DOI:10.16356/j.1005-2615.2025.05.007

视觉语义增强的联合小样本开集识别分类器

丁相舒[1,2], 耿传兴[1,2], 陈松灿[1,2]
(1.南京航空航天大学计算机科学与技术学院,南京 211106; 2.模式分析与机器智能工业和信息化部重点实验室,南京 211106)

摘要: 探究了视觉-语言预训练模型对比语言-图像预训练(Contrastive language-image pre-training,CLIP)在小样本开集识别(Few-shot open-set recognition, FSOR)任务中的潜力。实验发现基于CLIP图像编码特征的视觉原型分类器通常不如传统FSOR基线方法;基于CLIP语义编码特征的语义原型分类器虽然在闭集分类上显著优于传统基线,但在开集识别方面表现不佳。本文分析造成这些问题的主要原因可能是CLIP的训练数据与FSOR目标数据之间的分布差异及CLIP语义原型分类器为已知类别划分了过大的决策边界。本文提出了一种简单有效的视觉语义增强的联合小样本开集分类器,其不仅充分利用CLIP语义原型分类器的闭集分类优势,还巧妙挖掘了传统FSOR预训练模型构建的视觉原型分类器的潜力,以更紧密的决策边界进一步提升开集识别的精准度。在4个基准数据集上的实验结果表明,该方法在准确率(Accuracy,ACC)和受试者工作特征曲线下的面积(Area under the receiver operating characteristic,AUROC)指标上相比最优基线平均提升了2.9%和2.6%。

关键词: 小样本开集识别;视觉-语言模型;原型分类器;分布差异;决策边界
中图分类号: TP391 **文献标志码:** A **文章编号:** 1005-2615(2025)05-0861-09

Visual-Semantic Enhanced Joint Classifier for Few-Shot Open-Set Recognition

DING Xiangshu[1,2], *GENG Chuanxing*[1,2], *CHEN Songcan*[1,2]
(1. College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China; 2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China)

Abstract: This paper investigates the potential of the vision-language pretrained model contrastive language-image pre-training (CLIP) in few-shot open-set recognition (FSOR). The experiments reveal that the visual-prototype classifier based on CLIP's image encoding features generally performs worse than traditional FSOR baseline methods; although the semantic-prototype classifier based on CLIP's semantic encoding features significantly outperforms traditional baselines in closed-set performance, it underperforms in open-set performance. The primary reasons for these issues may be the gap between CLIP's training data and the FSOR target data, as well as the overly large decision boundaries assigned by the CLIP semantic prototype classifier to known classes. To tackle these problems, a simple yet effective joint few-shot open-set classifier enhanced with visual-semantic integration is proposed, which not only leverages the closed-set classification advantages of the semantic-prototype classifier based on CLIP but also skillfully exploits the potential of the visual-prototype classifier constructed by traditional FSOR pretrained models, which further enhances the open-set performance by establishing tighter decision boundaries. Experiments on four benchmark datasets

基金项目: 江苏省自然科学青年基金(BK20210292)。
收稿日期: 2024-08-26;**修订日期:** 2024-11-25
通信作者: 陈松灿,男,教授,博士生导师,E-mail: s.chen@nuaa.edu.cn。

引用格式: 丁相舒,耿传兴,陈松灿. 视觉语义增强的联合小样本开集识别分类器[J]. 南京航空航天大学学报(自然科学版),2025,57(5):861-869. DING Xiangshu, GENG Chuanxing, CHEN Songcan. Visual-semantic enhanced joint classifier for few-shot open-set recognition[J]. Journal of Nanjing University of Aeronautics & Astronautics(Natural Science Edition),2025, 57(5):861-869.

demonstrate that this method achieves average improvements of 2.9% in accuracy (ACC) and 2.6% in area under the receiver operating characteristic (AUROC) compared to the best traditional baselines.

Key words: few-shot open-set recognition(FSOR); vision-language model; prototype classifier; distribution discrepancy; decision boundary

小样本学习(Few-shot learning, FSL)旨在模仿人类从少量样本中学习并快速推广到新概念的能力,近年来取得了显著进展[1-3]。然而,传统的小样本学习方法主要遵循封闭集假设,即训练样本和测试样本均来自相同的标记空间,因此难以适用于测试样本可能来自未知类别的开放场景。为了解决这一挑战,小样本开集识别(Few-shot open-set recognition, FSOR)作为小样本学习和开集识别(Open-set recognition, OSR)的交集应运而生,引起了广泛的关注[4-9]。

小样本开集识别的目标是利用极少量的训练样本,快速适应新的分类任务,在准确分类已知类别样本的同时有效识别并拒绝未知类别的样本,这要求模型具有高度有效的特征表示能力[4]。近年来,在大规模的互联网数据集上预训练的视觉-语言模型(Vision-language models, VLMs)[10-12],如对比语言-图像预训练(Contrastive language-image pre-training, CLIP)[10],在目标检测[13]和视频理解[14]等多种下游视觉任务中表现出色,展示了强大的表示学习能力,这为FSOR的发展提供了新的可能性。

