计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(2): 432−442, 2025

基于图提示的半监督开放词汇多标记学习

李仲年[1,3] 皇甫志宇[1] 杨凯杰[1] 营 鹏[1] 孙统风[1,3] 许新征[1,2,3]
[1](中国矿业大学计算机科学与技术学院 江苏徐州 221116)
[2](计算机辅助设计与图形系统全国重点实验室(浙江大学) 杭州 310058)
[3](矿山数字化教育部工程研究中心(中国矿业大学) 江苏徐州 221116)
(zhongnianli@cumt.edu.cn)

Semi-supervised Open Vocabulary Multi-label Learning Based on Graph Prompting

Li Zhongnian[1,3], Huangfu Zhiyu[1], Yang Kaijie[1], Ying Peng[1], Sun Tongfeng[1,3], and Xu Xinzheng[1,2,3]
[1] (*School of Computer Science and Technology*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116)
[2] (*State Key Laboratory of CAD & CG* (*Zhejiang University*), *Hangzhou* 310058)
[3] (*Mine Digitization Engineering Research Center* (*China University of Mining and Technology*), *Ministry of Education*, *Xuzhou*, *Jiangsu* 221116)

Abstract Semi-supervised multi-label learning employs labeled and unlabeled data to train a model, which effectively achieves good results and reduces the labeling cost of multi-label data. Therefore, semi-supervised multi-label learning has attracted many researchers dedicated to this field. However, in the semi-supervised annotation process, due to the large number of labels, it is a common situation that some labels lack of samples, and these labels are called open vocabulary. It is difficult for the model to learn the label information of the open vocabulary, which leads to the degradation of its performance. To address the above problem, we propose a semi-supervised open vocabulary multi-label learning method based on graph prompting. Specifically, this method uses a graph neural network via prompt to fine-tune the pre-trained model and explore the relationship between open vocabulary and supervised samples. By using images and text, we construct a graph neural network as the input of text for the pre-trained model. Furthermore, by leveraging the generalization ability of the pre-trained model on open vocabulary, pseudo-labels are generated for unsupervised samples. Then we use pseudo-labels to train the classification layer and to enable the model to achieve better performance in classifying open vocabulary. Experimental results on multiple benchmark datasets, including VOC, COCO, CUB, and NUS, consistently demonstrate that the proposed method outperforms existing methods and achieves state-of-the-art performance.

Key words semi-supervised multi-label learning; pre-trained model; graph neural network; open vocabulary; prompt

摘 要 半监督多标记学习利用有标记数据和无标记数据进行模型的训练,降低了多标记数据的标记成本并取得了不错的结果,吸引了很多研究者不断进行研究.然而,在半监督标注过程中,由于标记的数量较多,往往会出现某些标记缺失标注样本的情况,这些标记被称为开放词汇.开放词汇会导致模型无法学

收稿日期: 2024-03-01; **修回日期**: 2024-10-31
基金项目: 国家自然科学基金项目(61976217, 62306320); 江苏省自然科学基金项目(BK20231063); 浙江大学计算机辅助设计与图形系统全国重点实验开放课题(A2424); 中国矿业大学研究生创新计划项目(2024WLJCRCZL262)
This work was supported by the National Natural Science Foundation of China (61976217,62306320), the Natural Science Foundation of Jiangsu Province (BK20231063), the Open Project Program of the State Key Laboratory of CAD&CG of Zhejiang University (A2424), and the Graduate Innovation Program of China University of Mining and Technology (2024WLJCRCZL262).
通信作者: 许新征(xxzheng@cumt.edu.cn)

习到该类别的标记信息，使得模型性能下降. 针对上述问题，提出了基于图提示的半监督开放词汇多标记学习方法. 具体地，该方法利用基于提示的图神经网络对预训练大模型进行微调，挖掘和探索开放词汇与监督样本之间的关系. 通过使用包含图像与文本的多模态数据构造图神经网络作为预训练大模型的文本输入进行学习. 其次利用预训练大模型在开放词汇上的泛化能力，对无监督样本生成伪标记，实现对输出分类层的微调，使模型在对开放词汇进行分类时能获得更加理想的效果. 多个基准数据集上的实验结果均显示，基于图提示的半监督开放词汇多标记学习方法优于目前的主流方法，在 VOC, COCO, CUB, NUS 等基准数据集上均取得了最优的效果.

关键词 半监督多标记学习；预训练模型；图神经网络；开放词汇；提示

中图法分类号 TP391

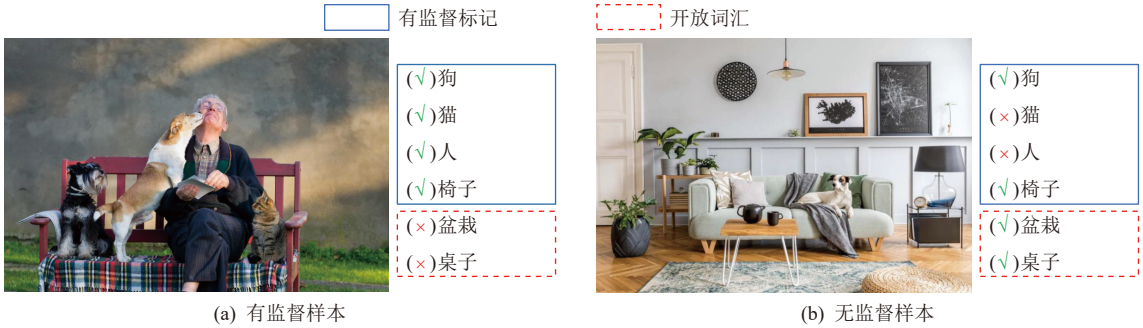
DOI: 10.7544/issn1000-1239.202440123 **CSTR:** 32373.14.issn1000-1239.202440123

在传统的机器学习任务中，一个样本通常对应一个明确的固定标记. 但随着机器学习技术的进步，许多机器学习任务中的样本可以具有多个标记，这被称为多标记学习. 多标记学习中的样本与多个标记相关联，其标记数量众多且具有复杂的语义特征. 这使得在实际应用中获取数据的真实标记常常代价过高，需要投入大量的时间和人力物力成本. 因此，许多研究者选择研究一种新的机器学习范式，即弱监督多标记学习.

