

基于多模态一致性协同提示微调的判断标记学习方法

韩日东^{1,3}, 许靖昊¹, 李仲年^{1,3}, 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/人工智能学院, 江苏徐州 221116; 2. 浙江大学计算机辅助设计与图形系统
全国重点实验室, 浙江杭州 310058; 3. 中国矿业大学矿山数字化教育部工程研究中心, 江苏徐州 221116)

摘要: 近期利用大规模未标注数据进行提示学习受到广泛关注。其中, 判断标记学习范式通过从候选集中随机采样类别, 并借助预训练视觉-语言模型进行二值化(True/False)判断, 从而生成高精度的弱监督标记, 大幅提升了标注效率, 推动了提示学习领域的快速发展。然而, 其随机均匀的采样机制不可避免地导致部分类别监督样本缺失, 致使下游任务模型难以构建精确的类别语义空间, 并且现有方法切断了文本与视觉特征之间的跨模态联合优化路径, 破坏了预训练模型原有的跨模态泛化能力。针对上述局限性, 本文提出一种基于多模态一致性协同提示微调的判断标记学习方法(true-false labels learning method based on Multi-modal Consistency Collaborative Prompt tuning, MCCP)。该方法在文本和视觉双模态上分别引入一致性约束, 并设计了跨模态协同优化机制。具体而言, 在文本模态, 设计了大语言模型驱动的知识协同机制, 利用大语言模型为各个类别生成细粒度的文本描述, 通过约束原始文本提示与类别文本描述之间的语义一致性, 补齐监督样本缺失类别的语义信息盲区, 并且隐式缓解监督信息不足的不利影响; 在视觉模态, 提出了基于自监督学习的一致性提示学习机制, 通过构建多视图, 约束原始图像、弱增强视图与强增强视图两两之间的特征分布一致性, 迫使视觉提示网络提取对数据增强操作不敏感的深层语义特征, 显著提升了视觉语义表征的鲁棒性与泛化能力; 在跨模态联合优化方面, 引入了基于训练轮次的余弦动态加权策略, 以实现预训练先验知识与领域适应性学习表征之间的平滑过渡与渐进式融合, 同时采用了分阶段梯度更新策略, 分别优化文本、视觉特定模态的一致性损失与跨模态图像分类的经验风险一致性交叉熵损失, 有效避免了多模态学习中的梯度冲突问题, 促进了跨模态语义空间的深度协同对齐。在 CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101、Stanford Cars、DTD (Describable Textures Dataset)、FGVC-Aircraft (Fine-Grained Visual Classification of Aircraft)、ImageNet-A 和 ImageNet-R 九个多类别图像分类数据集上的实验结果表明, 该方法在利用对比语言-图像预训练模型(Contrastive Language-Image Pre-training, CLIP)标注和人工标注的两种判断标记数据集上, 分别取得了 73.58% 和 74.20% 的最优平均分类准确率, 同时在跨数据集迁移设置下的绝大多数目标数据集上均展现出最优的迁移泛化性能, 验证了该方法的有效性。本文有效解决了判断标记学习范式的监督信号缺失挑战, 并为实现高效、鲁棒的多模态跨域知识迁移提供了方法支持。

关键词: 判断标记; 提示学习; 视觉-语言模型; 弱监督学习; 多模态一致性; 图像分类

基金项目: 国家自然科学基金(No.61976217, No.62306320); 江苏省自然科学基金(No.BK20231063); 浙江大学计算机辅助设计与图形系统全国重点实验室开放课题(No.A2424)

中图分类号: TP181; TP391

文献标识码: A

文章编号: 0372-2112(2026)06-2846-19

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20250727

Learning From True-False Labels Via Multi-Modal Consistency Collaborative Prompt Tuning

HAN Ridong^{1,3}, XU Jinghao¹, LI Zhongnian^{1,3}, SUN Tongfeng^{1,3}, XU Xinzheng^{1,2,3*}

(1. School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China; 2. State Key Laboratory of Computer Aided Design and Computer Graphics (CAD & CG), Zhejiang University, Hangzhou, Zhejiang 310058, China; 3. Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China)

Abstract: Prompt learning utilizing large-scale unlabeled data has recently received widespread attention. Among these, the true-false label learning paradigm randomly samples a category from the candidate set and leverages pre-trained vision-language models to make binary judgments, i.e., True or False, thereby generating high-precision weak supervision signals, substantially improving annotation efficiency, and driving the rapid development of the prompt learning field. How-

ever, its random uniform sampling mechanism inevitably leads to missing supervision samples for certain categories, making it difficult for downstream task models to construct an accurate category semantic space, and existing methods sever the cross-modal joint optimization path between textual and visual features, destroying the original cross-modal generalization capability of pre-trained models. To address the aforementioned limitations, this paper proposes a true-false labels learning method based on multi-modal consistency collaborative prompt tuning (MCCP). This method introduces consistency constraints in both textual and visual modalities, and designs a cross-modal collaborative optimization strategy. Specifically, in the textual modality, a large language model-driven knowledge collaboration mechanism is designed to generate fine-grained textual descriptions for each category. By constraining the semantic consistency between original textual prompts and category textual descriptions, it fills the semantic information blind spots of categories with missing supervision samples, and implicitly mitigates the effects of insufficient supervision information. In the visual modality, a consistency prompt learning mechanism based on self-supervised learning is proposed. By constructing multiple views, it constrains the pairwise feature distribution consistency among original images, weakly augmented views, and strongly augmented views, thereby forcing the visual prompt network to extract deep semantic features insensitive to data augmentation operations, significantly enhancing the robustness and generalization of visual-semantic representations. In terms of cross-modal joint optimization, an epoch-based cosine dynamic weighting strategy is introduced to achieve a smooth transition and gradual fusion between pre-trained prior knowledge and domain-adaptive feature representations, while a staged gradient update scheme is adopted to separately optimize the consistency losses of text- and vision-specific modalities and the empirical risk-consistency cross-entropy loss of cross-modal image classification, effectively avoiding the gradient conflict problem in multi-modal learning and promoting the deep collaborative alignment of cross-modal semantic spaces. Experimental results on nine multi-class image classification datasets, including CIFAR-100, Tiny-ImageNet, Caltech-101, Food-101, Stanford Cars, describable textures dataset (DTD), fine-grained visual classification of aircraft (FGVC-Aircraft), ImageNet-A and ImageNet-R, demonstrate that this method achieves state-of-the-art average classification accuracies of 73.58% and 74.20% on the True-False labels datasets annotated by the contrastive language-image pre-training model (CLIP) and human annotators, respectively, and simultaneously exhibits optimal transfer generalization performance on the vast majority of target datasets under the cross-dataset transfer setting, verifying the effectiveness of MCCP. This paper effectively addresses the missing supervision signal challenge of true-false labels learning paradigm, providing methodological support for achieving efficient and robust multi-modal cross-domain knowledge transfer.

Keywords: true-false labels; prompt learning; vision-language models; weakly supervised learning; multi-modal consistency; image classification

Foundation Item(s): National Natural Science Foundation of China (No.61976217, No.62306320); Natural Science Foundation of Jiangsu Province (No.BK20231063); Open Project Program of the State Key Laboratory of CAD&CG of Zhejiang University (No.A2424)

0 引言

近期研究表明,预训练视觉-语言模型(Vision-Language Models, VLMs)通过海量数据的预训练实现了跨模态特征的精确对齐,展现了强大的零样本(zero-shot)和小样本(few-shot)学习能力^[1-4]。然而,这些多模态模型在适应下游任务时仍面临一定挑战。传统全参数微调(full fine-tuning)方法虽然能够提升模型性能,但其高昂的计算成本与庞大的存储需求在实际应用中往往难以承受^[5]。在此背景下,参数高效迁移学习(如提示学习^[6-9]与适配器微调^[10])通过冻结预训练权重、单独优化轻量级附加模块的策略,成为适配下游任务的主流方法。其中,提示学习方法通过在输入空间引入可学习的提示向量(如文本侧上下文优化^[6]、视觉模态嵌入增强^[7]或双模态协同提示^[8]),将下游任务的语义隐式映射至预训练语义空

间。然而,现有工作表明,提升VLMs模型在下游各类任务上的适应能力,仍然需要大量标注数据进行微调^[11]。由于高质量标注数据难以获取,此类微调需求将导致高昂的标注成本。为应对这一挑战,已有研究开始探索使用海量无标注数据进行提示学习的方法^[11-13],旨在深度挖掘VLMs模型固有的零样本能力,从而显著减少VLMs模型适应下游任务时对数据标注的依赖。然而,现有方法的性能高度依赖于VLMs模型生成的硬伪标记(hard pseudo-labels)的准确性。若VLMs模型在特定下游任务中零样本能力欠佳,则提示学习模型在该任务上的性能表现将显著下降^[11,14]。该现象促使研究者开始探索判断标记(True-False Labels, TFL)^[13]范式,即在缓解VLMs模型零样本标注噪声的同时,结合提示学习实现参数高效知识迁移。其核心思想是:从候选标记集中随机采样

得到一个类别标记,标注者仅需判断待标注样本是否属于该随机类别,即可完成标注过程。图1展示了硬伪标记和判断标记的对比示意图。

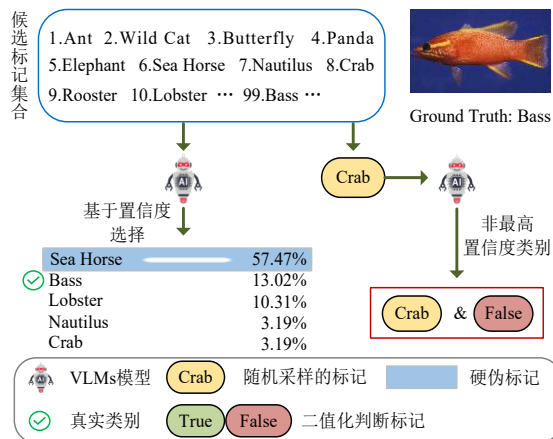


图1 硬伪标记与判断标记的对比示意图

Figure 1 Comparison illustration of hard pseudo-labels and True-False labels

当样本真实类别为“Bass”时,硬伪标记由VLMs基于置信度选择,可能得到错误标记“Sea Horse”;而在判断标记范式中,首先随机采样一个类别标记,若该类别标记恰为“Bass”,则标注为“Bass & True”;若该类别标记为“Crab”,则标注为“Crab & False”。判断标记范式具有双重优势:既能够有效利用VLMs模型生成高质量的标注数据,又能避免人工遍历完整的候选标记集,从而大幅提升标注效率。判断标记范式标注的各数据集上监督样本的分布信息见表1。

表1 判断标记范式标注的各数据集上监督样本的分布信息

Table 1 Distribution information of supervised samples across different datasets annotated by True-False labels paradigm

数据集	类别总数	监督样本缺失的类别数目	True标记样本数目
CIFAR-100	100	2	498
Tiny-ImageNet	200	10	507
Caltech-101	102	55	65
Stanford Cars	196	156	42

然而,如表1所示,在判断标记范式^[13]中,随机均匀的采样机制可能导致部分类别的监督样本缺失,而现有方法^[13]并未针对这一问题设计相应的补偿策略。这可能导致下游任务的分类模型难以构建精确的类别语义空间,从而削弱模型的判别能力并影响整体性能。此外,文献^[13]虽通过引入线性分类器简化了训练流程并在一定程度上提升了下游任务的分类准确率,但却切断了视觉-文本嵌入空间的联合优化路径,缺乏跨模态的联合优化,进而导致VLMs模

型原有的跨模态泛化能力部分丧失。因此,本文提出了一种基于多模态一致性协同提示微调的判断标记学习方法(true-false labels learning method based on Multi-modal Consistency Collaborative Prompt tuning, MCCP)。具体地,在文本模态,提出了大语言模型(Large Language Models, LLMs)驱动的知识协同机制,利用大语言模型为各类别生成细粒度文本描述,并结合文本提示网络与文本适配器,约束原始文本提示与生成的文本描述之间的语义一致性,实现对对比语言-图像预训练模型(Contrastive Language-Image Pre-training, CLIP)的文本提示(类别名称)的动态优化,从而有效缓解部分类别监督样本缺失的影响;在视觉模态,设计了基于自监督学习的一致性提示学习机制,通过计算原始图像、弱增强视图、强增强视图两两之间的特征分布一致性约束,优化视觉提示网络和视觉适配器,从而提升视觉语义表征的鲁棒性与泛化能力;在跨模态联合优化方面,通过引入余弦动态加权策略重构了判断标记学习框架^[13]中的条件概率估计机制,实现预训练先验知识与领域适应性表征之间的交互平衡,同时采用分阶段梯度更新策略,优化多类别图像分类任务的经验风险一致性交叉熵损失,以实现跨模态语义空间的协同对齐。

鉴于本文与文献^[13]高度相关,本节将两者之间的联系和区别进行了总结。联系在于:本文继承了文献^[13]的基础设定,包括判断标记学习范式及其一致性风险估计方法。区别在于以下三点:

