示$p_r$的真实值，$L_{CE}$表示交叉熵损失，$N_{cls}$表示区域提议的总数。

[0094] S4、在推理阶段，输入需要进行开放世界目标检测的图像，采用基于视觉大模型增强的开放世界目标检测模型SAM-OWOD识别已知和未知类别；

[0095] 在推理阶段，输入需要进行开放世界目标检测的图像，采用基于视觉大模型增强

的开放世界目标检测模型SAM-OWOD,根据已知类别的标签识别已知类别,同时根据训练得到的未知对象的的标签识别未知类别,输出检测图像。

[0096] S5、根据提供的未知类标签,利用增量学习方法学习新类,进而循环实现开放世界目标检测。

[0097] 根据提供的未知类标签,输入新的未知类别标签,增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD,进而循环实现开放世界未知类识别;

[0098] 所述增量学习得到基于视觉大模型增强的开放世界目标检测模型SAM-OWOD,利用基于样本回放的增量学习方法学习新类,即存储一部分具有代表性的旧数据,并在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整;

[0099] 将基于视觉大模型增强的开放世界目标检测模型SAM-OWOD除了输出层外其他层参数冻结,只对最后输出层的参数进行调整。

[0100] 基于样本回放的增量学习是一种机器学习方法,具体包括以下步骤:

[0101] S5.1、初始化模型:在增量学习开始之前,初始化SAM-OWOD模型,并将其用于训练一部分数据;

[0102] S5.2、训练模型:使用一部分新的数据进行SAM-OWOD模型的训练,得到第一模型;

[0103] S5.3、样本回放:将之前训练过的数据集中的设定比例的样本存储在一个缓冲区中,称为回放缓冲区,随后从回放缓冲区中随机抽取设定比例的的样本,将随机抽取的样本与当前训练数据一起用于SAM-OWOD模型的训练,得到第二模型;

[0104] S5.4、模型更新:将第一模型与第二模型进行合并,得到训练完成的SAM-OWOD模型;

[0105] S5.5、测试模型:使用测试数据集对训练完成的SAM-OWOD模型进行测试;

[0106] S5.6、如果还有新的数据需要进行训练,返回步骤S5.2,否则,结束增量学习。

[0107] 所述在每个增量步骤之后对基于视觉大模型增强的开放世界目标检测模型进行调整是在接收到未知类别的标签时,为了避免模型重新训练,使用一部分代表性的历史数据和新数据训练模型;利用预训练模型在大规模数据上学习到的通用特征,只对SAM-OWOD模型的最后几层进行微调,从而使得SAM-OWOD模型在新的任务上能够更好地适应,具体实现的流程如下:

[0108] A1、加载预训练SAM-OWOD模型:使用已经在大规模数据上预训练好的SAM-OWOD模型作为初始模型;

[0109] A2、冻结模型参数:对于不需要微调的层,将它们的参数冻结,使得它们在训练过程中不会发生变化;

[0110] A3、替换输出层:将SAM-OWOD模型的最后一层输出层替换为新的适应任务的输出层,该输出层包括新任务所需的类别数;

[0111] A4、只训练新的输出层:只对新的输出层进行训练,使得SAM-OWOD模型能够更好地适应新的任务;

[0112] A5、解冻参数:如果需要调整其他层的参数,则解冻需要调整的网络层的参数,让需要调整的网络层能够在调整中发生变化;

[0113] A6、微调模型:对整个SAM-OWOD模型进行调整,直到SAM-OWOD模型在新的任务上收敛。

[0114] 在一个实施例中,得到的效果图如图4中的a图和b图所示。

[0115] 在一个实施例中,为了证明本申请所提出的方法的有效性,下面进行验证实验:

[0116] 提出了一项全面的评估标准来探讨SAM-OWOD(基于视觉大模型增强的开放世界目标检测模型)的性能,包含对未知类别对象的识别,检测已知类别,以及对未知类提供标签时逐渐学习新类别。

[0117] 数据分割:在任务集$T=\{T_1,T_2,T_3,T_4\}$上评估SAM-OWOD模型。如表1所示,COCO的 80 个类别被分成四组,每组数据被视为一个流式任务的数据集。在任务 Tt 的训练过程中,一个特定任务的所有类将在时间点$t$被引入系统,对于任务$T_t$,$\{T_T:T<t\}$的为已知的,$\{T_T:T>t\}$将被视为未知。模型仅通过使用每个任务的已知类别注释的图像进行增量训练,而不是使用整个数据集。下表1显示了开放世界目标检测评估标准中的任务组成:

[0118] 表1

[0119]

| | Task 1 | Task 2 | Task 3 | Task 4 |
|---|---|---|---|---|
| Semantic split | Animals, Person, Vehicles | Appliances, Accessories, Outdoor, Furniture | Sports, Food | Electronic, Indoor, Kitchen |
| # training images | 89,490 | 55,870 | 39,402 | 38,903 |
| # train instances | 421,243 | 163,512 | 114,452 | 160,794 |
| # test images | 4,952 | | | |
| # test instances | 36,781 | | | |

[0120] 评估指标:我们采用 mAP(平均精度均值)作为评估已知类别检测性能的指标。至于未知类别,我们使用未知对象召回率(U-Recall)作为主要评估指标。

[0121] 如表2所示,在任务1、任务2、任务3、任务4在开放世界目标检测上的性能,本发明与Faster-RCNN+基准模型进行了比较,以及四个最先进的开放世界目标检测器,包括ORE 、OW-DETR 、CAT 和PROB 。Faster-RCNN+微调表示使用示例重放对Faster-RCNN进行微调。它们只能检测已知对象,因此它们对未知对象的召回结果均为零。本发明没有应用ORE的基于能量的未知标识符(EBUI),因为它需要弱的未知标签监督。ORE、OW-DETR、CAT和PROB的结果取自它们的论文;从实验结果可以看出,本发明SAM-OWOD模型,在未知召回率上有了巨大的提升,在任务1上超过CAT 26.9%,在任务2上超过CAT 27.7%,在任务3上超过PROB 18.8%,并且在已知检测精度上也有很好的提升。

[0122] 表2

[0123]

| Task-IDs (→) | Task-1 | | Task-2 | | Task-3 | | Task-4 |
|---|---|---|---|---|---|---|---|
| | U-Recall | mAP | U-Recall | mAP | U-Recall | mAP | mAP |
| Faster-RCNN+ | 0.0 | 74.4 | 0.0 | 24.7 | 0.0 | 14.5 | 10.5 |
| ORE-EBUI | 1.5 | 71.4 | 3.9 | 45.6 | 3.6 | 39.5 | 31.8 |
| OW-DETR | 5.7 | 73.1 | 6.2 | 46.0 | 6.9 | 39.7 | 33.1 |
| PROB | 17.6 | 73.5 | 22.3 | 50.4 | 24.8 | 42.0 | 39.9 |
| CAT | 24.0 | 74.2 | 23.0 | 50.7 | 24.6 | **45.0** | 42.8 |
| SAM-OWOD (Ours) | **50.9** | **74.4** | **50.7** | **52.3** | **43.6** | 44.1 | **43.2** |

[0124] 需要说明的是,任何本发明所属技术领域内的技术人员,在不脱离本发明所揭露的精神和范围的前提下,可以在实施的形式上及细节上进行变更和修改。因此,本发明的一些等同修改和变更也应该在本发明的权利要求的保护范围内。此外,尽管本说明书中使用了一些特定的术语,但这些术语只是为了方便说明,并不对本发明构成任何限制。

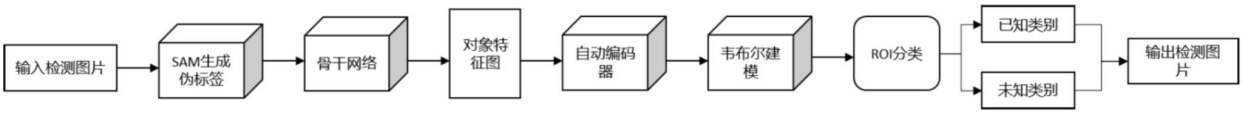
输入检测图片
SAM生成伪标签
骨干网络
对象特征图
自动编码器
韦布尔建模
ROI分类
已知类别
未知类别
输出检测图片