常见的弱监督多标记学习包括偏多标记学习[1]、标记缺失学习[2]、半监督多标记学习[3]. 在这些多标记问题中，半监督多标记学习吸引了大量研究者的关注. 半监督学习是指在训练模型时，同时利用有标记数据和无标记数据进行学习，其目标是通过利用有标记数据和无标记数据的信息来提高模型的性能. 通常情况下，有标记数据的数量相对较少，而无标记数据的数量相对较多. 目前，半监督多标记学习在计算机视觉[4-6]、自然语言处理[7-8]、社交媒体分析[9]、生物医学[10]等方面均具有广泛应用.

针对半监督学习中缺乏标记的问题，预训练大模型和图神经网络被用于解决这种问题. 在大模型使用方面，Song 等人[11]提出了 FD-Align(feature discrimination alignment)方法，通过在训练过程中保持虚假特征的一致性以增强模型的可推广性来提高 CLIP(contrastive language-image pre-training)[12]大模型在半监督任务上的表现效果. Mo 等人[13]则提出了 S-CLIP(semi-supervised vision-language pre-training)方法，使用专门设计的伪标记策略，通过配对的图像文本提供的关键字级伪标记，以及微调 CLIP 将 CLIP 的零样本分类性能提高了 10%. 同时，在图网络方面，Wang 等人[14]提出的 HSPNet(hierarchical semantic prompt network)则通过对标记进行建图来解决单一正标记问题.

然而在传统的半监督多标记学习中，每个标记都有被标注的样本，这与真实的情况不符. 在真实情况中，由于多标记任务中标记的数量较多、被标记的样本比例很小，导致一些标记在所有样本中都未被标记，如图 1 所示. 在图 1 中，由于盆栽和桌子标记不常见，这导致了它们在有监督样本中未被标记，但出现在无监督样本中，本文称这些标记为开放词汇(open vocabulary). 然而，在传统的半监督学习中，如图 2 所示，对所有标记来说都有被标注到的样本，其有监督标记集与无监督标记集一致，即不存在开放词汇.



(a) 有监督样本 (b) 无监督样本

Fig. 1 Open-vocabulary multi-label learning

图 1 开放词汇多标记学习



(a) 开放词汇半监督数据集



(b) 传统半监督数据集

Fig. 2 Difference between open vocabulary semi-supervised and traditional semi-supervised

图 2 开放词汇半监督与传统半监督的区别

在这个设定下, 现有的模型在面对开放词汇的标记时, 没有挖掘和探索开放词汇的信息, 因此会带来模型性能的下降. 这激发了本文对于该问题的探索. 为解决半监督多标记任务中的开放词汇问题, 提出了一种基于图提示的微调大模型方法. 通过使用包含图像与文本的多模态数据构造图神经网络, 将图神经网络作为预训练大模型 RAM(recognize anything model)[15] 的文本输入进行学习. 图神经网络由节点和边组成, 它使用标记与图像作为节点, 使用边存储节点间的信息, 在图神经网络中存储标记间的联系, 以提高模型在开放词汇半监督学习中的表现效果.

本文的主要贡献包括 3 个方面:

1)提出了一种基于图卷积网络的微调大模型算法. 该算法在训练过程中利用大模型的输出作为伪标记来参与损失计算, 以提高大模型在开放词汇半监督多标记学习中的分类性能;

2)构建了基于提示的图神经网络, 通过对数据集的文本特征和图像特征进行建图, 使用图神经网络来学习图像与文本之间的联系, 以达到提高模型分类性能的效果;

3)多个基准数据集上的实验结果表明, 基于图提示的大模型微调方法在开放词汇半监督多标记学习中具有最佳的效果.

1 相关工作

1.1 半监督多标记学习

多标记学习是机器学习领域的一个重要研究方向, 旨在解决实际问题中 1 张图像或样本存在多个相关标记的情况. 在过去的几十年里, 研究者提出了许多方法和算法以处理多标记学习问题. 其中, 一些方法基于传统的机器学习技术, 如二分类器链[16]、分类器融合[17] 和标记空间分割[18] 等. 这些方法通过将

多标记问题转化为一系列单标记问题的组合或分割，来实现多标记的预测.

随着深度学习的兴起，越来越多的研究工作将深度神经网络应用于多标记学习. 不同设置下的多标记分类问题在过去几年中取得了很大进展，对损失函数和骨干网络等的改进在完全标记和部分标记的设置下都取得了很大成功. 其中，Wang 等人[19]提出了一种双重关系半监督多标记学习方法，提出了一种双分类器域自适应策略，在生成伪标记的同时对齐特征以提高学习性能. Xie 等人[20]提出了一种局部多标记学习方法来同时恢复真值信息和识别噪声标记，结合标记相关性挖掘和特征诱导噪声模型，对多标记分类器和带噪声标记识别器进行联合优化. Duan 等人[21]提出了一种基于类转移跟踪的伪整流制导方法，统一了伪校正过程引起的类分布和类转移的历史信息，提高了标记非随机丢失问题中的伪标记质量.

1.2 开放词汇

开放词汇[22]原指在自然语言处理任务中，不事先限定词汇表的大小或内容，而允许使用任意的、未知的词汇. 相比于封闭词汇(closed vocabulary)[23]的模型，开放词汇模型具有更大的词汇表覆盖范围，可以处理未见过的词汇和新兴词汇. 在计算机视觉任务中，开放词汇模型可以对未见过的物体进行标记. 传统的封闭词汇模型通常仅能识别预定义的有限标记，而开放词汇模型可以对新的、未知的物体进行分类和识别.

在目标检测方面，Gu 等人[24]使用大模型强大的泛化性能，利用知识蒸馏技术对 CLIP 进行蒸馏，以解决开放词汇目标检测问题. 在目标分割方面，Zhang 等人[25]通过引入一个预训练的文本编码器，在分割任务中对所有视觉概念进行编码，取得了优异的结果. 在多标记学习方面，He 等人[26]提出了一种采用知识蒸馏来保证图像和标记嵌入的一致性的方法，使用提示调整以进一步更新标记嵌入. 为了进一步实现对多个物体的识别，开发了一个简单但有效的双流模块来捕获局部和全局特征.

1.3 预训练大模型

大模型的泛化能力使得其在零样本学习中能够取得较好的结果. 零样本学习属于迁移学习，其目的是使训练的模型能够正确识别，仅存在于测试集中，不存在于训练集中的样本中. 这与开放词汇的概念不谋而合.

CLIP 通过对比学习的训练方式，同时处理图像和文本 2 种模态的数据，有效地提高了模型的跨模态理解能力，更好地挖掘出语言和图像之间的内在联系，提高模型的泛化能力，使得其在开放词汇任务中具有较好的性能.