尽管视觉-语言模型在FSOR中具有巨大的应用潜力,目前针对这一交叉领域的研究仍然较为有限。为填补这一空白,本文以CLIP作为研究对象,探讨了其在FSOR中的应用,通过两种无需训练的原型分类器进行了研究。一种是通过CLIP的图像编码器结合支撑样本生成的视觉原型分类器,充分利用CLIP的视觉表示能力,另一种是通过CLIP的语义编码器结合文本提示生成的语义原型分类器,利用其文本语义理解能力增强图像识别,如图1所示。实验结果显示,视觉原型分类器通常不及传统最优方法,而语义原型分类器虽然在闭集分类方面显著优于传统最优方法,但在开集识别性能方面表现不佳。对于前者,主要原因可能是CLIP的训练数据与FSOR目标任务数据之间的分布差异。对于后者,语义原型不依赖于数据分布的特性使得分类器为已知类别形成了更广泛的决策边界,这在闭集设定下具有优势,但对需要紧凑决策边界以有效拒绝未知类别的开集场景却是不利的[15]。上述两种原型分类器决策边界的对比如图2所示,其中蓝色和绿色的圆分别代表了语义原型分类器和视觉原型分类器的决策边界,其半径之比由图1中的余弦相似度获得。可以看出,语义原型分类器较宽的决策边界虽然能够更好地分类已知类别样本,容纳更大的类内差异性,但也削弱了模型区分已知和未知类别的能力,增加了混淆的风险。

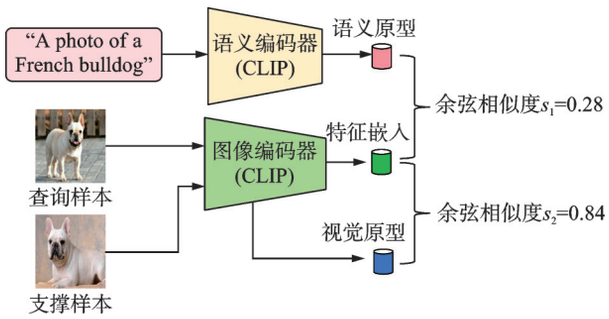

图1 两种原型分类器示意图

Fig.1 Overview of the two prototype classifiers

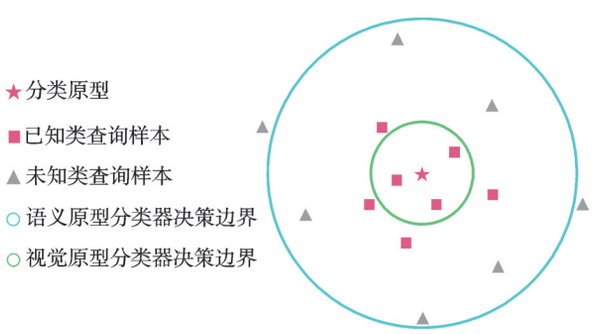

图2 两种原型分类器决策边界对比

Fig.2 Comparison of decision boundaries of two prototype classifiers

为了解决这些问题,本文提出了一种简单有效且无需训练的视觉语义增强的联合小样本开集分类器(Visual-semantic enhanced joint classifier, VS-JC),巧妙结合了两种不同的原型分类器。一种是基于文本提示增强的CLIP语义原型分类器,具有出色的闭集分类能力;另一种是利用传统FSOR预训练模型构建的先验增强的视觉原型分类器,包含更多对目标任务数据分布的先验信息,能够为已知类别提供更紧凑的决策边界。通过两者的有机结合实现优势互补,获得了性能的双重提升。

1 相关工作

1.1 小样本开集识别

作为小样本学习与开集识别的交叉领域,小样本开集识别因其内在的挑战性和实际意义,正逐渐成为研究的热点。现有方法大致可以分为两类:基于阈值的方法和基于负原型的方法。基于阈值的

方法通常设置分类置信度阈值，当模型预测的置信度低于该阈值时，便将该样本视为未知类。Liu等[4]借鉴并扩展了在小样本学习中已被证明有效的元学习策略，同时提出了一种基于最大熵的开集识别损失，显著增强了模型对未知类别的鲁棒性。Wang等[5]提出了一种基于全局能量分数的阈值方法，该方法通过计算类间和像素间的相似性来获得能量得分，并利用该得分设置阈值，从而实现对未知类别样本的有效识别。CHE等[6]引入了一种即插即用的多关系边缘训练损失，以更好地捕捉已知类和未知类的分布特征。Jeong等[9]利用变换一致性原理，通过比较原始原型与经过变换后将查询样本特征替换的原型之间的距离差异以有效地检测未知类别样本。基于负原型的方法为未知类生成一个或多个分类原型，将原本的N类开集识别任务转换为$(N+1)$类闭集分类任务。Huang等[7]创新性地引入了自注意力机制，通过将已知类原型输入Transformer模型以生成任务自适应的未知类原型，显著提升了模型在动态任务场景下的表现。Song等[8]则提出了一种独特的训练策略，将已知类别样本的背景区域视为未知类别，通过对背景分布的建模，进一步丰富了分类器对未知类别的区分能力。