图1

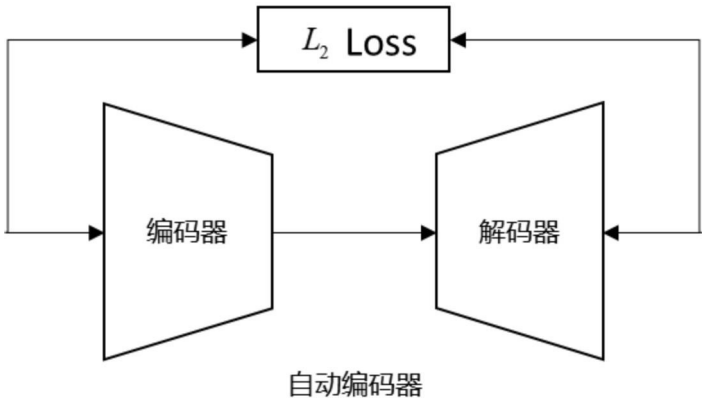
$L_2$ Loss
编码器
解码器
自动编码器

图2

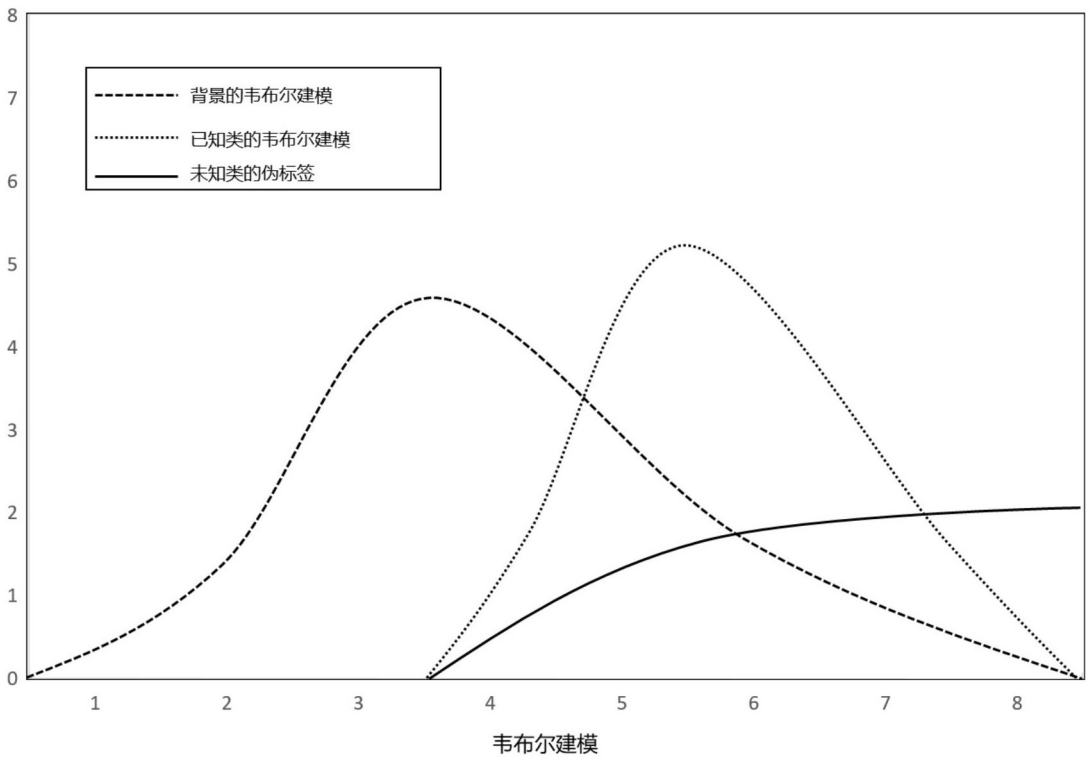
背景的韦布尔建模
已知类的韦布尔建模
未知类的伪标签
0
1
2
3
4
5
6
7
8
韦布尔建模

图3

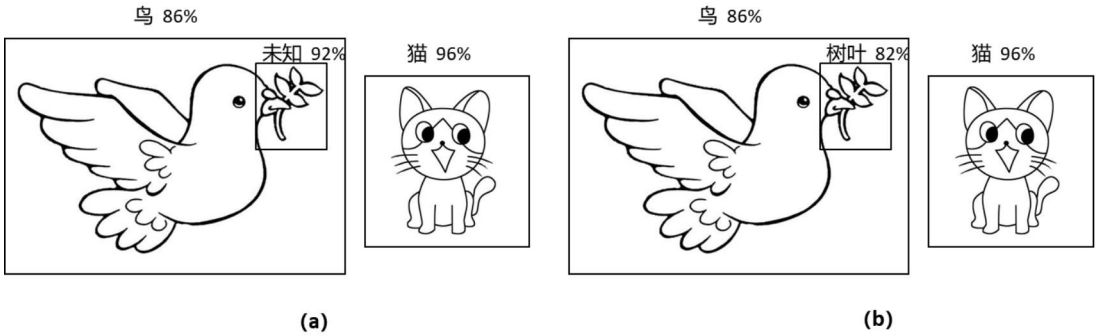


图4