BLIP(bootstrapping language-image pre-training)[27]不仅使用文本生成器合成文本，而且使用过滤器去除噪声文本，在预训练过程中实现了编码器-编码器的多模式混合，具有强大的泛化能力，同样在开放词汇任务中取得了较为理想的结果.

RAM 通过文本语义自动解析，使用大模型获取无标注的图像标记；随后通过统一标注，使用一个初步的自动标注模型产生自动标注，并使用数据引擎生成额外的注释；最后使用处理后的数据进行重新训练，以达到处理开放词汇多标记任务的效果.

CoOp(context optimization)[28]，CoCoOp(conditional context optimization)[29]使用提示学习的方法，尝试使用大模型更好地适应下游视觉任务，为微调大模型提出了崭新范式.

2 方法介绍

在本节中，提出了基于图提示的微调大模型方法. 主要介绍在开放词汇半监督多标记问题中如何使用大模型输出作为伪标记以计算损失以及如何通过图卷积神经网络提高模型的开放词汇多标记学习能力. 该方法旨在使用图卷积神经网络对图像和文本构造图神经网络，然后在图神经网络中学习开放词汇与封闭词汇之间的关系，从而实现对开放词汇样本的分类.

具体的，首先使用 CLIP 的编码器提取一个样本中的文本和图像特征. 然后将每个标记设置为 1 个节点，从而构建一张包含所有标记的图网络(见 2.2 节). 之后使用该图网络替换 RAM 中的文本输入(见 2.1 节). 在计算损失环节，对于有监督数据使用其原本标记的计算损失. 对于无监督数据使用大模型的输出作为伪标记计算其损失(见 2.3 节).

在本文中，M表示标记的个数. k表示构建多模态特征图使用的图像数量. D表示实例集. W表示标记集. T表示经过文本编码器获得的文本特征集. L表示实例集D经过图像编码器获得的图像特征集. O表示最终的多模态提示图. $G=\{g_1,g_2,\cdots,g_k\}$表示标记对应的多模态特征图. $\boldsymbol{A}_{ii}^{g_i}$表示图$g_i$的邻接矩阵，$\boldsymbol{B}_{ij}$表示图$O$对应的邻接矩阵. $X^{\mathrm{s}}=\{x_1^{\mathrm{s}},x_2^{\mathrm{s}},\cdots,x_{n_{\mathrm{s}}}^{\mathrm{s}}\}$表示有监督样本，$X^{\mathrm{u}}=\{x_1^{\mathrm{u}},x_2^{\mathrm{u}},\cdots,x_{n_{\mathrm{u}}}^{\mathrm{u}}\}$表示无监督样本，其中$n_{\mathrm{s}}$表示

有监督样本数量，n_u表示无监督样本数量. $Y^s=\{y_1^s, y_{2,}^s,\cdots,y_{n_s}^s\}$表示标记，$Y^u=\{y_1^u,y_2^u,\cdots,y_{n_u}^u\}$表示伪标记.

2.1 模型框架

如图 3 所示，本文提出的框架包含被冻结 RAM 预训练大模型、提示图网络、无监督损失与有监督损失. 其中，冻结 RAM 预训练大模型获取所有样本特征，包括有监督样本及无监督样本. 使用“A photo of a [Class]”的格式构建提示，与图像一起构建提示图网络，使用构建出的多模态提示图替换 RAM 大模型中的文本输入. 对于有监督样本，直接使用有监督标记参与损失计算，对于无监督样本，使用大模型的输出作为伪标记参与损失计算. 最后将得到的有监督损失与无监督损失相加微调大模型.

2.2 基于提示的图神经网络

关于提示的选取，在图的构建过程中共有 M 个标记，再针对每个标记获取数据集外的 k 张图像，将这 k 张图像组成实例集 $D_m=\{d_{m_1},d_{m_2},\cdots,d_{m_k}\}$. 同时获取其标记集 $W=\{w_1,w_2,\cdots,w_m\}$. 针对标记集中的每个标记，都有与其唯一对应的图像集 D. 然后将第 i 个标记通过模板“A photo of a [Class]”构造一个提示，并将提示输入到 CLIP 的文本编码器，提取 512 维的文本特征 T_i，将第 i 个标记的图像经过 CLIP 的图像编码器来提取 512 维的图像特征 L_i. 其次，将第 i 个标记的图像特征与文本特征构建为多模态特征图，记为 g_i. 其中，边的信息使用余弦相似度 $Sim(\cdot)$ 计算得到，使用 $\boldsymbol{A}_{ii}^{g_i}$ 表示其邻接矩阵. 具体计算方法如式(1)所示.

$$\boldsymbol{A}_{ii}^{g_i}=Sim(T_i\times L_i), \tag{1}$$

$$\boldsymbol{B}_{ij}=Sim\left(\boldsymbol{n}_i\cdot\boldsymbol{n}_j\right). \tag{2}$$

在得到所有标记的多模态特征图集合 $G=\{g_1,g_2,\cdots,g_m\}$ 后，将其经过 2 层图卷积层和 1 层池化层得到节点集 $N=\{\boldsymbol{n}_1,\boldsymbol{n}_2,\cdots,\boldsymbol{n}_m\}$，并对得到的节点集再次经过余弦相似度 $Sim(\cdot)$ 计算，得到最终的多模态提示图 O，使用 $\boldsymbol{B}_{ij}$ 表示其邻接矩阵(见式(2)). 最后，使用多模态提示图 O 替换 RAM 大模型中的文本输入. 具体构建过程如图 4 所示.

2.3 损失函数

通过构建出的图神经网络替换 RAM 文本嵌入，使得模型在面对开放词汇时能够通过图神经网络中的标记关系得到更多开放词汇的信息. 在计算损失时，面对已知类信息可以直接使用样本真实标记和直接参与计算，而未知标记信息则需使用伪标记作为真实标记进行损失计算.