(1)跨模态优化路径的连贯性。文献^[13]在得到文本与视觉嵌入表征后,引入了一个可训练的线性分类器进行最终预测,即文本与视觉嵌入特征简单拼接作为分类器输入。这种设计实质上切断了视觉与文本嵌入空间之间的跨模态联合优化路径,破坏了CLIP模型在预训练阶段建立的跨模态语义对齐能力,导致其泛化性受损。而本文保留了CLIP模型原有的基于余弦相似度的分类机制,采用文本与视觉双分支的适配器与提示网络协同微调,做到了跨模态语义空间的联合协同优化。

(2)对“监督信号缺失”的补偿机制:判断标记的随机均匀采样机制不可避免地会导致部分类别的监督样本缺失(如表1所示)。文献^[13]未设计任何机制来解决这一问题,可能导致模型难以准确构建完整的类别语义空间,而本文引入了大语言模型驱动的知识协同机制,通过让大语言模型生成细粒度的类别文本描述,为监督信号缺失的类别注入了丰富的先验语义信息。这种类别语义补偿机制,有助于模型准确构建完整的类别语义空间,能够有效弥补监督样本缺失导致的学习盲区。

(3)视觉嵌入表征提取的鲁棒性。文献[13]仅利用基于卷积神经网络的提示网络简单地提取视觉嵌入表征,在判断标记的极度稀疏正样本(即标注为“True”的样本)监督下,容易发生拟合,影响嵌入表征的鲁棒性。而本文引入了基于自监督学习的一致性提示学习机制,通过约束原始图像、弱增强视图与强增强视图嵌入表征之间的一致性,促使模型提取“不敏感于随机增强操作”的深层语义特征,能够有效提升视觉模态语义表征的鲁棒性。

为了验证提出的MCCP方法的有效性,本文将九个多类别图像分类数据集,包括CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101、Stanford Cars、DTD (Describable Textures Dataset)、FGVC-Aircraft (Fine-Grained Visual Classification of Aircraft)、ImageNet-A 和 ImageNet-R,依照判断标记学习范式重新标注,并开展了系统的对比实验、跨数据集迁移实验以及消融实验分析。实验结果表明,MCCP方法在CLIP模型标注和人工标注的数据集上均取得了最优平均分类准确率,分别为73.58%和74.20%;在跨数据集迁移设置下,几乎所有目标数据集上均获得最优性能,验证了MCCP方法出色的迁移能力;此外,消融实验进一步验证了各组成模块的有效性。综上,本文贡献可以概括为四个方面:

(1)在文本模态,提出了一种大语言模型驱动的知识协同机制,通过为各个类别生成文本描述,并约束原始提示与类别文本描述间的语义一致性,补充监督样本缺失类别的语义信息,从而缓解样本缺失的影响。

(2)在视觉模态,设计了一种基于自监督学习的一致性提示学习机制,通过约束原始图像、弱增强视图和强增强视图两两之间的特征分布一致性,提升视觉语义表征的鲁棒性与泛化能力。

(3)在跨模态联合优化方面,提出了基于余弦动态加权的条件概率估计机制,以动态融合预训练知识与领域适应性表征,并采用了分阶段梯度更新策略优化经验风险一致性损失,以有效对齐跨模态语义空间。

(4)在九个基准数据集上的实验结果表明,本文方法在多种设置下取得了优异的性能表现,不仅提升了模型对下游任务的适应性,而且较好保留了VLMs模型的泛化能力。

1 相关工作

1.1 面向视觉分类任务的提示学习方法

基于视觉-语言预训练模型的提示学习通过引入任务特定的引导信息,在保持预训练知识完整性的同

时,实现了对下游任务的参数高效迁移。该研究领域近年来取得了显著突破,尤其在图像-文本匹配模型上,形成了一套系统化的方法^[6-8,15-16]。此外,以CLIP模型^[1]为代表的对比学习模型,通过构建跨模态的语义对齐空间,为提示学习框架提供了强大的基础模型。具体地,CoOp模型^[6]创新性地将连续可学习的上下文向量引入VLMs,通过端到端地优化提示“a photo of a [CLS].”中的上下文词嵌入,显著提升了细粒度分类任务的性能。CoCoOp模型^[7]进一步提出了实例条件化提示机制,通过轻量级元网络动态生成与图像内容相关的提示向量。DCLIP模型^[17]通过大语言模型生成细粒度类别描述符,将图像-文本相似度计算扩展为多概念集成。CuPL模型^[18]系统化地利用大语言模型生成类别属性描述进行提示集成。TCP^[19]通过文本知识嵌入,将类别文本知识映射为动态可调的类感知提示符,有效增强了视觉语言模型在未知领域中的类别区分能力。Xie等人^[20]提出了一种即插即用的TextRefiner方法,通过构建局部缓存模块提取图像的局部细粒度视觉概念,并聚合对齐文本分支的输出,从而无需外部知识即可高效地优化现有方法的文本提示。VPT模型^[9]在Vision Transformer^[21]的输入层中添加可学习图像块,凭借不到1%的可训练参数实现了视觉特征的领域自适应。MaPLe模型^[8]则在文本和视觉侧协同优化,通过视觉-语言耦合函数,实现了跨模态表征对齐。除此之外,提示学习与弱监督学习的深度融合,正在开辟无监督提示学习的可能性。

1.2 弱监督标注范式

弱监督学习范式通过利用不完整、噪声或粗粒度标注数据来训练模型,显著降低了模型对精细标注数据的依赖^[22]。其核心在于标注形式革新与学习理论优化的深度融合,其中新型弱监督标注范式(如偏标记学习、补充标记学习)已成为近年来的研究焦点。弱监督标注范式通过重新定义监督信号的形式,降低了标注成本并提升了模型的适应性,主要包括偏标记学习^[23-27]、补充标记学习^[28-31]等。具体地,偏标记学习是指每个样本对应多个候选标记,其中有且仅有一个为真实类别标记。例如,一个图像被标注为候选标记集合{“猫”,“狗”,“老虎”},而真实标记为“老虎”。根据候选标记处理策略,偏标记学习方法由基于平均策略^[23,27]逐渐过渡到了基于识别策略^[24-26]。补充标记学习为每个样本标注一个“非真实类别”(如“非猫”),通过排除法隐式约束类别标签空间。例如,Ishida等人^[28]提出了一种补充标记学习方法,通过特定条件来获得无偏分类风险估计。Yu等人^[31]研究了存在偏差的补充标记情形。Ishida等人^[29]提出了一种

能够处理任意损失函数和模型的补充标记学习框架,通过无偏风险估计量进行模型学习。Gao 等人^[30]则提出了一种基于判别模型的补充标记学习方法,通过加权损失来最大化潜在真实类别标记和补充标记之间的预测差距。此外,针对如何进一步提升标注效率,也有研究探索了基于深度主动学习的半自动标注策略^[32]。

弱监督学习范式显著降低了数据标注成本。然而,如何在基于 VLMs 模型的提示学习方法中利用弱监督数据实现高效的下游任务适配,目前仍鲜有探索。本文针对这种情形,探索如何利用弱监督数据进行提示学习。

2 预备知识

2.1 多类别分类任务的分类风险

在一个多类别分类任务中,假设 $\mathbf{X} \in \mathbb{R}^d$ 为 d 维特征空间, $\mathbf{Y} = \{y_1, y_2, \dots, y_k\}$ 为类别标记空间,其中 $k > 2$ 表示类别标记数目。给定数据集 $D = \{(x_l, y_l)\} (l = 1, 2, \dots, N)$, 其中 $x_l \in \mathbf{X}$, $y_l \in \mathbf{Y}$, N 表示训练样本数目。假设所有样本 $\{(x_l, y_l)\} (l = 1, 2, \dots, N)$ 服从一个概率分布 $p(x, y)$, 传统监督学习范式的训练目标是利用多类别损失 $\mathcal{L}(f(x), y)$ 寻找最优分类器 $f(x): x \rightarrow \{y_1, y_2, \dots, y_k\}$ 以最小化下述分类风险:

$$\begin{aligned} R(f) &= \mathbb{E}_{p(x, y)} \mathcal{L}(f(x), y) \\ &= \mathbb{E}_{x \sim p(x, y)} \sum_{i=1}^k p(y = y_i | x) \mathcal{L}(f(x), y_i) \end{aligned} \quad (1)$$

其中, $\mathbb{E}$ 表示数学期望。

2.2 判断标记学习范式的一致性风险估计

在判断标记学习范式^[13]中,每个样本不再使用传统类别标记 y , 而是使用一个判断标记 $\hat{y}$ 标注。假设判断标记训练数据集 $D_{TF} = \{(x_l, \hat{y}_l)\} (l = 1, 2, \dots, N)$ 是从概率分布 $p(x, \hat{y})$ 中随机均匀采样得到的, 其中, $\hat{y}_l = (\bar{y}_l, s_l)$ 为二元组形式的判断标记; $\bar{y}_l \in \mathbf{Y}$ 表示随机采样的类别标记; $s_l \in \{0, 1\}$ 为二值化判断标记, 表示样本 x_l 是否属于标记 $\bar{y}_l$ 所指代的类别。 $s_l = 0$ 时表示样本 x_l 不属于 $\bar{y}_l$ 类别(图 1 中标注为“False”的情况); $s_l = 1$ 时表示样本 x_l 属于 $\bar{y}_l$ 类别(图 1 中标注为“True”的情况)。若一个类别没有 $s_l = 1$ 的样本, 则认为它是监督样本缺失的类别。与传统监督学习类似, 判断标记学习范式的训练目标是学习一个能够预测未知样本类别的分类器 $f(x): x \rightarrow \{y_1, y_2, \dots, y_k\}$ 。判断标记范式的一致性分类风险可表示为

$$R_{TF}(f) = \mathbb{E}_{p(x, \bar{y}, s=0)} \bar{\mathcal{L}}(f(x), \bar{y}) + \mathbb{E}_{p(x, \bar{y}, s=1)} \mathcal{L}(f(x), \bar{y}) \quad (2)$$

其中, $\mathcal{L}(f(x), \bar{y})$ 表示上述 2.1 节所述的多类别损失,

$\bar{\mathcal{L}}(f(x), \bar{y}) = \sum_{i=1, i \neq j}^k p(y = y_i | \bar{y} = y_j, s = 0, x) \mathcal{L}(f(x), y_i)$, y_j 是随机采样的类别标记。进一步地, 判断标记学习范式的经验风险估计量可表示为

$$\hat{R}_{TF}(f) = \frac{1}{N_F} \sum_{l=1}^{N_F} \bar{\mathcal{L}}(f(x_l), \bar{y}_l) + \frac{1}{N_T} \sum_{l=1}^{N_T} \mathcal{L}(f(x_l), \bar{y}_l) \quad (3)$$

其中, N_F 和 N_T 分别表示 $s = 0$ 和 $s = 1$ 的样本数量。通过最小化式(3)中的经验风险估计量, 可训练得到判断标记范式的分类器 $f(x): x \rightarrow \{y_1, y_2, \dots, y_k\}$ 。

3 基于多模态一致性协同提示微调的判断标记学习

在判断标记学习框架^[13]中, 随机均匀的采样机制可能导致长尾类别的训练样本缺失。为此, 本文提出了一种基于多模态一致性协同提示微调的判断标记学习方法, 整体框架如图 2 所示。首先, 在文本和视觉两个模态分别引入一致性协同提示学习机制, 以补充样本缺失类别的语义信息(文本模态), 并增强视觉语义表征的鲁棒性与泛化能力(视觉模态); 其次, 提出一种动态加权的条件概率估计机制, 通过余弦相似度驱动的动态加权策略, 以实现预训练模型的先验知识与待学习表征的渐进式融合; 最后, 设计了跨模态一致性损失函数进行跨模态联合优化, 以实现更有效的特征对齐。为便于表述, 本文后续采用 CLIP 模型^[1]作为预训练视觉-语言模型的代表。

上述方法能够有效缓解判断标记范式导致的监督样本缺失问题, 其核心逻辑在于: 深度结合“外部知识注入”与“多模态一致性对齐”, 为监督样本缺失类别补充细粒度类别语义并构建鲁棒特征, 具体地:

(1) 文本模态的 LLM 知识补偿: 利用大语言模型为每个类别生成细粒度的类别文本描述, 并基于一致性提示学习机制强制文本提示网络向这些全局语义表征靠拢, 通过来自大语言模型的富文本语义先验, 填补因监督信号缺失导致的类别语义盲区, 助力模型准确划定分类决策边界。

(2) 视觉模态的自监督一致性: 当监督信号缺失时, 视觉特征提取网络极易对少量样本产生过拟合, 难以获得鲁棒的视觉表征, 通过视觉模态多视图一致性提示学习机制, 可以迫使模型学习到“对数据增强操作不敏感”的鲁棒特征, 防止模式崩塌。