1.2　视觉-语言模型

近年来，将图像和文本映射到共同特征空间进行多模态表示学习的VLMs在计算机视觉领域引起了广泛关注。其中，CLIP[10]是一个具有代表性的模型，它通过自监督对比学习，在包含数百万图像文本对的互联网数据集上训练，并在各种下游任务中取得了出色的表现[16-18]。许多后续研究致力于开发更有效的方法，以便更好地将CLIP应用于下游的小样本分类任务。例如，Zhou等[19]提出用可学习的文本标记替代传统手工设计的文本提示词，通过小样本训练数据进行微调，以自适应生成更符合任务需求的文本提示。Zhou等[20]在此基础上加入了一个轻量级网络，为每个输入图像生成条件标记，从而动态适应不同图像的内容，有效缓解了对基类的过拟合问题。Gao等[21]引入了一个轻量级适配模块，该模块通过利用目标任务中的训练数据进行微调，从而生成更加适配任务需求的特征表示。Zhang等[22]通过引入键值缓存模型，在微调过程中加速了模型的收敛速度，以实现更高效的训练过程。为了充分挖掘更多的先验知识，Zhang等[23]将CLIP与DALL-E[11]和DINO[24]等预训练模型进行级联，通过集成这些强大的基础模型，有效提升了模型的泛化能力。He等[25]在CLIP的图像编码器中集成了一个基于Transformer的网络模块，并通过微调以生成判别性强且任务自适应的原型特征，提升了模型在特定任务中的表示能力。

尽管上述研究在小样本学习领域取得了显著进展，但基于VLMs的研究中，专门针对FSOR问题的工作仍然较少。因此本文以CLIP为例，深入探讨其在FSOR问题中的应用潜力。

2　问题定义

FSOR旨在通过有限的标注数据，既能精确分类已知类别样本，又能有效检测和识别未知类别样本。在一个具体的FSOR任务中，假设总类别数为N，每个类别的支撑样本（训练样本）数为K，则该任务被称为N-way K-shot任务。具体地，一个N-way K-shot的FSOR任务可以表示为$G=(D^{\mathrm{S}},D^{\mathrm{Q}})$，其中$D^{\mathrm{S}}=\{\boldsymbol{x}_i^{\mathrm{s}}\in\boldsymbol{X}^{\mathrm{S}},y_i^{\mathrm{s}}\in Y^{\mathrm{S}}\}_{i=1}^{N\times K}$表示支撑集（训练集），$D^{\mathrm{Q}}=\{\boldsymbol{x}_i^{\mathrm{q}}\in\boldsymbol{X}^{\mathrm{Q}},y_i^{\mathrm{q}}\in Y^{\mathrm{Q}}\}_{i=1}^{m}$表示查询集（测试集），$m$为查询样本的总数。与传统闭集小样本学习不同，查询集D^{Q}不仅包括已知类别的查询集$D^{\mathrm{K}}=\{\boldsymbol{x}_i^{\mathrm{k}}\in\boldsymbol{X}^{\mathrm{K}},y_i^{\mathrm{k}}\in Y^{\mathrm{K}}\}_{i=1}^{N\times n}$，还包括未知类别的查询集$D^{\mathrm{U}}=\{\boldsymbol{x}_i^{\mathrm{u}}\in\boldsymbol{X}^{\mathrm{U}},y_i^{\mathrm{u}}\in Y^{\mathrm{U}}\}_{i=1}^{N^{U}\times n}$，其中$Y^{\mathrm{S}}\cap Y^{\mathrm{U}}=\varnothing$，$N^{\mathrm{U}}$和$n$分别表示未知类别的总数和每个类别的查询样本数量。FSOR任务的目标是通过有限的支撑集D^{S}构建分类器，以准确分类来自D^{K}中的已知类查询样本，同时检测并拒绝来自D^{U}中的未知类查询样本。

3　CLIP在小样本开集识别任务中的研究

本章节主要探讨CLIP在FSOR中的表现，采用了两种无需训练的原型分类器：视觉原型分类器（CLIP-vis）和语义原型分类器（CLIP-sem）。

3.1　视觉原型分类器CLIP-vis

CLIP-vis是一个利用CLIP的图像编码器F_{img}和支撑样本构建的视觉原型分类器。具体来说，给定一个N-way K-shot的FSOR任务，首先通过支撑样本计算每个已知类别c_i的视觉分类原型。具体步骤是将类别c_i的支撑样本输入CLIP的图像编码器，对获得的所有特征嵌入进行平均，得到的特征即为视觉原型$\boldsymbol{V}_i$，表示为

$$\boldsymbol{V}_i=\frac{1}{K}\sum_{j=1}^{K}F_{\mathrm{img}}(\boldsymbol{x}_{i,j}^{\mathrm{s}}) \tag{1}$$

在获得每个已知类别的视觉原型后，通过比较它们与查询样本特征之间的余弦相似度进行分类。具体而言，查询样本$\boldsymbol{x}^{\mathrm{q}}$属于类别$c_i$的概率计算为

$$P(c_i|\boldsymbol{x}^{\mathrm{q}})=\frac{\exp(\cos(F_{\mathrm{img}}(\boldsymbol{x}^{\mathrm{q}}),\boldsymbol{V}_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(F_{\mathrm{img}}(\boldsymbol{x}^{\mathrm{q}}),\boldsymbol{V}_j)/\tau)} \quad (2)$$

