基于多模态元学习的小样本图像分类方法

王 雁[1]
WANG Yan

摘 要 针对小样本图像分类任务中模态信息单一，分类准确率欠佳的问题，文章提出了一种基于多模态元学习的小样本图像分类方法。首先，利用 CLIP 模型的视觉编码器和文本编码器分别提取图像的全局视觉特征与对应类别的文本语义特征，通过投影层进行特征映射；然后，通过元学习算法优化模型参数，使模型能够在少量样本上快速学习新类别特征。在 Mini-ImageNet 数据集上进行测试，结果表明，相较于传统元学习方法，该方法分类效果较好，为小样本图像分类提供了一种有效的解决思路。

关键词 小样本学习；元学习；多模态融合；图像分类

doi：10.3969/j.issn.1672-9528.2026.05.004

0 引言

传统的深度学习方法需要依靠大量经过标注的数据，然而在面对医疗影像、古文字识别这类数据稀缺的场景时模型极易陷入过拟合导致泛化能力和鲁棒性显著下降。小样本学习（few-shot learning, FSL）[1] 是当前计算机视觉领域研究的热点方向，目的是凭借少量样本训练出具有强大泛化能力的模型。小样本学习技术的核心难点在于怎样从数量极其有限的样本中归纳出能够推广的类别概念，此项研究对于推动人工智能在数据稀缺领域的应用有着重要意义。比如，在医疗影像分析中，对于某种罕见病的诊断可能只有很少的标注样本；在工业质检中，某些特定类型的缺陷产品样本也不容易大量获取；在古生物或天文领域，识别新颖类别物体同样面临样本不足的状况。在这些场景下，传统深度学习模型极易因过拟合而失去效用，研究如何利用先验知识快速适应新任务的小样本学习方法是解决上述问题的关键所在。

元学习通过“学会学习”的机制，能够快速适应新任务，成为解决样本稀缺问题的有效途径。其中，MAML[2]（model-agnostic meta-learning）是一种具有代表性的与模型无关的元学习方法，可用于任何基于梯度下降的模型。MAML 通过双层优化机制学习模型的初始参数，使模型在新任务上仅需少量梯度更新即可适应。传统元学习方法基于单一模态特征，忽略了不同模态信息（如文本描述、语义标签等）之间的互补性，限制了分类性能的提升。

人类对于世界的感知本质上是具有多模态性质的，能够通过综合视觉、听觉、文本描述等多种不同类型的信息来理解和识别新事物。受此启发，多模态学习致力于将不同模态中的信息加以融合，以此来获得相较于任何单一模态而言更全面且更鲁棒的数据表征。早期阶段的多模态研究大多集中在特征拼接或加权融合方面。近年来借助大规模互联网数据进行预训练的多模态模型，通过对比学习在统一的嵌入空间中对齐图像和文本特征，展现了强大的零样本识别和跨模态理解能力。这为小样本学习开拓了全新的思路：利用从海量图文对中学到的丰富视觉 - 语义先验知识来辅助模型快速理解只有少量样本的新类别。

CLIP[3]（contrastive language-image pre-training）是由 OpenAI 推出的多模态预训练神经网络，其主要思路是通过大规模图像 - 文本对数据联合训练一个图像编码器和一个文本编码器，进而学习图像和文本之间的语义对齐关系。模型具有多模态学习的能力，可以同时理解图像和文本两种不同模态的信息并在它们之间建立联系。CLIP 模型具有通用性和灵活性，在多个领域都展现出了强大的应用潜力包括图像分类、图像检索、文本生成、多模态搜索等。本文选用 CLIP 模型作为核心特征提取器，主要源于它的两大优势：CLIP 是在规模超大的 4 亿个图像 - 文本对上进行训练的，通过对比学习目标函数，使其学习到具有高度的语义一致性的视觉和文本表征，即语义相似的图像和文本在嵌入空间中距离更近。这种强大的跨模态对齐能力为零样本和小样本学习奠定了坚实基础。CLIP 的模型架构具有灵活性，它的视觉编码器能够采用 Vision Transformer (ViT) 或者 ResNet 等结构，文本编码器采用 Transformer。本文采用基于 ViT-B/32 的 CLIP 模型，输出特征维度为 512 维，为后续的特征融合与元学习优化提供了高密度高语义性的特征输入。

基于此，本文提出一种基于多模态元学习的小样本图像

1. 太原学院 山西太原 030032
[基金项目] 2024 年度太原学院院级科研项目（24TYQN25）

分类方法。首先，将强大的多模态预训练模型 CLIP 作为特征提取器引入元学习框架，弥补了传统元学习方法模态单一、先验知识不足的缺陷；其次，设计了特征投影层与联合损失函数，有效地将 CLIP 的通用特征适配到小样本分类任务中，实现了预训练知识与任务特定知识的融合；最后，通过系统的实验验证了所提方法（MAML-CLIP）在 Mini-ImageNet 数据集上的优越性，并通过可视化分析证明了多模态特征在提升类间区分度和类内紧凑性方面的作用。