(3) 跨模态联合: 通过跨模态协同优化, 使具备完整类别语义的文本特征去引导视觉特征的对齐, 由于文本侧已经通过大语言模型补齐了样本缺失类别的语义, 视觉网络便能在测试阶段凭借这种跨模态对齐

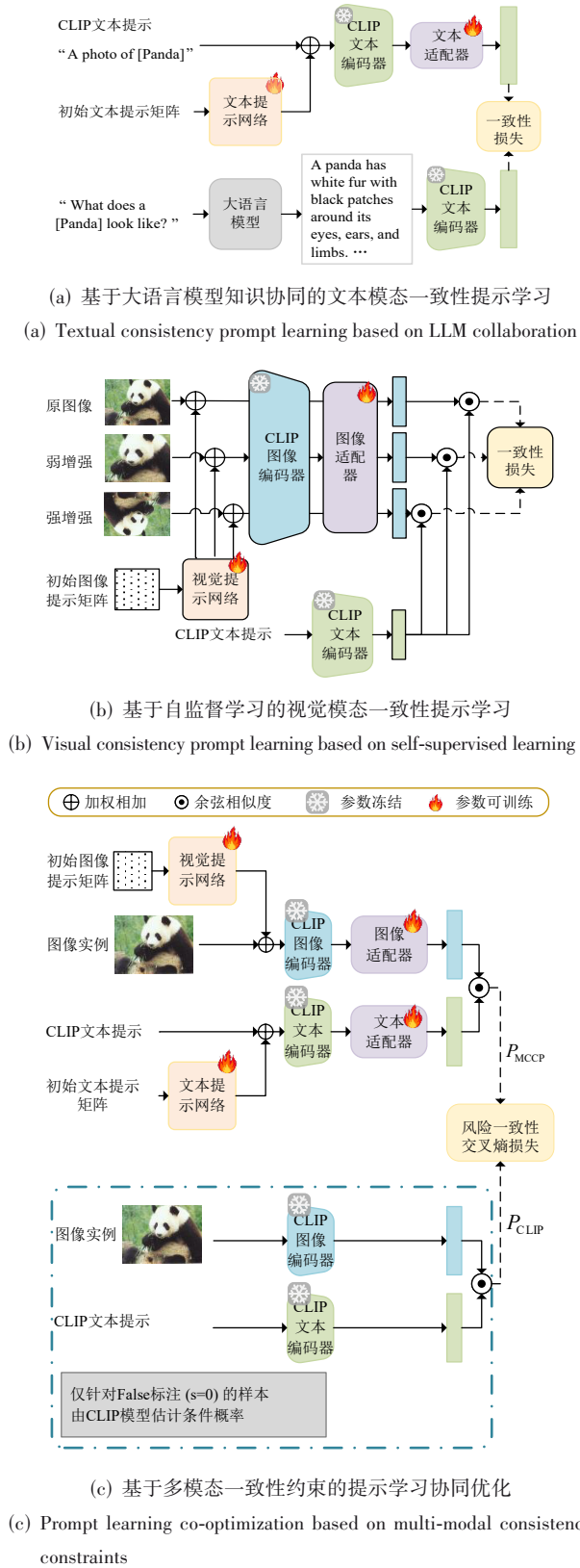


图2 基于多模态一致性协同提示微调的判断标记学习方法架构图
Figure 2 Architecture overview for learning from True-False labels via multi-modal consistency collaborative prompt tuning

能力,实现对未见过样本类别的有效识别。

3.1 基于大语言模型知识协同的文本模态一致性提示学习

判断标记学习中的随机均匀采样机制可能导致部分长尾类别缺失监督信号,为了给这些类别提供额外的语义信息,本节提出一种基于大语言模型知识协同的文本模态提示学习方法,其利用大语言模型作为外部知识源,为每个类别生成细粒度的类别描述文本,并通过文本模态的一致性约束损失优化“文本提示网络”和“文本适配器”,以生成融合类别语义信息的文本提示嵌入表征,动态补充监督样本缺失类别的语义信息。具体流程如图2(a)所示。

(1) 类别语义编码阶段

针对类别集合中的每个类别,构建结构化提示模板“*What does a [CLS] look like*”,其中[CLS]表示类别名称,并通过大语言模型生成细粒度的类别描述。受益于大语言模型在大规模语料上的预训练,其生成的文本描述不仅涵盖形态特征(如“斑马具有黑白条纹”),还包含场景关联信息(如“常出现在非洲草原”)等多维度语义内容,从而可以有效弥补监督样本缺失类别的语义知识空白。随后,将上述细粒度类别描述输入至CLIP文本编码器中,并通过平均池化(average pooling)操作获得该类别的语义嵌入表征。为了控制大语言模型LLMs的生成质量,采用了下述三种策略:多次生成、文本过滤、特征级去噪。

具体来说,设 $Y=\{y_1, y_2, \dots, y_k\}$ 为类别标记空间, $k>2$ 表示类别数目, $\{c_i\}(i=1, 2, \dots, k)$ 为类别名称集合,首先,对于每个类别名称 c_i ,向大语言模型进行多次提问“*What does a c_i look like*”来生成该类别的多个细粒度的类别文本描述,具体示例如表2所示。然后,针对上述类别描述进行文本过滤,一方面,过滤掉包含大语言模型常见拒绝响应词(如“*As an AI*”, “*I cannot*”, “*Sorry*”等)的无效输出;另一方面,考虑到CLIP模型文本编码器对输入文本的最大长度限制,设定合理的文本长度上下限阈值(如10~50词),剔除过短或过长的类别描述,最终保留 M 个类别文本描述。最后,针对保留的 M 个细粒度类别文本描述,进行特征级去噪,将这 M 个类别文本描述送入到CLIP模型文本编码器中,并计算特征平均值得到该类别的语义嵌入表征 q_{llm}^i :

$$q_{llm}^i = \frac{\sum_{j=1}^M g_T(\text{LLM}(\mathcal{Q}_j, c_i))}{M} \quad (4)$$

其中, $\mathcal{Q}_j$ 表示第 j 次提问的提示, c_i 表示第 i 个类别标记的名称,LLM($\cdot$)表示大语言模型, $g_T(\cdot)$ 表示CLIP模型的文本编码器。这种特征级别的平均池化(aver-

表2 大语言模型生成细粒度类别文本描述示例

Table 2 Examples of fine-grained text descriptions generated by large language models

Prompt 提示	What does a bear look like
文本描述 示例	1. Most bears are distinguished by their round faces, short snouts, small pointed ears, stocky bodies, fur-covered tails, and nonretractile claws.
	2. The identifying characteristics of a bear are that it is a large mammal with four legs and short, sharp claws.
	3. A bear typically has a large body, short legs, and a long snout.

age pooling)操作能够有效中和单次生成中偶然出现的离群语义噪声,起到了平滑和隐式去噪的作用。值得注意的是,此处大语言模型的多次生成以及文本过滤属于一次性离线预处理步骤,仅在数据准备阶段产生少量固定的时间开销,不会在后续模型的训练和推理阶段增加额外计算开销。

(2) 文本提示学习阶段

采用标准差为 0.02 的正态分布随机初始化一个初始文本提示矩阵 $V_T \in \mathbb{R}^{k \times d_c}$, 其中 k 为类别数目, d_c 表示嵌入向量维度, 将 V_T 输入到文本提示网络 $\phi_T(\cdot)$ 进行编码, 并将结果与 CLIP 模型原始文本提示(即“a photo of a [CLS].”)的嵌入表征加权相加, 最后输入到 CLIP 模型的文本编码器 $g_T(\cdot)$ 和可训练的文本适配器 $\text{adapter}_T(\cdot)$ 中, 以动态生成文本提示的嵌入表征 q_T^i :

$$q_T^i = \text{adapter}_T(g_T(\phi_T(V_T) \oplus Q_{\text{CLIP}}^i)) \quad (5)$$

其中, Q_{CLIP} 表示 CLIP 模型原始文本提示(即“a photo of a [CLS].”)的嵌入表征, $\oplus$ 表示向量加权相加操作。通过文本适配器 $\text{adapter}_T(\cdot)$ 的非线性映射, 来实现文本提示矩阵 V_T 与 CLIP 模型语义空间的适应性对齐。

本文遵循文献[10]的设定, 文本适配器 $\text{adapter}_T(\cdot)$ 采用两阶段线性映射结构。而文本提示网络 $\phi_T(\cdot)$ 由四部分构成: (i) 多头潜在注意力层: 通过低秩分解与位置编码融合技术, 在保留传统自注意力全局建模能力的同时, 显著降低计算复杂度; (ii) Dropout 层: 以 0.3 的概率随机屏蔽神经元, 增强模型的鲁棒性; (iii) 归一化层: 用于稳定特征分布, 提高模型收敛速度; (iv) 残差连接: 缓解梯度消失问题, 提高网络训练的稳定性。相较于直接优化文本提示矩阵 V_T , 引入注意力机制有助于充分建模向量序列内部的依赖关系。为保证后续对比实验的公平性, 上述文本提示网络采用类似 Transformer 编码器的结构, 并将多头自注意力机制替换为 DeepSeek^[33] 中的多头潜在注意力 (Multi-head Latent Attention, MLA), 以减少参数量。

(3) 一致性约束优化阶段

为了保证文本提示嵌入表征 q_T^i 与类别语义嵌入表征 q_{ilm}^i 之间的语义一致性, 提出文本模态的一致性

损失:

$$\mathcal{L}_T = \frac{1}{k} \sum_{i=1}^k (q_T^i - q_{\text{ilm}}^i)^2 \quad (6)$$

该约束损失能够确保在上述基于 LLMs 的类别语义信息注入过程中, 保持与 CLIP 模型预训练空间的语义一致性, 从而避免因领域偏移导致的性能退化。

3.2 基于自监督学习的视觉模态一致性提示学习

鉴于判断标记数据集中 True 标记样本(即 $s=1$) 数量稀少(如表 1 所示), 常规模型难以依赖如此稀疏的监督信号学习到鲁棒的特征嵌入表征。为了防止视觉提示过拟合于特定的噪声模式, 并提升模型对鲁棒视觉嵌入表征的学习能力, 本节在视觉模态中提出了一种基于自监督学习的视觉一致性正则化提示学习方法。该方法通过对原始图像、弱增强视图(如翻转、裁剪)和强增强视图(如色彩扰动)构建多视图, 并利用余弦相似度计算分类 logit 分数的一致性, 得到视觉模态一致性约束损失, 迫使视觉提示网络学习到“对随机增强操作不敏感”的深层、鲁棒语义特征, 从而避免了特征塌陷或学习无意义噪声。具体流程如图 2(b) 所示。

首先, 采用标准差为 0.02 的正态分布随机初始化一个初始图像提示矩阵 $V_I \in \mathbb{R}^{H \times W \times C}$, 其中, H 、 W 、 C 分别表示图像的高度、宽度和通道数, 并将图像提示矩阵 V_I 输入到可训练的视觉提示网络 $\phi_I(\cdot)$ 中, 以生成适应特定任务的视觉提示嵌入表征:

$$P_V = \phi_I(V_I) \quad (7)$$

然后, 对于一个图像实例 x , 利用随机数据增强技术得到该图像的强增强视图 x' 和弱增强视图 x'' 。随机数据增强的具体操作细节如表 3 所示。

表3 数据增强细节

Table 3 Details on data augmentation

增强视图	具体增强操作
弱增强	随机水平翻转→反射填充随机裁剪→尺寸归一化→标准化
强增强	弱增强流程→随机色彩扰动→矩形遮挡

弱增强操作包括随机水平翻转、反射填充随机裁剪、尺寸归一化和标准化。强增强操作在弱增强流程的基础上增加了随机色彩扰动和矩形遮挡两种操作, 以增强图像样本的多样性。

将原始图像和上述两种增强视图分别与视觉嵌入表征 $\mathbf{P}_V$ 加权相加,作为 CLIP 模型图像编码器 $\mathbf{g}_I(\cdot)$ 和视觉适配器 $\text{adapter}_V(\cdot)$ 的输入,得到原始图像嵌入表征 $\mathbf{q}_I$ 、强增强图像嵌入表征 $\mathbf{q}_{\text{strong}}$ 和弱增强图像嵌入表征 $\mathbf{q}_{\text{weak}}$:

$$\begin{aligned}\mathbf{q}_I &= \text{adapter}_V(\mathbf{g}_I(\mathbf{P}_V \oplus \mathbf{x})) \\ \mathbf{q}_{\text{strong}} &= \text{adapter}_V(\mathbf{g}_I(\mathbf{P}_V \oplus \mathbf{x}')) \\ \mathbf{q}_{\text{weak}} &= \text{adapter}_V(\mathbf{g}_I(\mathbf{P}_V \oplus \mathbf{x}''))\end{aligned}\quad (8)$$

其中, $\oplus$ 表示图像加权相加操作。视觉适配器 $\text{adapter}_V(\cdot)$ 的网络结构遵循文献[10]中的设定,采用两阶段线性映射结构。视觉提示网络 $\phi_I(\cdot)$ 与上一小节中的文本提示网络 $\phi_T(\cdot)$ 结构相同,但是其参数独立化设置,以适配各模态数据的特异性特征。