得益于大模型优秀的泛化性能和零样本推理能力，使用 RAM 大模型的无标记损失作为伪标记参与损失计算，提出了新的经验风险函数：

$$\hat{R}(\boldsymbol{f}_\theta)=\frac{1}{n_s}\sum_{i=1}^{n_s}L\left[\boldsymbol{f}_\theta\left[f_{\mathrm{RAM}}\left(x_i^s,P\right)\right],y_i^s\right]+\frac{1}{n_u}\sum_{i=1}^{n_u}L\left[\boldsymbol{f}_\theta\left[f_{\mathrm{RAM}}\left(x_i^u,P\right)\right],y_i^u\right], \tag{3}$$

其中 f_{RAM} 表示预训练大模型的输出，$\boldsymbol{f}_\theta$ 表示全连接层，P 表示图像特征，L 表示均方误差损失，计算方法如式(4)所示，其中，$\hat{c}_i$ 表示模型的输出，c_i 表示给定

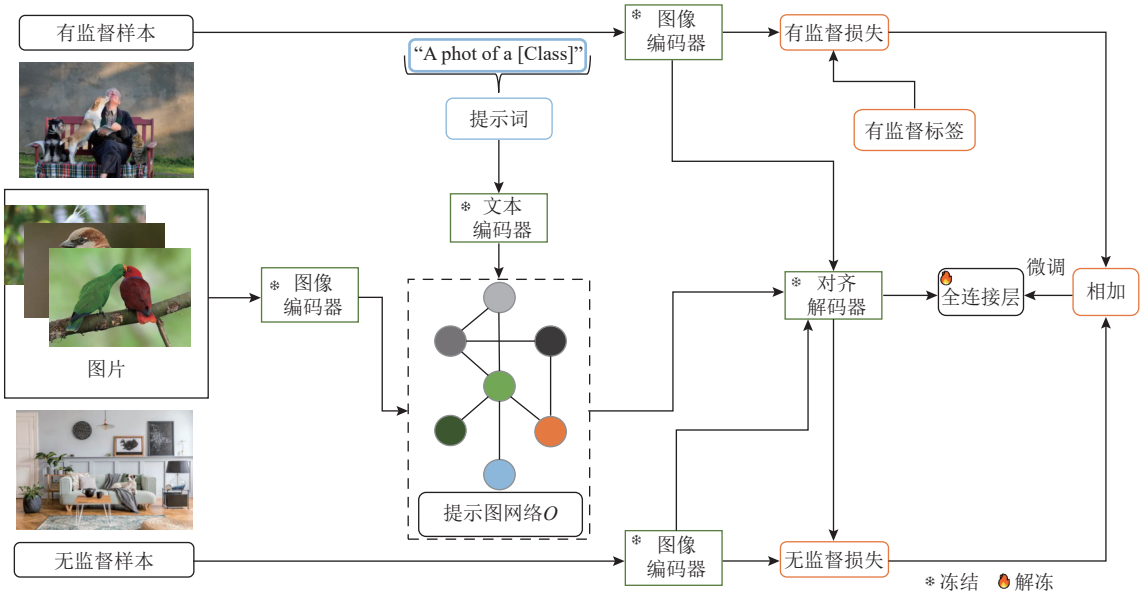


Fig. 3 Overall framework of the proposed model

图 3 本文提出的模型总体框架

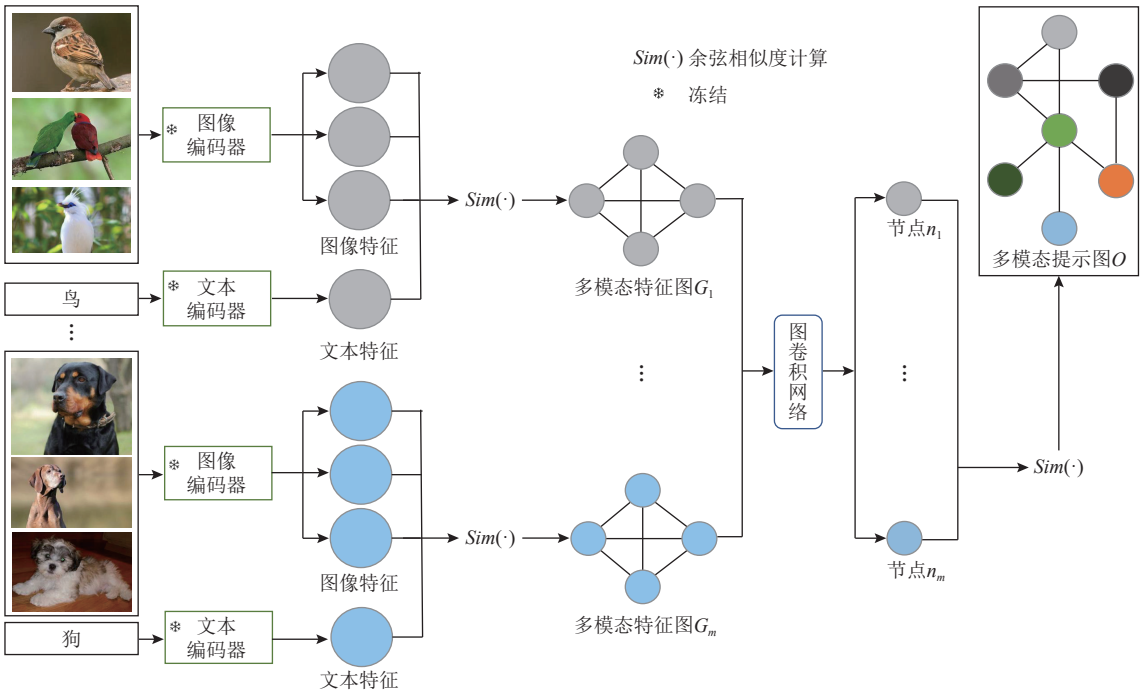


Fig. 4 Multi-modal prompt graph construction method

图 4 多模态提示图构建方法

的标记.

$$L[\hat{c}_i,c_i]=\frac{1}{M}\sum_{i=1}^{m}(\hat{c}_i-c_i)^2, \tag{4}$$

$$y_i^u=\mathrm{sgn}\left[f_{\mathrm{RAM}}\left(x_i^u,P\right)-\sigma\right]. \tag{5}$$

使用式(5)将预训练大模型的输出作为伪标记y_i^u，其中sgn表示符号函数，σ表示选择伪标记时的置信度阈值.

该经验风险函数在计算已知类信息时使用样本与其真实标记直接参与计算. 面对未知类时，使用大模型对其进行推理，将 RAM 大模型的预测输出当作该样本的伪标记，以计算标记缺失样本的损失.

2.4 算法复杂度分析

本文方法通过构造多模态提示图存储类间信息以实现开放词汇多标记任务，包括训练和测试 2 个过程.

计算空间复杂度时，由于本文方法针对每个数据集只需要构建 1 次多模态提示图，其空间占用不变，不会随训练样本数量的增加而增加，因此空间复杂度$S(n)=O(1)$.