式中:cos表示余弦相似度;τ为一个温度缩放因子,设置为0.01,与CLIP的设定一致。

3.2 语义原型分类器CLIP-sem

CLIP-sem是由CLIP的语义编码器F_{text}结合人工编写的文本提示构建的语义原型分类器。具体来说,对于每个已知类别c_i,首先生成一个形式为"a photo of a [CLASS]"的文本提示T_i,其中[CLASS]对应该类别的具体名称,如"cat""plane"或"apple"。然后将该文本提示输入CLIP的语义编码器,以获得该类别的语义分类原型$\boldsymbol{S}_i$,即

$$\boldsymbol{S}_i=F_{\mathrm{text}}(T_i) \quad (3)$$

通过比较语义原型与查询样本特征之间的余弦相似度,查询样本$\boldsymbol{x}^{\mathrm{q}}$属于类别$c_i$的概率计算为

$$P(c_i|\boldsymbol{x}^{\mathrm{q}})=\frac{\exp(\cos(F_{\mathrm{text}}(\boldsymbol{x}^{\mathrm{q}}),\boldsymbol{S}_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(F_{\mathrm{text}}(\boldsymbol{x}^{\mathrm{q}}),\boldsymbol{S}_j)/\tau)} \quad (4)$$

3.3 案例研究

以两个FSOR基准数据集TieredImageNet[26]和CIFAR-FS[27]为例对CLIP-vis和CLIP-sem的应用潜力进行了探究,结果如图3所示。与传统基线方法GEL[5]相比,CLIP-vis在TieredImageNet数据集5-way 1-shot任务上的闭集分类ACC和开集识别AUROC指标均明显落后,在CIFAR-FS数据集5-way 5-shot任务中同样表现不佳,这可能是由于CLIP的训练数据与FSOR目标任务数据之间存在分布差异。尽管CLIP-sem在两种任务设定下的ACC指标均明显优于GEL,但在AUROC指标上却明显不足,其主要原因可能是未能为已知类别建立紧凑的决策边界,这表明可能需要获得更多的数据分布信息以进一步提升性能。

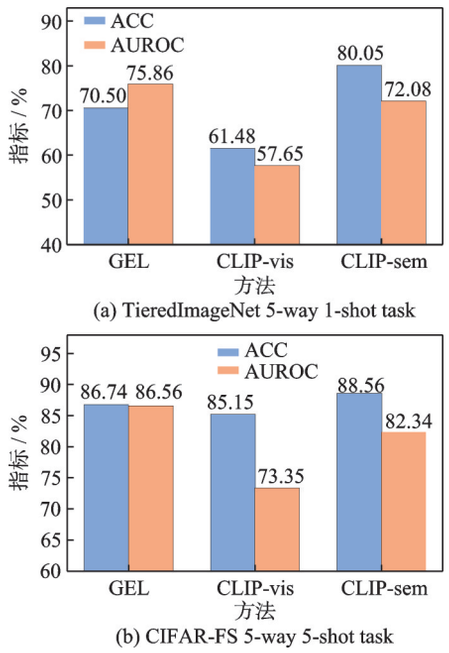

图3 两种原型分类器与传统基线GEL在不同任务设定下的性能对比

Fig.3 Performance comparison of two prototype classifiers with GEL across different settings

4 视觉语义增强的联合小样本开集分类器

为了克服上述的局限性,本文提出了一种简单有效且无需训练的VSJC,其核心思想是将基于CLIP的语义原型分类器和传统FSOR预训练模型的优势充分结合。前者在闭集分类中表现出色,而后者对目标任务的数据分布有更丰富的先验知识,能够为已知类别形成更紧凑的决策边界,从而提升开集识别性能。为实现这一目标,该方法结合了经过文本提示增强的CLIP语义原型分类器与由传统FSOR预训练模型生成的先验增强的视觉原型分类器,进行联合推理。图4展示了方法的整体框架。

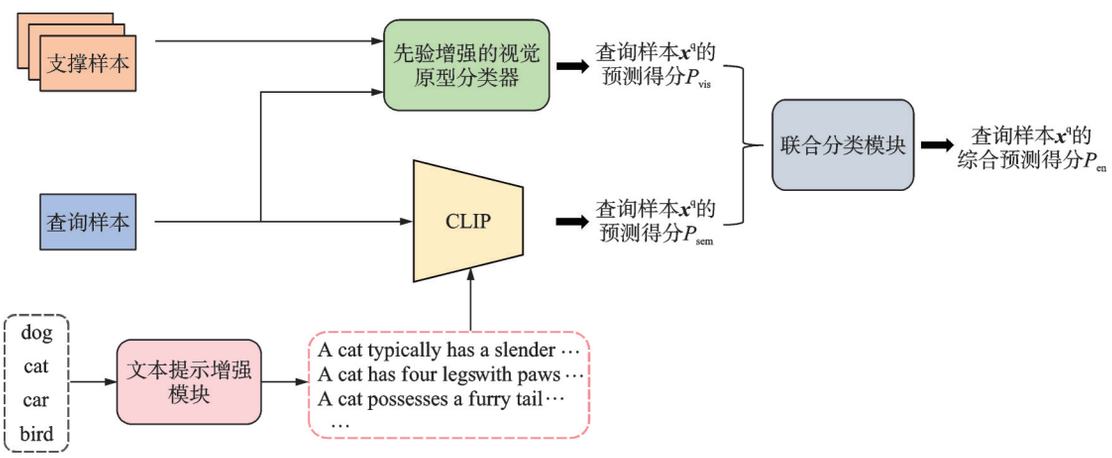

图4 视觉语义增强的联合小样本开集分类器方法框架图

Fig.4 Framework of VSJC method

4.1 先验增强的视觉原型分类器

鉴于CLIP-vis由于分布偏移问题而表现不佳，本文未将其作为VSJC的视觉原型分类器。相反，本文旨在挖掘传统FSOR预训练模型的潜力来缓解数据分布差异。具体而言，本文使用在基类数据集上预训练的图像编码器F_{base}来构建视觉原型分类器，充分利用其对目标任务数据分布的先验知识以增强性能。本文提出了一种先验增强的视觉原型分类器FSOR-vis。给定1个N-way K-shot的FSOR任务，首先使用支撑样本计算每个类别c_i的的视觉分类原型$\boldsymbol{V}_i$，即

$$\boldsymbol{V}_i=\frac{1}{K}\sum_{j=1}^{K}F_{\text{base}}(\boldsymbol{x}_{i,j}^{\text{s}}) \tag{5}$$