接下来,将 CLIP 模型的原始文本提示(即“a photo of a [CLS].”)输入到 CLIP 模型的文本编码器 $\mathbf{g}_T(\cdot)$ 中,得到类别文本提示的嵌入表征,并将该嵌入表征分别与原始图像嵌入表征 $\mathbf{q}_I$ 、强增强图像嵌入表征 $\mathbf{q}_{\text{strong}}$ 和弱增强图像嵌入表征 $\mathbf{q}_{\text{weak}}$ 进行矩阵相乘操作,得到下述原始图像分类 logit 分数 $\mathbf{S}_o$ 、强增强视图分类 logit 分数 $\mathbf{S}_s$ 和弱增强视图分类 logit 分数 $\mathbf{S}_w$:

$$\begin{aligned}\mathbf{S}_o &= \frac{\mathbf{q}_I \otimes (\mathbf{g}_T(\mathbf{Q}_{\text{CLIP}}))^T}{\tau} \\ \mathbf{S}_s &= \frac{\mathbf{q}_{\text{strong}} \otimes (\mathbf{g}_T(\mathbf{Q}_{\text{CLIP}}))^T}{\tau} \\ \mathbf{S}_w &= \frac{\mathbf{q}_{\text{weak}} \otimes (\mathbf{g}_T(\mathbf{Q}_{\text{CLIP}}))^T}{\tau}\end{aligned}\quad (9)$$

其中, $\mathbf{Q}_{\text{CLIP}}$ 表示 CLIP 模型的原始文本提示(即“a photo of a [CLS].”)的嵌入表征, $\otimes$ 表示矩阵相乘操作, T 表示矩阵转置操作, τ 为 CLIP 模型的温度超参数。

最后,对原始图像分类 logit 分数 $\mathbf{S}_o$ 、强增强视图分类 logit 分数 $\mathbf{S}_s$ 和弱增强视图分类 logit 分数 $\mathbf{S}_w$ 进行两两余弦相似度计算,并取平均值,以构建下述视觉模态一致性损失函数,促使 CLIP 模型提取“不敏感于随机增强操作”的深层语义特征:

$$\mathcal{L}_V = 1 - \frac{\cos(\mathbf{S}_o, \mathbf{S}_s) + \cos(\mathbf{S}_o, \mathbf{S}_w) + \cos(\mathbf{S}_s, \mathbf{S}_w)}{3} \quad (10)$$

其中, $\cos(\cdot, \cdot)$ 表示余弦相似度函数。

3.3 基于多模态一致性约束的提示学习协同优化

针对上述文本模态和视觉模态提示学习方法之间的跨模态空间解耦问题^[13],本节提出了基于多模态一致性约束的提示学习协同优化框架。如图 2(c) 所示,该方法在保留 CLIP 模型双塔架构的基础上,通过文本、视觉双模态上提示网络与适配器的协同优

化,实现适配下游任务与保留 CLIP 模型预训练知识之间的平衡。

为了利用 CLIP 模型完成图像分类任务,本节结合式(5)中的文本提示嵌入表征 $\mathbf{q}_T$ 与式(8)中的原始图像嵌入表征 $\mathbf{q}_I$,通过文本与视觉嵌入表征之间的跨模态交互,得到所提出 MCCP 方法的最终预测分类概率:

$$P_{\text{MCCP}} = \text{softmax}(\cos(\mathbf{q}_I, \mathbf{q}_T)/\tau) \quad (11)$$

其中, $\cos(\cdot, \cdot)$ 表示余弦相似度函数, $\text{softmax}(\cdot)$ 表示 softmax 激活函数。

为实现文本与视觉模态的跨模态协同优化,在判断标记学习范式下,需要最小化第 2.2 节中定义的分类经验风险估计量 $\hat{R}_{\text{TF}}$ 。在最小化分类经验风险估计量 $\hat{R}_{\text{TF}}$ 的过程中,关键在于对式(2)中定义的条件概率分布 $p(y=y_i|\bar{y}=y_j, s=0, x)$ 进行准确估计。为充分挖掘并利用视觉语言模型 VLMs 中的先验知识,本节借助 CLIP 模型对该条件概率分布进行精确建模。具体来说,本节设计了一种余弦退火动态加权融合策略,用于计算最终的经验条件概率分布,并实现式(11)中学习表征与 CLIP 模型先验知识的有效融合:

$$\hat{p}(y=y_i|\bar{y}=y_j, s=0, x) = \lambda P_{\text{MCCP}} + (1-\lambda)P_{\text{CLIP}} \quad (12)$$

其中, P_{MCCP} 为式(11)估计的条件概率分布, P_{CLIP} 为 CLIP 模型估计的条件概率分布, λ 为控制两个概率分布权重占比的超参数,在训练过程中依照下述公式动态生成:

$$\lambda = \alpha \cdot \left(1 - \cos\left(\frac{\text{epoch}}{N_e} \pi\right) \right) \quad (13)$$

其中, $\alpha \in [0, 1]$ 为控制 λ 的最大取值的超参数, epoch 代表当前训练轮次, N_e 代表训练总轮次。通过式(13),超参数 λ 在训练过程中逐渐从 0 增加到 α 。该策略使得分类模型在训练初期依赖于 CLIP 模型的零样本预测能力(即 $\lambda=0$),随着训练轮次的增加,逐步增强式(11)中学习表征的贡献(即 $\lambda \rightarrow \alpha$)。换句话说,条件概率分布 $p(y=y_i|\bar{y}=y_j, s=0, x)$ 最开始完全由 CLIP 模型估计,随着训练轮次的增加,所提出 MCCP 方法的估计占比逐渐增加。通过这种估计方式,分类模型可以在训练的过程中动态地利用 CLIP 模型的先验知识提升分类性能。条件概率分布 $p(y=y_i|\bar{y}=y_j, s=0, x)$ 可以被近似表示为经验条件概率分布 $\hat{p}(y=y_i|\bar{y}=y_j, s=0, x)$ 。

进一步地,基于此经验条件概率分布可以按照式(3)定义经验风险一致性交叉熵损失 $\mathcal{L}_{\text{cross}}$,从而协同优化第 3.1 节和第 3.2 节中提出的文本提示网络、文本适配器、视觉提示网络及视觉适配器。上述基于多模态一致性协同提示微调的判断标记学习方法的整

体流程如算法 1 所示。

算法 1 基于多模态一致性协同提示微调的判断标记学习方法 MCCP

输入: 判断标记训练集 $D_{TF} = \{(x_i, (\bar{y}_i, s_i))\} (i=1, 2, \dots, N)$, 训练总轮次 N_e , CLIP 图像编码器 $g_I(\cdot)$, CLIP 文本编码器 $g_T(\cdot)$, 视觉适配器 $\text{adapter}_V(\cdot)$, 文本适配器 $\text{adapter}_T(\cdot)$, 初始图像提示矩阵 V_I , 初始文本提示矩阵 V_T , 视觉提示网络 $\phi_I(\cdot)$, 文本提示网络 $\phi_T(\cdot)$

输出: 文本提示网络权重参数 θ_{TP} , 文本适配器权重参数 θ_{TA} , 视觉提示网络权重参数 θ_{VP} , 视觉适配器权重参数 θ_{VA}

1. 初始化 MCCP 的参数 $\theta_{TP}, \theta_{TA}, \theta_{VP}$ 和 θ_{VA} , 并通过式(4)计算类别语义嵌入表征 q_{lim}

2. FOR $t=0$ TO N_e

3. 将 D_{TF} 随机划分为 B 个批次

4. FOR $b=0$ TO B

5. 从 D_{TF} 中提取小批次数据 D_B

6. 通过式(6)计算文本模态一致性损失 $\mathcal{L}_T$

7. 通过式(10)计算视觉模态一致性损失 $\mathcal{L}_V$

8. 通过式(12)计算条件概率估计 $\hat{p}(y=y_i | \bar{y}=y_i, s=0, x)$

9. 通过式(3)计算分类风险一致性交叉熵损失 $\mathcal{L}_{cross}$

10. 基于文本模态一致性损失 $\mathcal{L}_T$ 更新文本提示网络权重参数 θ_{TP} 和文本适配器权重参数 θ_{TA}

11. 基于视觉模态一致性损失 $\mathcal{L}_V$ 更新视觉提示网络权重参数 θ_{VP} 和视觉适配器权重参数 θ_{VA}

12. 基于分类风险一致性交叉熵损失 $\mathcal{L}_{cross}$ 更新上述所有权重参数 $\theta_{TP}, \theta_{TA}, \theta_{VP}$ 和 θ_{VA}

13. END FOR

14. END FOR

15. 输出文本提示网络权重参数 θ_{TP} , 文本适配器权重参数 θ_{TA} , 视觉提示网络权重参数 θ_{VP} , 视觉适配器权重参数

“多模态耦合训练导致模态不平衡与被动跟随”是多模态学习领域一个非常关键且极具挑战性的问题。如果在模型优化时将视觉模态一致性损失 $\mathcal{L}_V$ 、文本模态一致性损失 $\mathcal{L}_T$ 和跨模态分类风险一致性交叉熵损失 $\mathcal{L}_{cross}$ 直接相加作为整体损失并进行反向传播, 确实极易引发梯度冲突, 导致收敛较快的模态主导整个训练过程, 从而使另一模态陷入过拟合或特征学习停滞的潜在风险。为了预防这种风险, 本算法采用分阶段参数更新策略, 在保证模型泛化能力的同时, 强化对下游任务的适应性。首先, 在初始化阶段计算类别语义提示嵌入表征 q_{lim} 。将训练集 D_{TF} 随机划分为 B 个批次, 从 D_{TF} 中提取小批次数据 D_B , 分别计算文本模态一致性损失 $\mathcal{L}_T$ 、视觉模态一致性损失 $\mathcal{L}_V$ 和分类风险一致性交叉熵损失 $\mathcal{L}_{cross}$ 。最后, 通过将模态特异性优化与跨模态联合优化进行解耦来实现三阶段梯度更新策略, 具体如下: (1) 在文本模态, 通过文本模态一致性损失 $\mathcal{L}_T$ 更新文本提示网络与文

本适配器的参数; (2) 在视觉模态, 通过视觉模态一致性损失 $\mathcal{L}_V$ 更新视觉提示网络和图像适配器的参数; (3) 通过分类风险一致性交叉熵损失 $\mathcal{L}_{cross}$ 协同更新文本、视觉双模态的各网络参数。值得注意的是, 该交叉熵损失 $\mathcal{L}_{cross}$ 同样隐式地约束了视觉提示表征的学习过程, 根据式(11), 视觉嵌入表征必须与具备明确语义信息(由大语言模型协同增强)的文本提示嵌入表征进行交互与对齐, 以完成类别分类。这种跨模态联合优化机制, 使得视觉提示的优化方向始终限定在 CLIP 模型的丰富语义空间内, 确保了其生成的表征不仅在视觉上鲁棒, 且具有明确的语义含义, 避免了特征模式崩塌或学习无意义噪声。

上述策略通过将各模态优化阶段与跨模态联合优化阶段进行解耦, 有效避免了不同模态间的梯度冲突问题。在训练过程中, CLIP 模型的文本编码器和图像编码器参数均保持冻结状态。通过高效学习文本与视觉双模态的提示网络和适配器, 实现了 VLMs 先验知识向下游任务的高效迁移, 同时最大程度地保留了其泛化性能。

3.4 测试推理过程

在测试阶段, 图像分类任务以一个测试图像 x 和一个候选类别 y (即 CLIP 原始文本提示 “a photo of a [CLS].”) 作为输入, 并依据式(11)计算该测试图像属于该候选类别的预测概率。图 3 展示了所提出的基于多模态一致性协同提示微调的判断标记学习方法的测试推理过程。

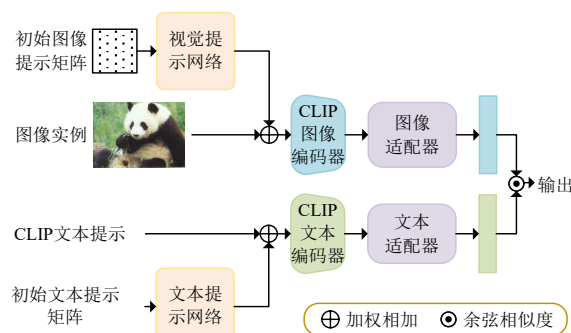


图3 测试推理流程示意图

Figure 3 Illustration of the inference process

与训练阶段类似, 在文本模态, 采用标准差为 0.02 的正态分布随机初始化初始文本提示矩阵 $V_T \in \mathbb{R}^{k \times d}$ 并输入到文本提示网络 $\phi_T(\cdot)$, 然后与原始文本提示 (即 “a photo of a [CLS].”) 加权相加后送入 CLIP 模型文本编码器 $g_T(\cdot)$ 和文本适配器 $\text{adapter}_T(\cdot)$ 中, 动态生成文本提示嵌入 q_T 。在视觉模态, 采用标准差为 0.02 的正态分布随机初始化初始图像提示矩阵 $V_I \in \mathbb{R}^{H \times W \times C}$, 经过视觉提示网络 $\phi_I(\cdot)$, 生成视觉