计算时间复杂度时，训练过程主要分为 2 个阶段，即多模态提示图构建阶段和反向传播优化阶段. 多模态提示图构建阶段与模型训练无关，因此本文针对每个数据集单独构建一次多模态提示图，即时间复杂度为图卷积计算复杂度$O(n^3)$，n表示样本数量. 反向传播优化过程中，根据样本数量和迭代次数，时间复杂度为$O(T\times n\times S)$，S表示单样本梯度更新时间，T表示迭代次数. 因此，本文方法训练阶段的时间复杂度为$O(n^3+T\times n\times S)$. 在测试过程中，不需要额外构建多模态提示图，所需的时间仅为前向传播的时间，时间复杂度为$O(n\times U)$，U表示单样本前向传播时间.

2.5 算法收敛性证明

理论 1[30]. 假设第t步目标函数f_t梯度有界，其中$t\in\{1,2,\cdots,T\}$. 对于所有$\boldsymbol{\theta}\in\mathbb{R}^d$都满足$\|\nabla f_t(\boldsymbol{\theta})\|_2\leqslant G$，$\|\nabla f_t(\boldsymbol{\theta})\|_\infty\leqslant G_\infty$. 假设使用 Adam 生成的任意$\theta_t$之间的距离是有界的，对于任意$m,n\in\{1,2,\cdots,T\}$，$\|\theta_n-\theta_m\|_2\leqslant D$，$\|\theta_n-\theta_m\|_\infty\leqslant D_\infty$，其中$d$表示参数量，$G$，$G_\infty$，$D$，$D_\infty$表示常数. 对于任意$T\geqslant 1$，Adam 实现如下保证：

$$\frac{R(T)}{T}=O\left(\frac{1}{\sqrt{T}}\right), \tag{6}$$

其中，$R(T)=\sum_{t=1}^{T}[f_t(\theta_t)-f_t(\theta^*)]$，$\theta^*=\arg\min_{\theta\in x}\sum_{t=1}^{T}f_t(\theta)$.

推论 1. 假设第t步经验风险$\hat{R}^t(\boldsymbol{f}_\theta)$梯度有界，其中$t\in\{1,2,\cdots,T\}$. 对于所有$\boldsymbol{f}_\theta\in\mathbb{R}^d$都满足$\|\nabla\hat{R}^t(\boldsymbol{f}_\theta)\|_2\leqslant G$，$\|\nabla\hat{R}^t(\boldsymbol{f}_\theta)\|_\infty\leqslant G_\infty$. 假设使用 Adam 生成的任意$\boldsymbol{f}_\theta^t$之间的距离是有界的，对于任意$m,n\in\{1,2,\cdots,T\}$，$\|\boldsymbol{f}_\theta^n-\boldsymbol{f}_\theta^m\|_2\leqslant D$，$\|\boldsymbol{f}_\theta^n-\boldsymbol{f}_\theta^m\|_\infty\leqslant D_\infty$. 在使用 Adam 作为优化器的情况下，本文的经验风险$\hat{R}(\boldsymbol{f}_\theta)$收敛.

证明. 根据理论 1 与推论 1 中假设可得式(7)：

$$\frac{1}{T}\sum_{t=1}^{T}\left[\hat{R}^t(\boldsymbol{f}_\theta)-\hat{R}^t\left(\boldsymbol{f}_\theta^*\right)\right]=O\left(\frac{1}{\sqrt{T}}\right), \tag{7}$$

其中 $f_\theta^* = \arg\min_{\theta\in x}\sum_{t=1}^{T}\hat{R}^t(f_\theta)$, 令 $T\to\infty$ 可得 $\lim_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\cdot\left[\hat{R}^t(f_\theta)-\hat{R}^t\left(f_\theta^*\right)\right]\right)=0$, 因此, 使用 Adam 作为优化器的情况下, 本文的经验风险 $\hat{R}(f_\theta)$ 收敛. 证毕.

2.6 算法伪代码

算法过程主要分为 2 个阶段, 即多模态图提示图构建阶段和模型训练阶段, 本节分别给出了 2 个阶段的算法伪代码, 其中多模态提示图的构建算法伪代码如算法 1 所示.

算法 1. 多模态提示图的构建.

输入: 实例集 D, 标记集 W, 标记数 M;

输出: O.

① for $m=1$ to M do

② 选取实例集中图像 D_m 经过图像编码器获得图像特征 L_i;

③ 选取标记集中标记 W_m 经过文本编码器获得文本特征 T_i;

④ 根据式(1)计算多模态特征图 G_m;

⑤ 将多模态特征图 G_m 经过图卷积和池化后得到节点集 N;

⑥ 根据式(2)计算多模态提示图 O;

⑦ end for

使用算法 1 单独为每一个数据集构建多模态提示图, 该构建过程独立于训练过程之外. 在完成多模态提示图构建后, 使用算法 2 进行模型微调, 具体模型微调过程伪代码如算法 2 所示.

算法 2. 模型微调过程.

输入: 多模态提示图 O, 训练集 Z, 迭代次数 t, $\varepsilon=10\times e^{-5}$;

输出: 模型 f_θ^t.

① while $\left(\hat{R}^t\left(f_\theta^t\right)-\hat{R}^{t-1}\left(f_\theta^{t-1}\right)\right)^2\geqslant\varepsilon$ do

② 选取训练集图像 Z 经过图像编码器得到图像特征 P_i;

③ 根据式(3)计算经验风险 $\hat{R}^t\left(f_\theta^t\right)$;

使用 Adam 更新参数 f_θ^t;

④ end while

3 实 验

3.1 实验设置

为验证本文方法, 在基准数据集 PASCAL VOC 2012(VOC), COCO-80, CUB-200-2011(CUB), NUSW-IDE(NUS)数据集上进行了实验. VOC 数据集共有 17 125 张图像, 20 个标记. COCO 数据集共有 158 957 张图像, 80 个标记. CUB 数据集共有 11 788 张图像, 200 个鸟类标记. NUS 数据集共有 269 648 张图像, 81 个标记. 数据集具体信息如表 1 所示.

Table 1 Benchmark Datasets

表 1 基准数据集

数据集	标记数	总标记数	训练集数目	验证集数目
VOC	20	27 450	5 717	5 823
COCO	80	117 266	118 287	40 670
CUB	200	11 788	5 994	5 794
NUS	81	5 018	161 789	107 859

针对每一个数据集, 在训练集中随机选出 25% 的标记定义为开放词汇, 将含有这些标记的样本全部置为不可见样本来模拟在实际工作中遇到的开放词汇情况. 针对剩余的 75% 的标记, 随机选出 50% 置为未知标记来实现半监督训练.