然后，通过计算查询样本特征与各类别视觉原型之间的余弦相似度，得到其预测得分。具体而言，对查询样本$\boldsymbol{x}^{\text{q}}$的预测得分$\boldsymbol{P}_{\text{vis}}$可以表示为

$$\boldsymbol{P}_{\text{vis}}=\{P_{\text{vis}}^{i}\}_{i=1}^{N} \tag{6}$$

式中：P_{vis}^{i}表示查询样本$\boldsymbol{x}^{\text{q}}$属于第$i$个类别的预测得分，具体计算为

$$P_{\text{vis}}^{i}=\cos(F_{\text{base}}(\boldsymbol{x}^{\text{q}}),\boldsymbol{V}_i) \tag{7}$$

4.2 文本提示增强的语义原型分类器

为了更好地将CLIP-sem应用于FSOR问题，本文在其基础上引入了一个文本提示增强模块，利用语言模型GPT-3[28]来生成更细致的文本提示，这有利于丰富上下文语义信息，获得更具代表性的语义原型，从而形成更紧凑的决策边界。本文提出了一种文本提示增强的语义原型分类器。具体而言，给定一个N-way K-shot的FSOR任务，使用统一的模板作为输入Input（例如“what a [CLASS] looks like”）来查询GPT-3，为每个已知类别获得n_t种更加丰富具体的文本提示。将为类别c_i生成的文本提示记作$\boldsymbol{T}_i^{\text{aug}}$，有

$$\boldsymbol{T}_i^{\text{aug}}=\text{GPT-3}(\text{input}_{c_i}) \tag{8}$$

然后通过对所生成文本提示编码后的特征取均值，计算每个类别的精细化语义原型，即

$$\boldsymbol{S}_i=\frac{1}{n_t}\sum_{j=1}^{n_t}F_{\text{text}}(\boldsymbol{T}_{i,j}^{\text{aug}}) \tag{9}$$

和FSOR-vis类似，查询样本$\boldsymbol{x}^{\text{q}}$的预测得分$\boldsymbol{P}_{\text{sem}}$可以表示为

$$\boldsymbol{P}_{\text{sem}}=\{P_{\text{sem}}^{i}\}_{i=1}^{N} \tag{10}$$

$$P_{\text{sem}}^{i}=\cos(F_{\text{img}}(\boldsymbol{x}^{\text{q}}),\boldsymbol{S}_i) \tag{11}$$

4.3 视觉-语义联合推断

为充分发挥上述两种原型分类器的优势，本文将其结合进行联合推断。具体来说，通过对两个分类器的预测得分$\boldsymbol{P}_{\text{vis}}$和$\boldsymbol{P}_{\text{sem}}$进行加权求和，然后通过softmax归一化，得到查询样本$\boldsymbol{x}^{\text{q}}$的综合预测得分$\boldsymbol{P}_{\text{en}}=\{P_{\text{en}}^{i}\}_{i=1}^{N}$，具体计算为

$$\boldsymbol{P}_{\text{en}}=\text{softmax}(\alpha\cdot\boldsymbol{P}_{\text{vis}}+(1-\alpha)\cdot\boldsymbol{P}_{\text{sem}}) \tag{12}$$

式中：α为一个取值范围在0～1的权重超参数。

最后，通过对综合预测得分$\boldsymbol{P}_{\text{en}}$进行阈值判断，确定查询样本$\boldsymbol{x}^{\text{q}}$的最终分类结果$y^{\text{q}}$。将阈值表示为$\eta$，有

$$y^{\text{q}}=\begin{cases}\hat{n} & P_{\text{en}}^{\hat{n}}>\eta\\ \text{unknown} & \text{Otherwise}\end{cases} \tag{13}$$

$$\hat{n}=\text{argmax}_i\{P_{\text{en}}^{i}\}_{i=1}^{N} \tag{14}$$

5 实验与分析

5.1 实验设置

5.1.1 数据集与评估指标

为验证所提方法的有效性，本文在FSOR领域广泛认可的4个基准数据集：MiniImageNet[29]、TieredImageNet[26]、CIFAR-FS[27]和CUB[30]上进行了实验。本文采用准确率（Accuracy，ACC）和受试者工作特征曲线下的面积（Area under the receiver operating characteristic，AUROC）作为主要评估指标，其中ACC用于衡量模型对已知类查询样本的闭集分类准确度，AUROC用于评估模型检测未知类别的能力。根据现有研究的设置[4-9]，本文在600个5-way 1-shot或5-way 5-shot的FSOR任务上对模型进行测试，并计算平均ACC和AUROC指标。