提示 P_v , 并将测试图像 x 与视觉提示 P_v 加权相加, 作为 CLIP 模型图像编码器 $g_i(\cdot)$ 和图像适配器 $\text{adapter}_v(\cdot)$ 的输入, 得到图像嵌入 q_i 。最终, 依据式 (11) 得到测试图像的分类概率 $P(y|x)$ 。

4 结果与分析

为验证所提出 MCCP 方法的有效性, 本节首先在九个基准数据集上系统比较了 MCCP 方法与传统弱监督学习方法、基于适配器或线性探针的方法, 以及其他提示学习方法的性能表现; 随后, 评估了 MCCP 方法在跨数据集迁移设置下的泛化能力; 接着, 对三种一致性约束损失函数进行了消融实验; 最后, 对视觉提示网络和文本提示网络的设计进行了消融分析。

4.1 实验设置

(1) 数据集

本文在涵盖 CIFAR-100^[34]、Tiny-ImageNet^[35]、Caletech-101^[36]、Food-101^[37]、Stanford Cars^[38]、DTD^[39]、FGVC-Aircraft^[40]、ImageNet-A^[41] 和 ImageNet-R^[42] 的九个多类别图像分类数据集上评估所提出 MCCP 方法的有效性。其中, FGVC-Aircraft 的文本提示为 “a photo of a [CLS], a type of aircraft.”; DTD 的文本提示为 “a photo of a [CLS] texture.”; 其他数据集的文本提示均为 “a photo of a [CLS].”。对于每个数据集, 训练集中所有图像的类别标记均被替换为判断标记, 而测试集则保留原始的真实标记。

(2) 对比方法

为全面评估所提出 MCCP 方法的有效性, 本文选取的对比方法涵盖传统弱监督学习方法、基于适配器或线性探针的方法, 以及提示学习方法。其中, 传统弱监督学习方法包括 ORCA^[43]、NACH^[44]、PaPi^[45]、CLL with WL^[30]、SCARCE^[46] 和 FineSSL^[47]; 基于适配器或线性探针的方法包括 Tip-Adapter^[48]、Tip-Adapter-F^[48]、CLIP 线性探针 (CLIP LP) 和 TMP^[13]; 提示学习方法包括 CoOp^[6]、CoCoOp^[7]、MaPLe^[8]、TCP^[19]、TextRefiner^[20] 和 CuPL^[18]。各方法的具体实验设置如下。

ORCA^[43] 和 NACH^[44]: 一种面向开放类别识别的半监督学习方法。在本文实验中, $s=1$ 的样本 (即判断标记为 “True”) 视为标注数据; 而 $s=0$ 的样本 (即判断标记为 “False”) 视为未标注数据。

PaPi^[45]: 一种利用层次化消歧方法引导正样本去噪的偏标记学习方法。在本文实验中, $s=1$ 样本视为标注数据; $s=0$ 样本视为偏标记数据, 并将随机采样类别外的所有类别视为偏标记学习的候选标记集。

CLL with WL^[30]: 一种使用新型加权损失的补充标记学习方法。在本文实验中, $s=1$ 样本视为标注数

据; $s=0$ 样本视为补充标记数据, 并将随机采样类别视为样本不属于的类别标记。

SCARCE^[46]: 一种利用风险校正 (risk-correction) 避免过拟合的补充标记学习方法。本文实验中, $s=1$ 样本视为标注数据; $s=0$ 样本视为补充标记数据, 并将随机采样类别视为样本不属于的类别标记。

FineSSL^[47]: 一种面向 VLMs 模型微调的半监督方法。本文实验中, $s=1$ 样本视为标注数据; $s=0$ 样本视为未标注数据。

Tip-Adapter^[48] 和 Tip-Adapter-F^[48]: 一种利用小样本训练集构建键-值缓存模型的适配器方法。Tip-Adapter 直接利用键-值缓存模型将小样本训练集知识与 CLIP 的预训练知识相结合。Tip-Adapter-F 则在 Tip-Adapter 的基础上利用训练集微调键-值缓存模型。本文实验中, 使用 $s=1$ 样本构建小样本训练集。

CLIP 线性探针 (CLIP LP)^[1]: 在 CLIP 视觉编码器后添加一个可训练线性分类器, 仅训练该线性分类器。本文使用多种变体: CLIP LPS、CLIP LPP 和 CLIP LPT。其中, CLIP LPS 表示使用人工标注的原始训练数据 (非判断标记标注方式) 进行全监督学习; CLIP LPP 表示使用 CLIP 标注的原始训练数据 (非判断标记标注方式) 进行全监督学习; CLIP LPT 表示仅使用 $s=1$ 个判断标记训练数据进行学习。

TMP^[13]: 一种基于判断标记一致性风险的多模态提示检索学习方法。本文实验遵循原论文的所有设置。

CoOp^[6]、CoCoOp^[7]、MaPLe^[8]、TCP^[19] 和 TextRefiner^[20] 在本文实验中: 当 $s=1$ 时, 样本视为标注数据; 当 $s=0$ 时, 样本视为未标注数据。此外, 损失函数遵循原论文使用交叉熵损失函数。

CuPL^[18]: 一种文本模态提示工程方法, 使用大语言模型生成类提示模板进行零样本推理。本文实验遵循原论文的所有设置。

(3) 实现细节与参数设置

为确保实验对比的公平性, 所有模型均采用 ViT-L/14 架构的 CLIP 模型作为视觉主干网络。对于文本模态一致性提示学习过程中的细粒度类别描述, 本文采用 GPT-4o mini 模型生成, 官方 API 版本为 “gpt-4o-mini-2024-07-18”。所有模型均在单张 NVIDIA RTX 4090 GPU 上训练 50 个轮次 (epochs)。文本提示网络和视觉提示网络采用 AdamW 优化器^[49] 进行优化, 初始学习率为 8×10^{-2} , 衰减率为 0.01, 最低学习率为 5×10^{-4} 。文本适配器和视觉适配器同样使用 AdamW 优化器进行优化, 初始学习率为 1×10^{-3} , 衰减率为 0.9, 最低学习率为 5×10^{-6} 。对于批次大小 (batch size), 弱监督学习方法及基于适配器/线性探针各方法均设定为 256; MCCP、CoOp、TCP 和 TextRefiner 方法的批次大

小设定为 32;CoCoOp 和 MaPLe 方法的批次大小遵循原文,分别设定为 1 和 4。对于式(13)中的权重 α ,本文在 $[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ 的范围内进行网格搜索(grid search),寻找能够平衡 CLIP 先

验知识与模型在式(11)中学习的表征的最优权重值。

4.2 CLIP 模型标注的判断标记数据上的对比实验

本节在 CLIP 模型标注的判断标记数据集上进行了对比实验,具体结果如表 4 所示。

表 4 在 CLIP 模型标注的判断标记数据上的分类准确率对比

单位: %

Table 4 Comparison of classification accuracy on the True-False labels dataset annotated by CLIP model

unit: %

模型	CIFAR-100	Tiny-ImageNet	Caltech-101	Food-101	Stanford Cars	DTD	FGVC-Aircraft	ImageNet-A	ImageNet-R	平均值
CLIP LPP	77.40	74.71	88.74	93.30	71.56	54.37	30.87	63.80	86.64	71.27
CLIP 零样本	75.58	72.66	87.24	92.76	70.76	52.36	<u>30.63</u>	62.49	85.12	69.96
CLIP LPT	25.25	18.24	22.63	66.69	4.96	10.58	4.11	2.80	5.28	17.84
Tip-Adapter	76.21	73.80	64.87	91.34	70.71	52.30	27.54	55.96	81.55	66.03
Tip-Adapter-F	76.07	74.91	83.20	91.42	70.71	53.61	27.90	57.24	83.93	68.78
TMP	78.22	75.14	89.14	93.52	<u>72.52</u>	<u>55.32</u>	28.86	59.16	86.24	<u>70.90</u>
ORCA	53.26	19.21	14.40	7.96	7.25	32.27	4.66	45.64	49.16	25.98
NACH	64.42	35.09	21.39	12.62	4.53	42.08	5.18	50.44	50.92	31.85
PaPi	63.73	41.50	43.27	81.94	10.19	33.40	6.03	52.04	35.96	40.90
CLL with WL	59.05	44.21	44.79	85.46	10.40	35.76	3.71	47.45	47.83	42.07
SCARCE	43.10	31.69	35.92	56.41	6.13	17.79	3.94	34.04	37.67	29.63
FineSSL	66.62	33.92	26.95	6.44	1.07	44.91	5.39	55.96	52.33	32.62
CoOp	74.97	73.36	85.31	92.30	67.33	51.54	29.07	<u>63.20</u>	82.14	68.80
CoCoOp	75.96	75.80	85.97	92.78	70.23	48.88	8.82	60.62	84.33	67.04
MaPLe	<u>77.60</u>	75.53	82.18	92.70	62.49	46.63	9.27	53.24	83.81	64.83
CuPL	76.75	<u>78.66</u>	86.33	92.80	71.65	48.52	9.05	57.11	81.56	66.94
TCP	74.79	71.24	48.43	92.19	58.34	35.64	9.75	51.77	72.91	57.23
TextRefiner	75.75	74.01	84.18	92.71	63.38	50.18	25.17	63.02	<u>86.63</u>	68.34
MCCP	77.40	79.21	<u>87.83</u>	<u>93.22</u>	72.65	60.40	35.64	67.96	87.93	73.58

注:粗体数值表示最优结果,下划线数值表示次优结果。

整体来看,本文提出的 MCCP 方法在 9 个数据集上取得了 73.58% 的平均分类准确率,超越了所有对比方法,证实了所提出基于多模态一致性协同提示微调的判断标记方法的有效性。具体来看,MCCP 方法在 9 个数据集中的 8 个(即 Tiny-ImageNet、Caltech-101、Food-101、Stanford Cars、DTD、FGVC-Aircraft、ImageNet-A 和 ImageNet-R)上取得了最优或次优的分类准确率,且与最优方法的准确率差距均小于 1.5%。

(1)与传统弱监督学习方法相比,MCCP 方法的平均分类准确率提升了超过 30%(73.58 vs. 42.07),表明传统弱监督学习方法难以有效利用 VLMs 模型中的先验知识,而 MCCP 方法能够充分利用这些知识。

(2)与其他提示学习方法相比,MCCP 方法在除 CIFAR-100 外的其他 8 个数据集上均取得了更优性能表现,且在 CIFAR-100 数据集上取得了相近性能(77.40 vs. 77.60),这说明 MCCP 方法能够更好地利用 VLMs 模型中的先验知识。

(3)与基于适配器或线性探针的方法相比,尤其是 TMP 方法,MCCP 方法在平均分类准确率上展现出

较大幅度的优势(73.58 vs. 70.90),说明 MCCP 方法在适应下游任务的同时,能够更大程度地保留预训练 VLMs 模型的泛化能力。相比之下,TMP 方法引入了线性分类器,其适应下游任务的能力是以牺牲 VLMs 模型泛化能力为代价而实现的。值得注意的是,在 FGVC-Aircraft 数据集上,由于不同飞行器的外观相近,在判别飞行器型号的分类任务上难度极大,传统弱监督方法和除 CoOp 和 TextRefiner 外的提示学习方法的性能表现都极差(准确率小于 10.0%),本文 MCCP 方法取得了 35.64% 的准确率,提升巨大(35.64 vs. 9.75)。

(4)与全监督学习范式下的 CLIP LPP 方法相比,MCCP 方法在所有数据集上均取得了相近或者更好的性能表现,平均准确率也更高(73.58 vs. 71.27),这说明 MCCP 方法能够从稀疏的判断标记样本中有效学习到高质量的类别语义信息,从而缩小与全监督学习方法间的性能差距,甚至超越全监督学习方法。此外,CLIP LPP 方法的性能明显低于表 6 中 CLIP LPS 方法的原因可能是:CLIP 模型在样本标注过程中引入了错误标签。

综上所述,本文提出的 MCCP 方法在下游任务的适应性与模型泛化能力之间实现了更优的平衡,展现出了明显优势。

为了进一步验证 MCCP 方法相较于基线模型在“监督样本缺失类别”上的性能表现。本节在 CLIP 模型标注的判断标记数据集上,将类别集合划分为“有监督样本的类别 S_{have} ”和“完全缺失监督样本的类别

S_{miss} ”两个互斥子集,对比分析了 MCCP 方法和次优基线方法 TMP 在这两个类别集合上的平均分类准确率。“有监督样本的类别”意味着该类别存在 $s=1$ 的标注样本,“完全缺失监督样本的类别”表示该类别没有 $s=1$ 的标注样本。平均分类准确率表示,先计算各个类别的分类准确率,再对上述两个集合内所有类别的准确率求平均值。具体结果如表 5 所示。