本文选取 RAM 预训练大模型作为基模型, 输入的图像大小为 384×384, Batch 大小设置为 128, 使用 Adam 优化器, 学习率设置为 0.01, 学习率衰减设置为 0.9, 每个数据集训练 30 轮, 使用平均精度均值(mean average precision, mAP)作为评估标准, 置信度阈值 σ 设置为 0.9.

将本文方法与现有的最先进的半监督多标记学习进行了比较, 包括 MKT[26], ASL[31], CSRA[32], CTran[33]. 其中, MKT 提出联合捕获局部和全局特征的双流特征提取模块, 应用知识迁移引入开放词汇的前置知识, 同时对文本特征使用提示学习, 以适应不同的下游任务. ASL 基于局部损失(focal loss)解决正负样本不平衡问题, 提出了局部损失的改进版, 一种非对称的损失, 即非对称损失(asymmetric loss). CSRA 通过提出一个简单的空间注意力得分来生成每个标记的特定标记特征, 然后将其与标记无关的平均池化特征相结合以解决多标记分类问题. CTran 提出了一个多标记图像分类的通用框架, 使用 Transformer 来利用视觉特征和标记之间的复杂依赖关系, 设定一个标记掩码训练目标来解决训练过程中标记状态为正、负或未知的问题.

为了确保公平性, 本文使用了以上方法发布的最新代码在表 1 中的数据集上进行实验.

3.2 对比实验

在 4 个基准数据集的对比实验如表 2 所示. 从表 2 中可以看出, 在所有基准数据集上, 本文方法均优于现有的最先进的方法, 其中方法中的第 1 行表示强监督下进行训练的结果, 第 2 行表示使用 RAM

大模型进行零次推理得到的结果，第 3 行表示对 RAM 添加线性探针得到的结果. 具体而言，使用本文方法得到的 mAP 在 VOC 数据集上比直接使用 RAM 大模型得到的 mAP 高出约 1.2 个百分点，比 MKT 方法高出约 20 个百分点，比 ASL 方法高出约 10 个百分点，比 CSRA 方法高出约 6 个百分点，比 CTran 方法高出约 4 个百分点. 同时在 COCO, CUB, NUS 数据集上，通过使用基于图提示的微调大模型方法，均具有更好的效果.

Table 2 Experimental Results on Baseline Datasets

表 2 基准数据集实验结果 %

方法	mAP			
	VOC	COCO	CUB	NUS
强监督	93.84	82.79	33.34	59.04
零次学习	90.88	75.71	21.67	41.34
线性探针	90.68	72.76	21.26	41.22
MKT[26]	68.74	62.38	10.61	32.44
ASL[31]	79.50	71.27	10.92	40.99
CSRA[32]	86.21	78.97	20.86	42.87
CTran[33]	88.60	77.89	10.86	43.23
本文方法	92.08	79.71	22.66	44.20

本文认为，性能的提升是由于使用了图神经网络挖掘了开放词汇与监督标记之间的关系，充分探索了不同标记间的相关信息，发挥了图神经网络学习标记间信息的能力，使得模型在面对开放词汇半监督多标记问题时表现出了更优的效果.

3.3 图网络性能评估

为了分析图卷积神经网络的有效性，本节进行了消融实验，消融实验通过去除图卷积神经网络，使用 RAM 直接进行开放词汇半监督多标记学习训练，得到的结果如表 3 所示.

Table 3 Graph Neural Network Ablation Experiments on VOC Dataset

表 3 VOC 数据集上图神经网络消融实验 %

方法	mAP
不使用提示图网络	90.45
使用提示图网络	92.08

从表 3 可以看出，在不使用提示图网络，利用 RAM 预训练大模型对 VOC 数据集进行直接推理时得到的 mAP 为 90.45%. 而在使用提示图网络替换 RAM 大模型文本输入，并对线性层进行训练后，得到的 mAP 提升到了 92.08%.

这个结果说明了多模态提示图在解决开放词汇半监督学习问题时，能够使用图充分挖掘出标记间的联系并保存. 使用保存了标记间关系的多模态提示图输入预训练模型进行训练，能够有效融合图像与文本的多模态信息，提高模型的开放词汇半监督多标记学习能力.

3.4 监督样本比例

同样地，根据已有标记数量的不同，在 20% 已有标记、50% 已有标记、80% 已有标记和完全标记 4 种情况下均设计了消融实验，得到的实验结果如表 4 所示.

Table 4 Ablation Experiments Based on the Proportion of Supervised Samples

表 4 基于监督样本比例的消融实验 %

标记比例	mAP			
	VOC	COCO	CUB	NUS
20	89.47	76.16	21.66	42.90
50	92.08	79.71	22.66	44.20
80	93.01	80.07	23.51	48.72
100	93.84	82.79	33.34	59.04

如表 4 所示，随着有监督样本比例的提高，本文模型在每个基准数据集上的 mAP 同样得到了提高. 同时在仅有 20% 有监督标记的情况下，模型仍能保持较高精度，这证明了本文方法具有鲁棒性.

但在 CUB 数据集及 NUS 数据集上半监督训练的结果要远低于全监督训练的结果. 这可能是由于 CUB 数据集及 NUS 数据集的特殊性，预训练大模型本身在这 2 个数据集上的基础性能不佳，在开放词汇比例为 25% 的情况下，预训练大模型很难做出准确的预测，这样就导致伪标记的 mAP 降低，最终模型的效果变差.

3.5 开放词汇比例与提示个数

为验证提示个数及开放词汇比例对结果的影响，在提示个数为 0, 1, 10, 20 和开放词汇比例为 0, 0.2, 0.4, 0.6, 0.8 的情况下，在 VOC 数据集上进行了消融实验. 得到的实验结果如图 5 所示.