5.1.2 实现细节

本文采用传统方法ATT-G[7]中基于残差网络（Residual network12，ResNet-12）[31]的预训练模型作为FSOR-vis中的图像编码器F_{base}。对于CLIP[10]，使用预训练的ResNet-50[31]作为图像编码器F_{img}，并使用与之对应的Transformer作为文本编码器F_{text}。对于文本提示增强策略，将由GPT-3[28]为每个类别生成的文本提示数量n_t设置为20。关于权重超参数α，其选择通过在验证集上进行网格搜索确定。对于5-way 1-shot任务，在MiniImageNet、TieredImageNet、CIFAR-FS和CUB数据集上的设置分别为0.04、0.15、0.06和0.07；对于5-way 5-shot任务，分别设置为0.07、0.28、0.1和0.05。

5.2 方法对比与分析

5.2.1 与无需训练的方法对比

为了验证本文方法的优势，将VSJC与两种典型的无需训练的方法CLIP-vis和CLIP-sem，以及6种经典的传统FSOR方法（PEELER[4]，SnaTCHer[9]，ATT-G[7]，MRM[6]和GEL[5]，FSOR-vis）进行

了比较,结果如表1和表2所示。通过结合由传统FSOR预训练模型构建的先验增强的视觉原型分类器FSOR-vis进行联合推理,VSJC在ACC和AUROC方面相比CLIP-sem取得了显著提升。例如,在TieredImageNet 5-way 1-shot任务中,VSJC在ACC上提升了5.5%,在AUROC上提升了6.9%;在CIFAR-FS 5-way 1-shot任务中,VSJC在ACC和AUROC上分别达到92.14%和86.51%,相较于CLIP-sem分别提高了3.6%和4.2%,充分体现了联合推理策略在小样本开集识别任务中的有效性。

表1 在MiniImageNet和TieredImageNet上与无需训练方法的对比

Table 1 Comparison results with training-free methods on MiniImageNet and TieredImageNet %

方法	MiniImageNet				TieredImageNet			
	1-shot		5-shot		1-shot		5-shot	
	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC
PEELER[4]	65.86	60.57	80.61	67.35	69.51	65.20	84.10	73.27
SnaTCHer[9]	67.02	68.27	82.02	72.42	70.52	74.28	84.74	82.02
ATT-G[7]	68.11	72.41	83.12	79.85	70.58	73.43	85.38	81.64
MRM[6]	67.03	71.20	82.00	80.39	71.13	75.59	85.27	**83.03**
GEL[5]	68.26	73.70	83.05	82.29	70.50	75.86	84.60	81.95
FSOR-vis	65.38	63.64	83.65	73.37	68.99	67.71	84.67	75.71
CLIP-vis	80.30	67.47	93.34	75.80	61.48	57.65	79.61	64.63
CLIP-sem	96.82	92.23	96.82	92.23	80.05	72.08	80.05	72.08
VSJC	**97.56**	**94.74**	**97.80**	**94.94**	**85.57**	**78.95**	**90.21**	82.43

表2 在CIFAR-FS和CUB上与无需训练方法的对比

Table 2 Comparison results with training-free methods on CIFAR-FS and CUB %

方法	CIFAR-FS				CUB			
	1-shot		5-shot		1-shot		5-shot	
	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC
PEELER[4]	71.47	71.28	85.46	75.97	62.62	57.26	82.44	65.01
SnaTCHer[9]	75.09	78.15	87.18	85.81	69.03	74.34	83.77	84.57
ATT-G[7]	72.43	76.72	86.52	84.64	79.27	81.73	90.02	90.01
MRM[6]	—	—	—	—	70.00	76.30	84.51	85.91
GEL[5]	76.77	78.67	86.74	86.56	79.00	81.78	89.75	89.95
FSOR-vis	70.45	70.46	86.47	80.35	67.54	68.10	85.38	82.72
CLIP-vis	65.25	61.67	85.15	73.35	78.67	66.69	94.61	79.24
CLIP-sem	88.56	82.34	88.56	82.34	95.63	81.98	95.63	81.98
VSJC	**92.14**	**86.51**	**94.30**	**88.65**	**96.20**	**88.39**	**97.16**	**90.41**

此外,与传统FSOR方法相比,VSJC在大多数数据集上均表现出显著的性能优势。例如,与传统最优方法GEL和ATT-G相比,VSJC在4个基准数据集上的ACC和AUROC指标分别平均提升了14%、6.8%和14.4%、8.1%。这些结果充分证明了VSJC在小样本开集识别任务中的卓越性能与广泛适用性。

5.2.2 与基于微调的方法对比

为了进一步证明本文方法的有效性,将两种流行的CLIP微调策略(CLIP-Adapter[21]和Tip-Adapter[22])适配到CLIP-sem上,并进行了对比分析,如表3和表4所示。结果显示,尽管CLIP-Adapter和Tip-Adapter通过在有限的支撑集上进行微调可以显著提升性能,但它们在4个基准数据集上的表现仍然落后于本文提出的VSJC方法,并且后者无需额外的训练开销。例如,在TieredImageNet数据集上,VSJC的ACC相较于CLIP-Adapter和Tip-Adapter分别平均提升了4.5%和5.9%,AUROC分别提升了7.0%和7.7%;在CIFAR-FS数据集上,VSJC的ACC和AUROC分别相较于CLIP-Adapter提高了3.2%和3.8%,相较于Tip-Adapter提高了3.4%和4.1%。这些结果进一步验证了VSJC在分类精度和开集识别性能方面的显著优势。