表 5 “有监督样本类别”和“完全缺失监督样本类别”的平均分类准确率

单位: %

Table 5 Average classification accuracy for “categories with supervised samples” and “categories without supervised samples” unit: %

模型	CIFAR-100		Tiny-ImageNet		Caltech-101		Food-101		Stanford Cars	
	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}
TMP	85.16	60.79	82.90	58.73	96.51	55.67	98.53	74.64	76.73	19.42
MCCP	84.72	64.43	83.19	64.85	95.78	59.96	97.80	78.13	76.98	26.08

注:粗体数值表示最优结果。

整体来看,在有监督样本类别 S_{have} 上, MCCP 方法与 TMP 的性能相近,差距在 1% 以内;但在完全缺失监督样本类别 S_{miss} 上, MCCP 方法相较于 TMP 展现出了较大幅度的性能提升,提升幅度大约在 3%~7%。需要说明的是,由于 Stanford Cars 数据集上“完全缺失监督样本的类别”多达 156 个(一共 196 个类别),远多于其他数据集, MCCP 和 TMP 方法在其上的平

均分类准确率要大幅度低于其他数据集(19.42 vs. 55.67)。上述结果证实了 MCCP 方法利用大语言模型补充语义信息以缓解监督样本缺失问题是极其有效的。

4.3 人工标注的判断标记数据上的对比实验

本节在人工标注的判断标记数据集上进行了对比实验,具体结果如表 6 所示。

表 6 在人工标注的判断标记数据上的分类准确率对比

单位: %

Table 6 Comparison of classification accuracy on the True-False labels dataset annotated by human annotators unit: %

模型	CIFAR-100	Tiny-ImageNet	Caltech-101	Food-101	Stanford Cars	DTD	FGVC-Aircraft	ImageNet-A	ImageNet-R	平均值
CLIP LPS	85.81	85.31	96.76	94.94	87.71	76.36	53.68	71.51	90.19	82.47
CLIP 零样本	75.58	72.66	87.24	92.76	70.76	52.36	<u>30.63</u>	62.49	85.12	69.96
CLIP LPT	25.60	21.37	21.36	68.23	3.69	10.05	3.09	2.71	5.17	17.92
Tip-Adapter	76.47	73.86	86.77	91.82	70.60	52.30	21.81	53.78	81.63	67.67
Tip-Adapter-F	77.40	75.39	86.81	92.10	70.61	53.25	22.98	55.91	84.07	68.72
TMP	<u>78.72</u>	75.84	<u>90.60</u>	93.55	<u>72.60</u>	<u>55.73</u>	28.89	58.49	<u>86.44</u>	<u>71.21</u>
ORCA	50.54	17.65	14.14	7.82	7.25	28.37	6.44	23.37	25.17	20.08
NACH	63.46	31.78	15.09	8.51	5.57	30.98	4.01	27.93	55.15	26.94
PaPi	60.69	40.80	47.06	80.31	6.00	28.33	4.98	27.65	42.52	37.59
CLL with WL	63.25	51.54	50.73	87.85	9.64	27.62	4.74	23.82	29.72	38.77
SCARCE	44.06	33.13	37.39	57.64	4.37	17.58	2.59	21.20	35.84	28.20
FineSSL	67.28	35.18	28.72	6.57	1.21	19.25	5.46	35.02	56.48	28.35
CoOp	77.48	77.64	88.05	92.92	66.21	52.36	26.88	61.16	81.86	69.40
CoCoOp	77.59	77.70	89.91	92.94	62.80	47.46	6.39	<u>61.47</u>	84.67	66.77
MaPLe	78.90	77.84	88.05	93.37	61.96	36.05	15.54	46.22	81.73	64.41
CuPL	76.75	<u>78.66</u>	86.33	92.80	71.65	47.16	18.90	49.78	82.41	67.16
TCP	76.48	75.82	90.99	92.59	58.36	35.05	13.80	49.56	69.84	62.50
TextRefiner	77.32	77.23	88.88	92.86	59.08	52.60	27.72	59.24	85.67	68.96
MCCP	78.12	79.47	89.76	<u>93.38</u>	72.94	61.52	36.18	68.40	88.01	74.20

注:粗体数值表示最优结果,下划线数值表示次优结果。

整体来看,本文 MCCP 方法在九个数据集上取得了 74.20% 的平均分类准确率,优于所有对比方法,证实

了所提出基于多模态一致性协同提示微调的判断标记方法的有效性。值得注意的是,在 Food-101 数据集

上, MCCP方法的分类准确率已接近全监督学习的性能表现(93.38 vs. 94.94), 这表明 MCCP方法借助大语言模型的知识协同, 有效学习到了高精度的类别语义信息。

具体来看, MCCP方法在九个数据集上的七个(即 Tiny-ImageNet、Food-101、Stanford Cars、DTD、FGVC-Aircraft、ImageNet-A 和 ImageNet-R)上取得了最优或次优的分类准确率, 且与最优方法的准确率差距不超过 1.2%。

(1) 与传统弱监督学习方法相比, MCCP方法的平均分类准确率提升了超过 30%(74.20 vs. 38.77), 表明 MCCP方法能够充分利用 CLIP 模型和大语言模型的先验知识, 从稀疏监督信息中实现高效提示学习, 而传统弱监督学习方法难以有效利用这些先验知识。在 Stanford Cars 数据集上, 传统弱监督学习方法的分类准确率极低, 而 MCCP方法的分类准确率达到 72.94%, 表明其能够捕捉细粒度的判别性特征。相比之下, 传统弱监督学习方法因无法建模跨模态语义关联, 在类别间相似度较高的 Stanford Cars 数据集上几乎无法进行有效分类。

(2) 与其他提示学习方法相比, MCCP方法在除 CIFAR-100 和 Caltech-101 外的七个数据集上均取得了更优的性能表现。在 Stanford Cars 上, MCCP方法的分类准确率为 72.94%, 较 CuPL 方法提升 1.29%(72.94 vs. 71.65), 较 CoOp 方法提升 6.73%(72.94 vs. 66.21), 较其他提示学习方法提升超过 10%(72.94 vs. 62.80)。尽管 MaPLe 和 TCP 方法分别在 CIFAR-100 和 Caltech-101 数据集上取得了 0.78%(78.90 vs. 78.12) 和 1.23%(90.99 vs. 89.76) 的性能领先, 但它们在 Stanford Cars 数据集上性能表现不佳, 分类准确率降低了超过 10%。在 DTD、FGVC-Aircraft、ImageNet-A 和 ImageNet-R 这 4 个数据集上, MCCP方法较其他提

示学习方法分别至少提升了 8.92%(61.52 vs. 52.60)、8.46%(36.18 vs. 27.72)、6.93%(68.40 vs. 61.47) 和 2.34%(88.01 vs. 85.67)。这表明 MCCP方法依托三阶段优化策略和分模态提示学习, 不仅在 Tiny-ImageNet 等粗粒度分类任务上表现出竞争力, 也在 Stanford Cars、FGVC-Aircraft 等类别间相似度极高的细粒度分类任务上展现出显著的性能优势。

(3) 与基于适配器或线性探针的方法相比, 尤其是 TMP 方法, MCCP方法在 ImageNet-A 数据集上取得了 9.91% 的显著性能提升(68.40 vs. 58.49), 在其他数据集上要么表现接近, 要么有不同程度的性能提升, 验证了其性能竞争力。

(4) 与全监督学习方法 CLIP LPS 相比, MCCP方法在整体上还是存在性能差距。这是因为 MCCP 采用判断标记标注方式, 属于弱监督学习范畴; 而 CLIP LPS 方法采用全监督范式进行训练, 可以看作模型的性能上限。MCCP方法的设计初衷并非在绝对性能上超越全监督方法, 而是在利用判断标记标注方式大幅度降低标注成本的前提下, 尽可能地缩小弱监督与全监督范式之间的性能差距。

如表 6 所示, 尽管存在监督信息上的天然劣势, MCCP方法在 Food-101 数据集上取得了 93.38% 的准确率, 已经极其逼近全监督 CLIP LPS 方法的 94.94%, 这有力地证明了 MCCP方法能够从极度稀疏的标注样本中有效学习到高质量类别语义表征。为了进一步验证 MCCP方法相较于基线模型在“监督样本缺失类别”上的性能表现。本节在人工标注的判断标记数据集上, 将类别集合划分为“有监督样本的类别 S_{have} ”和“完全缺失监督样本的类别 S_{miss} ”两个互斥子集, 对比分析了 MCCP方法和次优方法 TMP 在两个类别集合上的平均分类准确率。具体结果如表 7 所示。

表 7 “有监督样本类别”和“完全缺失监督样本类别”的平均分类准确率

单位: %

Table 7 Average classification accuracy for “categories with supervised samples” and “categories without supervised samples”

unit: %

模型	CIFAR-100		Tiny-ImageNet		Caltech-101		Food-101		Stanford Cars	
	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}	S_{have}	S_{miss}
TMP	85.69	61.17	82.98	59.26	96.83	55.49	98.69	74.86	76.80	20.62
MCCP	84.92	64.33	83.35	65.51	95.55	60.70	97.54	78.46	77.43	28.25

注: 粗体数值表示最优结果。

整体来看, 在有监督样本类别 S_{have} 上, MCCP方法与 TMP 的性能相近, 差距在 0.8% 以内; 但在完全缺失监督样本类别 S_{miss} 上, MCCP方法相较于 TMP 展现出了大幅度的性能提升, 幅度大约在 3%~8%, 证实了 MCCP方法利用大语言模型补充语义信息以缓解监督样本缺失问题是极其有效的。

4.4 跨数据集迁移对比实验

为进一步验证 MCCP方法在跨数据集迁移场景

下的泛化能力, 本节在 Tiny-ImageNet 数据集上完成模型训练, 并直接在不相关数据集 (CIFAR-100、Caltech-101、Food-101 和 Stanford Cars) 上进行评估。这些数据集之间存在显著的任务异质性, 例如从通用对象分类到细粒度车辆分类的转变, 这为模型可迁移性评估提供了严格的验证环境, 能够有效验证 MCCP在开放环境下的泛化能力。由于对比方法 TMP 引入了线性分类器, 无法进行跨数据集迁移, 因此本节未

将其纳入实验对比。具体结果如表 8 所示。

表 8 跨数据集迁移场景下各方法分类准确率对比 单位: %

Table 8 Comparison of classification accuracy among different methods in cross-dataset transfer scenarios unit: %

判断标记来源	模型	CIFAR-100	Caltech-101	Food-101	Stanford Cars	平均值
CLIP 模型标注	CoOP	71.85	80.32	89.77	57.29	74.81
	CoCoOp	74.43	82.98	91.45	64.99	78.46
	MaPLe	75.01	83.56	91.77	65.97	79.08
	TCP	55.03	62.39	26.74	6.74	37.73
	TextRefiner	73.01	81.63	90.85	60.78	76.57
	MCCP	76.74	85.93	92.11	70.63	81.35
人工标注	CoOP	75.65	83.71	92.25	65.38	79.25
	CoCoOp	75.46	86.59	92.42	69.02	80.87
	MaPLe	73.93	76.49	91.34	62.26	76.01
	TCP	57.95	76.20	49.38	11.35	48.72
	TextRefiner	74.60	84.07	91.70	65.40	78.94
	MCCP	76.34	85.93	92.43	70.59	81.32

注:粗体数值表示最优结果。

整体来看, MCCP 方法在跨数据集迁移实验中表现出优异的表征鲁棒性,无论是在 CLIP 模型标注的判断标记数据集之间,还是在人工标注的判断标记数据集之间,均取得了稳定优异的性能表现。具体而言,在人工标注的判断标记数据集之间,可以看到: (1) MCCP 方法在 CIFAR-100、Food-101 和 Stanford Cars 这 3 个数据集上取得了最优迁移分类准确率,在 Caltech-101 数据集上仅以 0.66% 的性能差距位列次优; (2) 在 Stanford Cars 细粒度分类任务中, MCCP 方法的迁移分类准确率达 70.59%, 较次优方法 CoCoOp 提升 1.57%; (3) 在 CIFAR-100 数据集上, MCCP 方法取得了 76.34% 的迁移分类准确率, 较次优方法 CoOP 提高 0.69%; (4) 在 Food-101 食物纹理分类任务中, MCCP 方法的迁移分类准确率为 92.43%, 与表 6 中“在 Food-101 数据集上训练并测试”的源域测试结果相比, 仅相差 0.95% (92.43 vs. 93.38)。上述稳定迁移性能验证了 MCCP 方法所学习的视觉语义表征的鲁棒性, 证实了 MCCP 方法的有效性。