在图 5 中，横坐标表示开放词汇比例，即在整个数据集中完全未标记到的标记占整个数据集的比例. 纵坐标表示 mAP. 在不同的开放词汇比例下，当提示个数从 0 增加至 1 时，模型的表现有较大提升，但当提示个数从 1 增加至 10 及从 10 增加至 20 时，模型的表现变化并不明显. 比如开放词汇比例为 0.6 时，

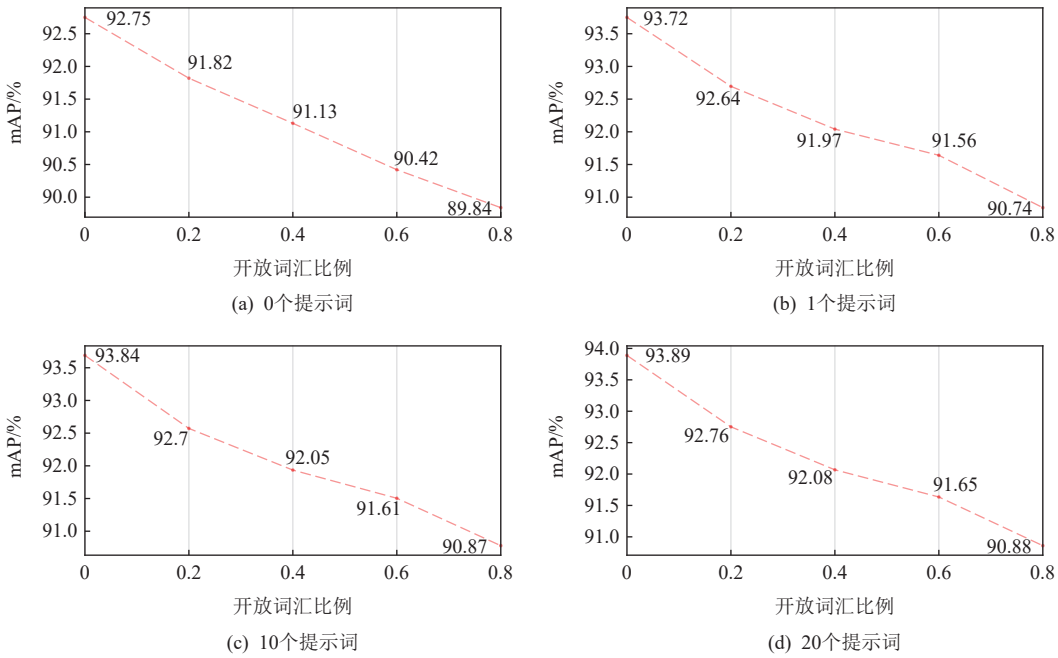


Fig. 5 mAP under the proportion of open vocabulary and number of prompts

图 5 开放词汇比例及提示个数下的 mAP

当图像个数从 0 增加至 1 时，mAP 由 90.42% 提升至 91.56%. 但当图像个数从 1 增加至 20 时，mAP 提升了 0.09 个百分点. 这个结果说明了本文方法对于提示的图像个数是鲁棒的.

3.6 算法复杂度分析

为了验证算法的复杂度，本节在 VOC 数据集和 COCO 数据集上对每个对比方法进行了算法复杂度分析. 得到的实验结果如表 5 所示.

Table 5 Execution Time of Algorithm

表 5 算法执行时间 s

数据集	执行方式	MKT[26]	ASL[31]	CSRA[32]	CTran[33]	本文方法*	本文方法
VOC	训练	54	53	57	50	55	55
	测试	55	53	57	50	55	55
COCO	训练	1 190	1 105	1 259	1 130	1 215	1 168
	测试	51	51	51	51	51	51

注：本文方法*表示未使用多模态提示图时所需训练与测试时间.

在 COCO 和 VOC 数据集上，本文方法在提升了 mAP 的同时并没有占用过多的时间进行训练. 与 CSRA 方法相比，本文方法在提升了 mAP 的同时占用更少的训练时间. 同时本节设计了不使用多模态提示图时进行训练与测试的消融实验，实验结果表明多模态提示图并未占据更多时间.

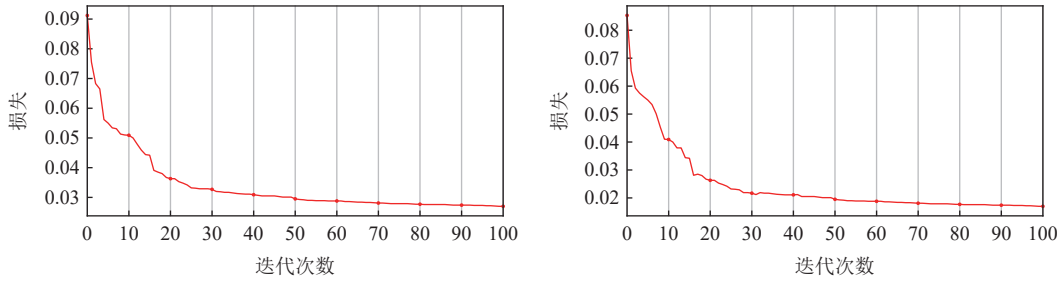
3.7 算法收敛性分析

通过在 VOC 数据集和 COCO 数据集上进行实验，得到的实验结果如图 6 所示. 从图 6 可以看出，在 VOC 数据集上，本文方法在 25 个迭代次数内即完成收敛；在 COCO 数据集上，本文方法在 30 个迭代次数内即完成收敛. 同时在 2 个数据集上，本文方法都在 30 个迭代次数内完成收敛，这也表明了本文提出的算法具有收敛性.

4 结 论

本文提出了一种基于图提示的微调大模型方法，基于提示图神经网络对 RAM 预训练大模型进行微调，使用图神经网络学习开放词汇与有监督样本之间的关系. 同时使用预训练大模型的泛化能力以及使用大模型的输出作为伪标记参与计算无监督样本的损失，使模型在对开放词汇进行分类时能取得更加理想的效果.

在基准数据集上的实验结果表明了本文方法在解决开放词汇多标记问题时具有较好的效果. 此外，



(a) VOC数据集 (b) COCO数据集

Fig. 6 Convergence of algorithm

图 6 算法收敛性

本文还设计了不同的消融实验验证了本文方法具有可信服性. 在未来的工作中, 将考虑对图构造算法进一步优化, 同时研究自适应的伪标记阈值选择方法, 以进一步增强模型的开放词汇分类能力.

作者贡献声明: 李仲年提出研究课题、确定研究方案; 皇甫志宇撰写论文; 杨凯杰负责实验的实施; 营鹏负责调研和整理文献; 孙统风负责实验的指导; 许新征负责论文的修改.