5.3 消融实验

为了评估所提出的VSJC模型各个组件的有效性,本文在TieredImageNet的5-way 1-shot任务和CIFAR-FS的5-way 1-shot任务上进行了消融

表 3　MiniImageNet 和 TieredImageNet 上与基于微调方法的对比

Table 3　Comparison results with fintuning methods on MiniImageNet and TieredImageNet　%

方 法	MiniImageNet				TieredImageNet			
	1-shot		5-shot		1-shot		5-shot	
	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC
CLIP-sem	96.82	92.23	96.82	92.23	80.05	72.08	80.05	72.08
CLIP-Adapter	96.99	92.37	97.08	92.45	81.85	72.82	85.03	74.63
Tip-Adapter	97.19	92.60	97.25	92.70	81.39	72.50	82.56	73.45
VSJC	**97.56**	**94.74**	**97.80**	**94.94**	**85.57**	**78.95**	**90.21**	**82.43**

表 4　CIFAR-FS 和 CUB 上与基于微调方法的对比

Table 4　Comparison results with fintuning methods on CIFAR-FS and CUB　%

方 法	CIFAR-FS				CUB			
	1-shot		5-shot		1-shot		5-shot	
	ACC	AUROC	ACC	AUROC	ACC	AUROC	ACC	AUROC
CLIP-sem	88.56	82.34	88.56	82.34	95.63	81.98	95.63	81.98
CLIP-Adapter	89.62	83.41	90.38	84.20	95.79	82.13	96.16	87.25
Tip-Adapter	89.50	83.23	90.18	83.78	95.96	82.25	96.09	82.30
VSJC	**92.14**	**86.51**	**94.30**	**88.65**	**96.20**	**88.39**	**97.16**	**90.41**

实验，结果如表 5 和表 6 所示。

表 5　TieredImageNet 5-way 1-shot 任务上的消融实验

Table 5　Ablation experiment on TieredImageNet 5-way 1-shot task

组成部分			性能	
CLIP-sem	文本提示增强	联合推断	ACC/%	AUROC/%
√	—	—	80.05	72.08
√	√	—	81.09	72.71
√	—	√	85.08	78.67
√	√	√	85.57	78.95

表 6　CIFAR-FS 5-way 1-shot 任务上的消融实验

Table 6　Ablation experiment on CIFAR-FS 5-way 1-shot task

组成部分			性能	
CLIP-sem	文本提示增强	联合推断	ACC/%	AUROC/%
√	—	—	88.56	82.34
√	√	—	90.10	83.88
√	—	√	91.22	86.20
√	√	√	92.14	86.51

实验表明，文本提示增强与联合推断策略均能有效提升闭集分类准确率 ACC 和开集识别性能 AUROC。具体而言，文本提示增强策略通过借助 GPT-3 丰富语义上下文信息，在不降低 ACC 的前提下，使 AUROC 在两项任务中平均提升了 1.1%。同时，通过联合推断策略，充分结合两种原型分类器的优势，ACC 和 AUROC 指标分别在两项任务中平均提升了 3.8% 和 5.2%。此外，结合这两种策略可以实现最佳性能表现。

5.4　超参数敏感性分析

为了评估 VSJC 模型对于权重超参数 α 的敏感性，本文在 MiniImageNet 的 5-way 5-shot 任务和 CIFAR-FS 的 5-way 1-shot 任务上进行了实验，其中 α 的值分别被设置为 {0.05, 0.06, 0.07, 0.08, 0.09} 和 {0.04, 0.05, 0.06, 0.07, 0.08}，结果如图 5 所示。实验结果表明，在给定范围内，权重超参数 α 的选择对性能的影响总体上并不显著。然而，精细调整 α 的取值仍然能够带来性能的提升。

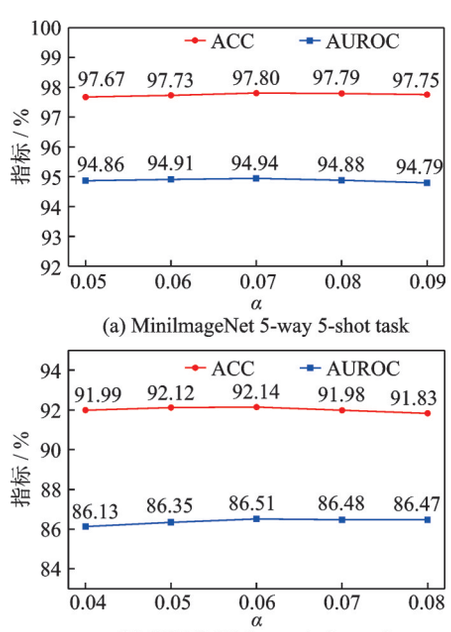

图 5　在不同任务下对超参数 α 的敏感性分析

Fig.5　Sensitivity analysis of hyperparameter α across different tasks

6 结 论

本文探究了视觉-语言预训练模型CLIP在FSOR中的应用潜力,实验发现:

(1)基于CLIP图像编码特征的视觉原型分类器通常不如传统FSOR基线方法,这可能是由于CLIP的预训练数据集与FSOR目标任务数据集之间存在数据分布差异。

(2)基于CLIP语义编码特征的语义原型分类器虽然在闭集分类上显著优于传统基线,但开集识别性能不佳,这可能是因为语义原型分类器为已知类别划分了过大的决策边界。

针对这些问题,本文提出了一种简单有效且无需训练的视觉语义增强的联合小样本开集分类器。该方法充分利用了基于文本提示增强的CLIP语义原型分类器在闭集分类中的优势,同时发挥了由传统FSOR预训练模型构建的先验增强的视觉原型分类器的潜力,通过其对目标任务数据的先验知识,缓解数据分布差异,并为已知类别建立更紧密的决策边界,从而更好地检测未知类别,实现性能的双重提升。实验结果表明,本文所提出的方法在ACC和AUROC指标上取得了显著的提升,并在大多数数据集上取得了较先进的性能。

参考文献:

[1] 张玉, 尚志华, 郭晓楠, 等. 小样本图像分类中的类别信息融合网络[J]. 南京航空航天大学学报, 2022, 54(4): 715-722.
ZHANG Yu, SHANG Zhihua, GUO Xiaonan, et al. A category information fusion network for few-shot image classification[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(4): 715-722.

[2] 徐惠灵, 尚政国, 董胜波, 等. 面向深度神经网络应用的小样本学习技术研究[J]. 南京航空航天大学学报, 2022, 54(S1): 80-86.
XU Huiling, SHANG Zhengguo, DONG Shengbo, et al. A study on few-shot learning techniques for deep neural network applications[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(S1): 80-86.

[3] ZHOU Y, HAO J, HUO S, et al. Automatic metric search for few-shot learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023,35(7): 10098-10109.

[4] LIU B, KANG H, LI H, et al. Few-shot open-set recognition using meta-learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2020.

[5] WANG H, PANG G, WANG P, et al. Glocal energy-based learning for few-shot open-set recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2023: 7507-7516.

[6] CHE Y, AN Y, XUE H. Boosting few-shot open-set recognition with multi-relation margin loss[C]//Proceedings of the IJCAI. China: IJCAI, 2023: 3505-3513.

[7] HUANG S, MA J, HAN G, et al. Task-adaptive negative envision for few-shot open-set recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2022: 7171-7180.

[8] SONG N, ZHANG C, LIN G. Few-shot open-set recognition using background as unknowns[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5970-5979.

[9] JEONG M, CHOI S, KIM C. Few-shot open-set recognition by transformation consistency[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2021: 12566-12575.

[10] 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. New York: PMLR, 2021: 8748-8763.

[11] REDDY M D M, BASHA M S M, HARI M M C, et al. DALL-E: Creating images from text[J]. UGC Care Group I Journal, 2021, 8(14): 71-75.

[12] WU X, ZHU F, ZHAO R, et al. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2023: 7031-7040.

[13] ZHAO S, ZHANG Z, SCHULTER S, et al. Exploiting unlabeled data with vision and language models for object detection[C]//Proceedings of the European Conference on Computer Vision.Heidelberg: Springer-Verlag GmbH, 2022: 159-175.

[14] WU W, SUN Z, OUYANG W. Revisiting classifier: Transferring vision-language models for video recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC: AAAI Press, 2023.

[15] ROADY R, HAYES T L, KEMKER R, et al. Are open set classification methods effective on large-scale datasets[J]. PLoS ONE, 2020, 15(9): e0238302.

[16] WANG J, CHAN K C, LOY C C. Exploring clip for assessing the look and feel of images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC: AAAI, 2023.

[17] WASIM S T, NASEER M, KHAN S, et al. Vita-clip: Video and text adaptive clip via multimodal prompting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2023: 23034-23044.

[18] LUO H, JI L, ZHONG M, et al. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning[J]. Neurocomputing, 2022, 508(C): 293-304.

[19] ZHOU K, YANG J, LOY C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[20] ZHOU K, YANG J, LOY C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2022.

[21] GAO P, GENG S, ZHANG R, et al. Clip-adapter: Better vision-language models with feature adapters [J]. International Journal of Computer Vision, 2024, 132(2): 581-595.

[22] ZHANG R, FANG R, ZHANG W, et al. Tip-adapter: Training-free clip-adapter for better vision-language modeling[EB/OL]. (2021-11-15). https://arxiv.org/abs/2111.03930.

[23] ZHANG R, HU X, LI B, et al. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2023: 15211-15222.

[24] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos, CA: IEEE, 2021: 9650-9660.

[25] HE F, LI G, SI L, et al. Prototype Former: Learning to explore prototype relationships for few-shot image classification[EB/OL]. (2023-10-07). https://arxiv.org/abs/2310.03517.

[26] REN M, TRIANTAFILLOU E, RAVI S, et al. Meta-learning for semi-supervised few-shot classification[EB/OL]. (2018-03-02). https://arxiv.org/abs/1803.00676.

[27] KRIZHEVSKY A, NAIR V, HINTON G. Cifar-10 (canadian institute for advanced research)[EB/OL]. (2025-10-03). http://www.cs.toronto.edu/kriz/cifar.html.

[28] BROWN T B. Language models are few-shot learners [EB/OL]. (2020-05-28). https://arxiv. org/abs/2005.14165.

[29] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016: 3637-3645.

[30] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset: CNS-TR-2011-001[R].[S.l.]:[s.n.], 2011.

[31] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2016: 770-778.

（编辑：刘彦东）