在 CLIP 模型标注的判断标记数据集之间, 可以得到以下观察结果: (1) MCCP 方法在 CIFAR-100、Caltech-101、Food-101 和 Stanford Cars 这 4 个数据集上均取得了最优迁移分类准确率, 分别较次优方法提升了 1.73% (76.74 vs. 75.01)、2.37% (85.93 vs. 83.56)、0.34% (92.11 vs. 91.77) 和 4.66% (70.63 vs. 65.97), 验证了其出色的跨数据集迁移能力; (2) 值得注意的是, 部分方法的迁移分类准确率甚至超过了人工标记数据集上的表现。例如, MaPLe 方法的平均迁移分类

准确率高达 79.08%, 较其在人工标注数据集上的结果 76.01% 提升了 3.07%; MCCP 方法以 81.35% 的平均迁移分类准确率实现最优性能, 略高于其在人工标注数据集上的结果 81.32%。其原因可能在于, 判断标记的随机采样机制可看作一种隐式正则化手段 (类似于 Dropout 和随机数据增强), 在人工重新标注判断标记数据集时, 原始数据集中存在的噪声标记往往不可避免地继承; 而由 CLIP 模型标注的判断标记数据集可能在一定程度上缓解了这些固有的标注噪声, 从而在部分情况下表现出更优性能。

4.5 多模态一致性损失的消融实验分析

本节基于上述人工标注的判断标记数据集, 对所提出 MCCP 方法中的三个一致性约束损失进行了消融实验分析, 具体采用以下 4 种实验设置: (a) 去除视觉模态一致性损失 $\mathcal{L}_V$; (b) 去除文本模态一致性损失 $\mathcal{L}_T$; (c) 同时去除视觉模态一致性损失 $\mathcal{L}_V$ 和文本模态一致性损失 $\mathcal{L}_T$; (d) 移除分阶段参数更新策略, 将视觉模态一致性损失 $\mathcal{L}_V$ 、文本模态一致性损失 $\mathcal{L}_T$ 和跨模态分类风险一致性交叉熵损失 $\mathcal{L}_{\text{cross}}$ 直接相加作为整体损失, 进行模型训练。由于跨模态分类风险一致性交叉熵损失 $\mathcal{L}_{\text{cross}}$ 为分类任务所必需, 因此无法对其进行消融分析。具体结果如表 9 所示。

表 9 多模态一致性损失消融实验的分类准确率 单位: %

Table 9 Classification accuracy of ablation experiments on multi-modal consistency loss unit: %

$\mathcal{L}_{\text{cross}}$	$\mathcal{L}_T$	$\mathcal{L}_V$	CIFAR-100	Tiny-ImageNet	Caltech-101	Food-101	Stanford Cars	平均值
√	√	√	78.12	79.47	89.76	93.38	72.94	82.73
√	√		77.24	79.15	89.54	93.14	72.44	82.30
√		√	77.81	79.17	87.79	93.18	72.68	82.33
√			77.19	79.07	89.18	93.05	72.40	82.18
移除分阶段参数更新策略			73.22	75.83	84.74	89.05	64.89	77.55

注:粗体数值表示最优结果,√表示保留相应损失函数。

实验结果表明, MCCP 方法中的三种一致性约束损失对模型性能提升具有不同程度的贡献, 逐步移除一致性约束损失会导致性能呈阶梯式下降。具体而言: (1) 移除视觉模态一致性损失 $\mathcal{L}_V$ 导致模型在 CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101 和 Stanford Cars 数据集上的分类准确率分别下降 0.88% (78.12→77.24)、0.32% (79.47→79.15)、0.22% (89.76→89.54)、0.24% (93.38→93.14)、0.50% (72.94→72.44); (2) 移除文本模态一致性损失 $\mathcal{L}_T$ 导致模型在 CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101 和 Stanford Cars 数据集上的分类准确率分别下降 0.31% (78.12→77.81)、0.30% (79.47→79.17)、1.97% (89.76→87.79)、

0.20% (93.38→93.18)、0.26% (72.94→72.68); (3)同时去除视觉和文本模态一致性损失 $\mathcal{L}_V$ 和 $\mathcal{L}_T$ 导致模型在 CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101 和 Stanford Cars 数据集上的分类准确率分别下降 0.93% (78.12→77.19)、0.40% (79.47→79.07)、0.58% (89.76→89.18)、0.33% (93.38→93.05)、0.54% (72.94→72.40), 平均分类准确率下降 0.55% (82.73→82.18); (4)移除分阶段参数更新策略在 CIFAR-100、Tiny-ImageNet、Caltech-101、Food-101 和 Stanford Cars 数据集上的分类准确率均导致了大幅度的下降, 最大降幅 8.05% (72.94→64.89), 证实了本文提出的分阶段参数更新策略的有效性和必要性。这些结果也验证了三个一致性损失的有效性。

4.6 提示网络的消融实验分析

为验证本文使用的文本提示网络 $\phi_T(\cdot)$ 和视觉

提示网络 $\phi_I(\cdot)$ 的有效性, 本节基于人工标注的判断标记数据集, 设计并对比了以下两种实验设置: (1) 本文 MCCP 方法采用的是通过优化视觉提示网络 $\phi_I(\cdot)$ 和文本提示网络 $\phi_T(\cdot)$ 间接生成提示嵌入表征; (2) 直接优化尺寸为 $224 \times 224 \times 3$ 的图像以及长度为 77 的文本序列, 其中 77 是 CLIP 文本提示所允许的最大长度。实验结果如表 10 和图 4 所示。

表 10 文本和视觉提示网络消融实验的分类准确率 单位: %

方法/模型	CIFAR-100	Tiny-ImageNet	Caltech-101	Food-101	Stanford Cars	平均值
MCCP	78.12	79.47	89.76	93.38	72.94	82.73
直接优化	77.17	79.04	88.01	93.30	71.36	81.78

注: 粗体数值表示最优结果。

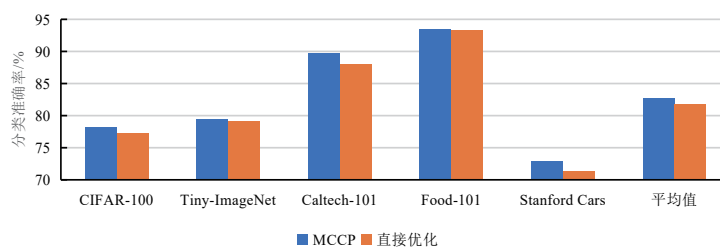


图 4 文本和视觉提示网络消融实验分类准确率的折线图

Figure 4 Line chart showing classification accuracy of ablation experiment on textual and visual prompt networks

实验结果表明, 在所有数据集上, MCCP 方法的分类准确率均优于“直接优化图像和文本”的方法。尤其是在 Caltech-101 数据集上, MCCP 方法的分类准确率较直接优化方法提升了 1.75 个百分点 (89.76 vs. 88.01)。此外, 如图 4 所示, MCCP 方法在 CIFAR-100、Caltech-101 和 Stanford Cars 三个数据集上相较于直接优化方法展现出更加明显的性能提升。综合来看, 尽管两种实验设置在参数规模上相近, 但本文采用的文本提示网络 $\phi_T(\cdot)$ 和视觉提示网络 $\phi_I(\cdot)$ 在捕获图像局部细节特征及建模文本内部依赖关系方面表现出更明显的优势。

4.7 文本模态一致性约束损失定义方式的影响

为了探究 3.1 节式 (6) 中一致性约束损失 $\mathcal{L}_T$ 的不同定义方式对 MCCP 模型性能的影响, 本节基于人工标注的判断标记数据集, 对两种定义方式进行了实验对比分析, 具体结果如表 11 所示。

两种定义方式如下: (1) 均方误差, 具体如式 (6) 所述; (2) KL 散度, 具体定义为

$$\mathcal{L}_T = \frac{1}{k} \sum_{i=1}^k \text{KL}(q_T^i, q_{\text{ilm}}^i) \quad (14)$$

其中, $\text{KL}(\cdot, \cdot)$ 表示 KL 散度 (Kullback-Leibler divergence) 计算函数, 计算时首先要利用 softmax 函数将 q_T^i

表 11 文本模态一致性约束损失不同定义方式的分类准确率

单位: %

损失定义方式	CIFAR-100	Tiny-ImageNet	Caltech-101	Food-101	Stanford Cars	平均值
均方误差	78.12	79.47	89.76	93.38	72.94	82.73
KL 散度	75.45	78.11	86.94	91.60	69.22	80.26

注: 粗体数值表示最优结果。

与 q_{ilm}^i 转换为概率分布形式, 满足“非负且总和为 1”的特性要求。

可以看到, 相较于均方误差定义方式, KL 散度定义方式在所有数据集上均导致了明显的性能下降, 证实了式 (6) 中均方误差的定义方式更加适合 q_T^i 与 q_{ilm}^i 这种特征级别的对齐。

4.8 超参数敏感性分析

为探究 MCCP 方法对式 (13) 中权重 α 的敏感性, 本节基于人工标注的判断标记数据集, 在 $[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ 的范围内, 对权重 α 的不同取值进行了实验对比分析, 具体结果展示在图 5 中。

可以看到, 当权重 α 取值较小 (如 0.0~0.2) 时,

MCCP模型在整个训练周期内都高度依赖CLIP模型的零样本条件概率分布 P_{CLIP} 。这种设定虽然最大程度地保留了预训练先验知识,但也限制了提示网络和适配器网络学习下游任务特异性特征的能力,导致性能遭遇瓶颈。当权重 α 取值过大(如0.8~1.0)时,MCCP模型在训练后期会大幅度地抛弃CLIP模型的概率估计 P_{CLIP} 。考虑到在判断标记学习范式下,

$s=1$ 的正样本监督信号极其稀少(如表1所示),完全依赖MCCP模型当前学习到的嵌入表征极易引发模型在长尾类别上的过拟合,从而削弱预训练CLIP模型原本强大的泛化能力,导致下游任务性能下降。因此,一个适中的 α 取值是确保预训练先验知识与下游任务适应性表征之间平滑过渡与动态融合的关键,一般取值0.3或0.4即可获得较优性能。

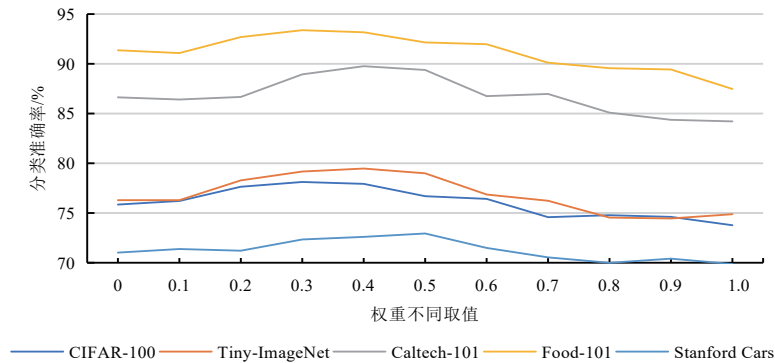


图5 各数据集上不同权重 α 设置下的分类准确率折线图

Figure 5 Line chart showing classification accuracy under different weight α across different datasets

4.9 计算开销的对比分析

本方法在训练阶段并不存在反复调用大模型带来的高昂开销。(1)在文本模态一致性提示学习中,的确需要调用LLM为每个类别生成细粒度的类别文本描述,但该过程属于数据预处理步骤。正如3.3节的算法1所示,利用LLM生成类别文本描述并获取类别语义嵌入表征的过程,发生在主训练循环之前,是一次性的离线预处理步骤,仅在数据准备阶段产生极少量的固定时间开销,完全不会增加模型在训练和推理阶段的计算开销。(2)在训练文本或视觉的提示网络和适配器时,的确会反复调用CLIP模型的文本编码器和视觉编码器,但是在整个训练与推理过程中,这两部分的参数处于完全冻结(frozen)状态,不参与

反向传播和梯度更新,计算开销很小。实际参与梯度计算和参数更新的,仅是本文引入的轻量级的提示网络和适配器模块。因此,上述过程中反复调用CLIP和LLM,并不会带来高昂的计算开销。

为了展现MCCP方法的计算效率,本节以相同的硬件环境(单张NVIDIA RTX 4090 GPU)和主干网络(ViT-L/14架构的CLIP)为基准,在CLIP模型标注的Caltech-101数据集上,批次大小设为1,将本文提出的MCCP方法与次优基线方法TMP进行了全方位对比。具体指标包括:总参数量(parameters)、可训练参数量(trainable parameters)、浮点运算次数(Floating Point Operations, FLOPs)、峰值显存占用(peak GPU memory)和单轮训练时间(time per epoch)。具体结果如表12所示。

表12 MCCP方法与主流基线方法的计算效率对比

Table 12 Comparison of computational efficiency between MCCP and standard baseline models

模型	总参数量	可训练参数量	浮点运算次数	峰值显存占用/GB	单轮训练时间/min
TMP	4.2949×10^8	1.87×10^6	9.753×10^{10}	9.95	17.22
MCCP	4.2921×10^8	1.59×10^6	9.948×10^{10}	7.63	28.37