1 方法

针对小样本图像分类中单一视觉模态信息有限的问题，本文的核心思想是：利用大规模预训练的多模态模型（CLIP）所提供的丰富视觉 - 语义先验知识，并通过元学习（MAML）框架使其快速适配到新任务上，以此提升模型的泛化能力和分类精度。该方法的工作流程如图 1 所示，主要包括数据预处理、多模态特征提取、特征融合、元学习训练和分类与评估五个部分。

图 1 模型框架图

其主要创新点体现在：

（1）多模态特征提取与对齐：利用 CLIP 模型强大的图文对齐能力，分别提取图像的全局视觉特征和类别的文本语义特征，并通过一个可学习的投影层将它们映射到同一个任务相关的特征空间，实现更深层次的模态融合。

（2）元学习快速适配机制：采用 MAML 元学习算法，以“学会学习”为目标优化模型初始参数。使得模型在面对仅有少量样本（*K*-shot）的新类别（*N*-way）任务时，通过内循环的少量几步梯度更新，就能快速生成高性能的任务特定模型。

（3）数据预处理：本文采用 Mini-ImageNet 数据集 [4]，共有 100 个物体类别，每个类别含 600 张 84 px×84 px 的 RGB 图像，为适配 CLIP-ViT 模型的输入要求，将分辨率统一调整为 224 px×224 px。

（4）多模态特征提取和特征融合：流程图如图 2 所示。将大小为 3×224×224 的图像经过标准化后，通过基于 ViT-B/32[5] 架构的视觉编码器，获得图像特征的向量表示，然后经过投影层得到与任务相关的特征；同时，将类别描述通过文本编码器 Transformer[6] 和文本投影层，映射到与图像特征相同的多模态空间中，从而产生融合特征和分类头。

设 CLIP 预训练的共享嵌入空间为 $S \subseteq \mathbb{R}^{512}$，图像与文本特征通过投影后在该空间内对齐：

图像分支：$\boldsymbol{X} \to S$

文本分支：$\boldsymbol{T} \to S$

其中：$\boldsymbol{X}$ 为图像输入空间；$\boldsymbol{T}$ 为文本输入空间。

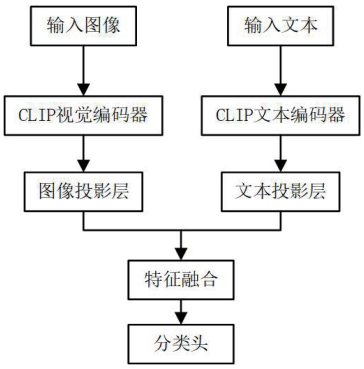

图 2 多模态特征提取流程图

为进一步提升特征的任务相关性，引入一个可学习的投影网络（MLP），该网络由线性层、LayerNorm 层和 GELU 激活函数构成，即图像特征 $\boldsymbol{P}_{\text{img}}$ 为：

$$\boldsymbol{P}_{\text{img}} = \text{GeLU}\left(\text{LayerNorm}\left(\boldsymbol{W}_{\text{proj}}^{\text{img}} \cdot \boldsymbol{E}_{\text{img}}^{\text{clip}}\right)\right) \tag{1}$$

式中：$\boldsymbol{W}_{\text{proj}}^{\text{img}}$ 为投影权重矩阵，本文将投影维度设置为 512，旨在保持特征丰富度的同时完成特征变换；$\text{GeLU}(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}\left(1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$ 为高斯误差线性单元激活函数 [7]。GeLU 激活函数因其平滑且非线性的特性，有助于模型学习更复杂的特征映射。

该投影过程将通用视觉特征 $\boldsymbol{E}_{\text{img}}^{\text{clip}} = \boldsymbol{W}_{\text{vis}} \cdot \left[\text{ViT}\left(\frac{X-\mu}{\sigma}\right)\right]_0$ 转换为任务相关的视觉特征 $\boldsymbol{P}_{\text{img}}$。

本文采用提示工程（Prompt Engineering）策略，将类别名称嵌入模板“a photo of a [class]”中，形成文本输入，以更好地激活 CLIP 的文本先验知识。文本输入 $\boldsymbol{T}$ 首先通过分词器（Tokenize）转换为词元序列，然后输入基于 Transformer 架构的 CLIP 的文本编码器，并提取文本编码器最后一层输出的特征作为整个句子的语义表示。即：

$$\boldsymbol{E}_{\text{txt}}^{\text{clip}} = \boldsymbol{W}_{\text{txt}} \cdot [\text{Transformer}(\text{Tokenize}(\boldsymbol{T}))]_{-1} \tag{2}$$