参 考 文 献

[1] Tian Yingjie, Yu Xiaotong, Fu Saiji. Partial label learning: Taxonomy, analysis and outlook[J]. Neural networks, 2023, 161(1): 708−734

[2] Rastogi R, Kumar S. Discriminatory label-specific weights for multi-label learning with missing labels[J]. Neural Processing Letters, 2023, 55(2): 1397−1431

[3] Li Yufeng, Huang Shengjun, Zhou Zhihua. Regularized semi-supervised multi-label learning[J]. Journal of Computer Research and Development, 2012, 49(6): 1272−1278 (in Chinese)
(李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习方法[J]. 计算机研究与发展, 2012, 49(6): 1272−1278)

[4] Lai Zhengfeng, Wang Chao, Cheung S C , et al. SAR: Self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2022: 4091−4100

[5] Zhang Yong, Chen Rongrong, Zhang Jing. Safe tri-training algorithm based on cross entropy[J]. Journal of Computer Research and Development, 2021, 58(1): 60−69 (in Chinese)
(张永, 陈蓉蓉, 张晶. 基于交叉熵的安全 Tri-training 算法[J]. 计算机研究与发展, 2021, 58(1): 60−69)

[6] Fan Yue, Kukleva A, Dai Dengxin, et al. Revisiting consistency regularization for semi-supervised learning[J]. International Journal of Computer Vision, 2023, 131(3): 626−643

[7] Qiu Ye, Gong Xiaolong, Ma Zhiyi, et al. MixLab: An informative semi-supervised method for multi-label classification[C]//Proc of the 9th CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2020: 506−518

[8] Huang Teng, Jia Binbin, Zhang Minling. Progressive label propagation for semi-supervised multi-dimensional classification [C]//Proc of the 32nd Int Joint Conf on Artificial Intelligence. San Francisco, CA: Margan Kaufmann, 2023: 3821−3829

[9] Liu Xuan, Shi Tianyi, Zhou Guohui, et al. Emotion classification for short texts: An improved multi-label method[J]. Humanities and Social Sciences Communications, 2023, 10(1): 1−9

[10] Zhang Shuo, Zhang Jiaojiao, Tian Biao, et al. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation[J]. Medical Image Analysis, 2022, 83(1): 102656−102656

[11] Song Kun, Ma Huimin, Zou Bochao, et al. FD-Align: Feature discrimination alignment for fine-tuning pre-trained models in few-shot learning[C]//Proc of the 37th Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 43579−43592

[12] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[13] Mo S, Kim M, Lee K, et al. S-CLIP: Semi-supervised vision-language learning using few specialist captions[C]//Proc of the 37th Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 61187−61212

[14] Wang Ao, Chen Hui, Lin Zijia, et al. Hierarchical prompt learning using CLIP for multi-label classification with single positive labels[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 5594−5604

[15] Zhang Youcai, Huang Xinyu, Ma Jinyu, et al. Recognize anything: A strong image tagging model[J]. arXiv preprint, arXiv: 2306.03514, 2023

[16] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3): 333−359

[17] Liu Zhunga, Pan Quan, Dezert J, et al. Classifier fusion with contextual reliability evaluation[J]. IEEE Transactions on Cybernetics, 2017, 48(5): 1605−1618

[18] Zhao Xiangyun, Vemulapalli R, Mansfield P A, et al. Contrastive learning for label efficient semantic segmentation[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Los Alamitos, CA: IEEE Computer Society, 2021: 10623−10633

[19] Wang Lichen, Liu Yunyu, Qin Can, et al. Dual relation semi-

supervised multi-label learning[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 6227−6234

[20] Xie Mingkun, Huang Shengjun. Partial multi-label learning with noisy label identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(7): 3676−3687

[21] Duan Yue, Zhao Zhen, Qi Lei, et al. Towards semi-supervised learning with non-random missing labels[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Los Alamitos, CA: IEEE Computer Society, 2023: 16121−16131

[22] Zhou Zhihua. Open-environment machine learning[J]. National Science Review, 2022, 9(8): nwac123

[23] Eichstaedt J C, Kern M L, Yaden D B, et al. Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations[J]. Psychological Methods, 2021, 26(4): 398−398

[24] Gu Xinyue, Lin T Y, Kuo Weicheng, et al. Open-vocabulary object detection via vision and language knowledge distillation[J]. arXiv preprint, arXiv: 2104.13921, 2021

[25] Zhang Hao, Li Feng, Zou Xueyan, et al. A simple framework for open-vocabulary segmentation and detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Los Alamitos, CA: IEEE Computer Society, 2023: 1020−1031

[26] He Sunan, Guo Taian, Dai Tao, et al. Open-vocabulary multi-label classification via multi-modal knowledge transfer[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 808−816

[27] Li Junnan, Li Dongxu, Xiong Caiming, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proc of the 39th Int Conf on Machine Learning. New York: ACM, 2022: 12888−12900

[28] Zhou Kaiyang, Yang Jingkang, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337−2348

[29] Zhou Kaiyang, Yang Jingkang, Loy C C, et al. Conditional prompt learning for vision-language models[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2022: 16816−16825

[30] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014

[31] Ridnik T, Ben-Baruch E, Zamir N, et al. Asymmetric loss for multi-label classification[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2021: 82−91

[32] Zhu Ke, Wu Jianxin. Residual attention: A simple but effective method for multi-label recognition[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2021: 184−193

[33] Lanchantin J, Wang Tianlu, Ordonez V, et al. General multi-label image classification with transformers[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2021: 16478−16488



Li Zhongnian, born in 1990. PhD, lecturer. Member of CCF. His main research interests include data mining, machine learning, and information system.

李仲年, 1990 年生. 博士, 讲师. CCF 会员. 主要研究方向为数据挖掘、机器学习、信息系统.



Huangfu Zhiyu, born in 2000. Master. His main research interests include artificial intelligence and machine learning.

皇甫志宇, 2000 年生. 硕士. 主要研究方向为人工智能、机器学习.



Yang Kaijie, born in 2001. Master. His main research interests include artificial intelligence and machine learning.

杨凯杰, 2001 年生. 硕士. 主要研究方向为人工智能、机器学习.



Ying Peng, born in 2001. PhD. Student member of CCF. His main research interests include artificial intelligence, data mining, and machine learning.

营 鹏, 2001 年生. 博士. CCF 学生会员. 主要研究方向为人工智能、数据挖掘、机器学习.



Sun Tongfeng, born in 1977. PhD, associate professor. Member of CCF. His main research interests include artificial intelligence, machine learning, and data mining.

孙统风, 1977 年生. 博士, 副教授. CCF 会员. 主要研究方向为人工智能、机器学习、数据挖掘.



Xu Xinzheng, born in 1980. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include artificial intelligence, machine learning, and medical image processing.

许新征, 1980 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为人工智能、机器学习、医学图像处理.