可以看到,与TMP方法相比,本文的MCCP方法在总参数量、可训练参数量两个指标上相差无几,差距在 3.0×10^5 参数以内。由于MCCP方法需要多次调用CLIP模型的文本编码器和视觉编码器,因此,相较于TMP方法在浮点运算次数和单轮训练时间两个指标上有所提升。值得注意的是,MCCP方法的绝对训练时间成本仍处于轻量级微调的范畴,其单轮训练耗

时28.37 min,按照本文设定的50个训练轮次(epochs)来计算,模型在单张NVIDIA RTX 4090 GPU上的总训练时间不超过24 h。在当今视觉-语言模型微调动辄消耗整个计算集群资源并耗时数天甚至数十天的背景下,单卡单日内完成收敛,客观证明了MCCP方法的时间开销仍处于轻量级的较低水平。更进一步,虽然MCCP方法的单轮训练时间相较于TMP方法增加

了 11.15 min, 但是其仍存在以下优势, 以佐证 MCCP 方法的必要性: (1) 如表 4 和表 6 所示, MCCP 方法相较于 TMP 在九个数据集上的平均性能提升了大约 2%~3% (73.58 vs. 70.90, 71.21 vs. 74.20)。在 ImageNet-A 这样的复杂数据集上, MCCP 方法相较于 TMP 甚至实现了高达 9.91% 的绝对性能提升。(2) 如表 5 和表 7 所示, MCCP 方法相较于 TMP 方法在“监督样本缺失类别”上取得了大幅度的性能提升, 提升幅度大约为 3%~8%。尤其是在细粒度分类的 Stanford Cars 数据集上, 分类准确率从 20.62% 提升至 28.25%, 相对提升比例为 37%。这证实了 MCCP 方法引入的大语言模型知识协同机制对于解决监督信号缺失问题具有无可替代的必要性。(3) MCCP 方法具备更低的显存硬件门槛。如表 12 所示, MCCP 方法相较于 TMP 方法的峰值显存占用更低 (7.63 GB vs. 9.95 GB), 能够稳定运行在广泛普及的 8 GB 显存的消费级显卡上 (如 NVIDIA RTX 3070 和 NVIDIA RTX 4060), 而 TMP 方法会面临显存溢出风险, 这客观降低了 MCCP 方法的训练与复现门槛。

5 结论

本文针对现有判断标记学习方法面临的部分类别监督样本缺失以及缺乏跨模态联合优化的问题, 提出了一种基于多模态一致性协同提示微调的判断标记学习方法 MCCP。具体而言, 在文本模态, 设计了大语言模型驱动的知识协同机制, 利用大语言模型生成细粒度的类别文本描述, 并结合文本提示网络和文本适配器, 约束原始文本提示与类别文本描述之间的语义一致性, 实现了文本提示的动态优化, 从而缓解了类别监督信号缺失的影响; 在视觉模态, 设计了基于自监督学习的一致性提示学习机制, 通过计算原始图像、弱增强视图、强增强视图特征分布之间的一致性约束, 优化视觉提示网络和视觉适配器, 提升了视觉语义表征的鲁棒性与泛化能力; 在跨模态联合优化方面, 采用了分阶段梯度更新策略优化多类别图像分类任务的经验风险一致性损失, 并重构了判断标记学习框架中的条件概率估计机制, 通过引入余弦动态加权策略有效平衡了预训练先验知识与领域适应表征之间的交互融合, 实现了跨模态语义空间的协同对齐。在九个图像分类数据集上的实验结果表明, MCCP 方法取得了极具竞争力的性能表现, 并在跨数据集迁移设置下的四个目标数据集上实现了最优的平均分类准确率, 验证了其有效性。消融实验亦验证了 MCCP 方法各组成模块的有效性和必要性。

参考文献

- [1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2021: 8748-8763.
- [2] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2022: 12888-12900.
- [3] Jia Chao, Yang Yinfei, Xia Ye, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2021: 4904-4916.
- [4] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 34892-34916.
- [5] 李鑫尧, 李晶晶, 朱磊, 等. 资源受限的大模型高效迁移学习算法研究综述[J]. 计算机学报, 2024, 47(11): 2491-2521.
- [6] Li Xinyao, Li Jingjing, Zhu Lei, et al. Efficient transfer learning of large models with limited resources: A survey[J]. Chinese Journal of Computers, 2024, 47(11): 2491-2521. (in Chinese)
- [7] Zhou Kaiyang, Yang Jingkan, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.
- [8] Zhou Kaiyang, Yang Jingkan, Loy C C, et al. Conditional prompt learning for vision-language models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 16795-16804.
- [9] Khattak M U, Rasheed H, Maaz M, et al. Maple: Multi-modal prompt learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 19113-19122.
- [10] Jia Menglin, Tang Luming, Chen B C, et al. Visual prompt tuning[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 709-727.
- [11] Gao Peng, Geng Shijie, Zhang Renrui, et al. CLIP-adapter: Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2024, 132(2): 581-595.
- [12] Menghini C, Delworth A, Bach S. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 60984-61007.
- [13] Zhang Jiahao, Wei Qi, Liu Feng, et al. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 60004-60020.
- [14] Li Zhongnian, Xu Jinghao, Ying Peng, et al. Learning from

- true-false labels via multi-modal prompt retrieving[C]//Proceedings of the 42nd International Conference on Machine Learning. New York: ACM, 2025: 36652-36666.
- [14] Wang Xudong, Wu Zhirong, Lian Long, et al. Debaised learning from naturally imbalanced pseudo-labels[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 14627-14637.
- [15] 李仲年, 皇甫志宇, 杨凯杰, 等. 基于图提示的半监督开放词汇多标记学习[J]. 计算机研究与发展, 2025, 62(2): 432-442.
- Li Zhongnian, Huangfu Zhiyu, Yang Kaijie, et al. Semi-supervised open vocabulary multi-label learning based on graph prompting[J]. Journal of Computer Research and Development, 2025, 62(2): 432-442. (in Chinese)
- [16] 孙仁科, 许靖昊, 皇甫志宇, 等. 基于视觉-语言预训练模型的零样本迁移学习方法综述[J]. 计算机工程, 2024, 50(10): 1-15.
- Sun Renke, Xu Jinghao, Huangfu Zhiyu, et al. Survey of zero-shot transfer learning methods based on vision-language pre-trained models[J]. Computer Engineering, 2024, 50(10): 1-15. (in Chinese)
- [17] Menon S, Vondrick C. Visual classification via description from large language models[C]//Proceedings of the 11th International Conference on Learning Representations. PMLR, 2023: 1-17.
- [18] Pratt S, Covert I, Liu R, et al. What does a platypus look like? Generating customized prompts for zero-shot image classification[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 15645-15655.
- [19] Yao Hantao, Zhang Rui, Xu Changsheng. TCP: Textual-based class-aware prompt tuning for visual-language model[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 23438-23448.
- [20] Xie Jingjing, Zhang Yuxin, Peng Jun, et al. TextRefiner: Internal visual feature as efficient refiner for vision-language models prompt tuning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(8): 8718-8726.
- [21] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of the International Conference on Learning Representations. LCLR, 2021: 1-21.
- [22] 张震宇, 姜远. 面向特征演变环境的标记噪声鲁棒学习算法[J]. 计算机研究与发展, 2023, 60(8): 1740-1753.
- Zhang Zhenyu, Jiang Yuan. Label noise robust learning algorithm in environments evolving features[J]. Journal of Computer Research and Development, 2023, 60(8): 1740-1753. (in Chinese)
- [23] Cour T, Sapp B, Taskar B. Learning from partial labels[J]. Journal of Machine Learning Research, 2011, 12: 1501-1536.
- [24] Tang Caizhi, Zhang Minling. Confidence-rated discriminative partial label learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 2611-2617.
- [25] Zhang Minling, Zhou Binbin, Liu Xuying. Partial label learning via feature-aware disambiguation[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1335-1344.
- [26] Xu Ning, Lv Jiaqi, Geng Xin. Partial label learning via label enhancement[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 5557-5564.
- [27] Hüllermeier E, Beringer J. Learning from ambiguously labeled examples[J]. Intelligent Data Analysis, 2006, 10(5): 419-439.
- [28] Ishida T, Niu Gang, Hu Weihua, et al. Learning from complementary labels[C]//Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017: 5644-5654.
- [29] Ishida T, Niu Gang, Menon A, et al. Complementary-label learning for arbitrary losses and models[C]//Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 2971-2980.
- [30] Gao Yi, Zhang Minling. Discriminative complementary-label learning with weighted loss[C]//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 3587-3597.
- [31] Yu Xiyu, Liu Tongliang, Gong Mingming, et al. Learning with biased complementary labels[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 69-85.
- [32] 李自强, 杨薇, 杨先凤, 等. 基于弱标签争议的半自动分类数据标注方法[J]. 电子学报, 2024, 52(8): 2891-2899.
- Li Ziqiang, Yang Wei, Yang Xianfeng, et al. The semi-automatic classification data labeling method based on dispute about weak label[J]. Acta Electronica Sinica, 2024, 52(8): 2891-2899. (in Chinese)
- [33] DeepSeek-AI, Zhu Qihao, Guo Daya, et al. DeepSeek-coder-V2: Breaking the barrier of closed-source models in code intelligence[PP/OL]. V1. arXiv (2024-06-17) [2025-08-20]. <https://doi.org/10.48550/arXiv.2406.11931>.
- [34] Krizhevsky A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.
- [35] Le Ya, Yang Xuan. Tiny ImageNet visual recognition challenge[R]. Stanford: Stanford University, 2015.
- [36] Li Feifei, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 2007, 106(1): 59-70.
- [37] Bossard L, Guillaumin M, Van Gool L. Food-101-mining

- discriminative components with random forests[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2014: 446-461.
- [38] Krause J, Stark M, Jia Deng, et al. 3D object representations for fine-grained categorization[C]//2013 IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE, 2013: 554-561.
- [39] Cimpoi M, Maji S, Kokkinos I, et al. Describing textures in the wild[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 3606-3613.
- [40] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[PP/OL]. V1.arXiv (2013-06-21)[2025-08-20]. <https://doi.org/10.48550/arXiv.1306.5151>.
- [41] Hendrycks D, Zhao K, Basart S, et al. Natural adversarial examples[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 15257-15266.
- [42] Hendrycks D, Basart S, Mu N, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 8320-8329.
- [43] Cao Kaidi, Brbic M, Leskovec J. Open-world semi-supervised learning[C]//Proceedings of the 10th International Conference on Learning Representations. ICLR, 2022: 1-19.
- [44] Guo Lanzhe, Zhang Yige, Wu Zhifan, et al. Robust semi-supervised learning when not all classes have labels[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2022: 3305-3317.
- [45] Xia Shiyu, Lv Jiaqi, Xu Ning, et al. Towards effective visual representations for partial-label learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 15589-15598.
- [46] Wang Wei, Ishida T, Zhang Yujie, et al. Learning with complementary labels revisited: The selected-completely-at-random setting is more practical[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 50683-50710.
- [47] Gan Kai, Wei Tong. Erasing the bias: Fine-tuning foundation models for semi-supervised learning[C]//Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 14453-14470.
- [48] Zhang Renrui, Zhang Wei, Fang Rongyao, et al. Tip-adapter: Training-free adaption of CLIP for few-shot classification[C]//Proceedings of the 17th European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 493-510.
- [49] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]//Proceedings of the 7th International Conference on Learning Representations. OpenReview.net, 2019: 1-18.

作者简介



韩日东 男,1995年12月出生于黑龙江省绥化市。2024年博士毕业于吉林大学计算机科学与技术学院。现为中国矿业大学计算机科学与技术学院/人工智能学院讲师。主要研究方向为机器学习、自然语言处理。

E-mail: hanrd@cumt.edu.cn



许靖昊 男,2000年6月出生于江苏省徐州市。2025年硕士毕业于中国矿业大学计算机科学与技术学院/人工智能学院。主要研究方向为弱监督学习、预训练模型的提示学习。

E-mail: jinghaoxu@cumt.edu.cn



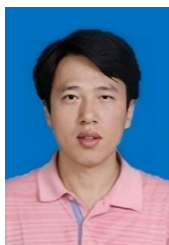
李仲年 男,1990年5月出生于江苏省徐州市。现为中国矿业大学计算机科学与技术学院/人工智能学院预聘副教授、硕士生导师。主要研究方向为机器学习、医学影像处理。

E-mail: zhongnianli@cumt.edu.cn



孙统风 男,1977年4月出生于江苏省徐州市。现为中国矿业大学计算机科学与技术学院/人工智能学院副教授、硕士生导师。主要研究方向为深度学习、知识重检索。

E-mail: suntf@cumt.edu.cn



许新征 男,1980年8月出生于安徽省宿州市。现为中国矿业大学计算机科学与技术学院/人工智能学院教授、博士生导师。主要研究方向为机器学习与数据挖掘、人工智能与模式识别、医学图像处理等。

E-mail: xxzheng@cumt.edu.cn