同样地，为保持模态对称并实现特征空间对齐，文本特征也需通过一个结构相似的投影网络，即文本特征 $\boldsymbol{P}_{\text{txt}}$ 为：

$$\boldsymbol{P}_{\text{txt}} = \text{GeLU}\left(\text{LayerNorm}\left(\boldsymbol{W}_{\text{proj}}^{\text{txt}} \cdot \boldsymbol{E}_{\text{txt}}^{\text{clip}}\right)\right) \tag{3}$$

式中：$\boldsymbol{W}_{\text{proj}}^{\text{txt}}$ 是文本投影层的权重矩阵。

为融合两种模态的信息，本文采用特征加法进行融合。即将投影后的图像特征与文本特征直接相加，这样不增加特征维度，有利于在元学习快速适配过程中保持计算高效性。故多模态融合特征 $\boldsymbol{F}$ 为：

$$\boldsymbol{F} = \boldsymbol{P}_{\text{img}} + \boldsymbol{P}_{\text{txt}} \tag{4}$$

（1）元学习训练：基于 MAML 算法通过两次梯度更新（内循环和外循环）训练一个能够快速适应新任务的模型。通常设置为 *N*-way *K*-shot[8]，即每个任务有 *N* 个类别，每个类别有 *K* 个样本（支持集）和 *Q* 个查询样本。元训练过程通过大量元任务来模拟测试时的小样本场景。每个元任务的构

建方式如下：从训练集类别中随机采样 N 个类别，从每个类别中随机采样 K 个样本构成支持集，再采样 Q 个查询样本用于评估和元更新。本文采用 episodic training 范式，使模型学会如何利用支持集中的 K 个样本（即“少量样本”）来快速适应新任务，并在查询集上做出准确预测。这种训练方式确保了学习到的初始参数对一系列新任务具有良好的泛化性。

其训练目标为：

$$\min_{\theta} \boldsymbol{E}_{T\sim p(T)}\left[L_T\left(U_T^k(\theta)\right)\right] \tag{5}$$

式中：$U_T^k(\theta)$是任务 T 上对模型初始参数 θ 的 k 步更新算子；L_T 是任务 T 的损失函数。

具体表现为通过内部循环（任务内优化）和外部循环（全局优化）优化多模态初始参数。

①内循环：在每个任务 T_i 上，使用支持集D_i^{sup}对模型进行几次梯度更新（称为适应步骤），得到一个针对该任务优化后的模型参数θ'_i，即任务特定模型。

$$\theta'_i=\theta-\alpha\nabla_{\theta}L_{T_i}(f_{\theta}, D_i^{\text{sup}}) \tag{6}$$

式中：α 为内循环学习率。

②外循环：使用查询集D_i^{query}计算任务特定模型θ'_i的损失，然后通过梯度下降更新原始模型的参数 θ，使得经过内循环更新后的模型在查询集上表现良好。

$$\theta\leftarrow\theta-\beta\nabla_{\theta}\sum_{T_i\sim p(T)}L_{T_i}(f_{\theta'_i}, D_i^{\text{query}}) \tag{7}$$

式中：β 为元学习率。

（2）分类与评估：在测试阶段，对于给定的新任务，模型利用学习到的初始参数，在支持集样本上进行少量几步梯度更新，快速得到一个针对该任务的分类器。最终，使用模型对查询集样本进行预测，输出分类结果，并采用准确率（ACC）、宏 F_1 值进行评估。

2 实验结果与分析

2.1 数据集及实验设置

训练集、验证集、测试集按照 16:4:5 的比例进行划分，即训练集共 64 种类别、验证集共 16 种类别，测试集共 20 种类别。三者之间没有种类重叠。

采用 CLIP-ViT 作为主干网络，正则化 Dropout 为 0.3。训练过程中，迭代次数为 50。元优化器采用 Adam，学习率为 0.000 3；内部优化器使用 SGD，学习率为 0.002。学习率调度采用余弦重启 [9]。元批量大小为 16。训练模式采用 5-way-1-shot 和 5-way-5-shot。

采用 CLIP-ViT 作为主干网络，正则化 Dropout 为 0.3，减少过拟合风险。训练过程中，迭代次数为 50。元优化器采用 Adam，学习率为 0.000 3；内部优化器使用 SGD，学习率为 0.002。学习率调度采用余弦重启 [9]。元批量大小为 16。训练模式采用 5-way-1-shot 和 5-way-5-shot。

本文使用交叉熵损失（cross-entropy loss）和对比损失（contrastive loss）联合优化目标。交叉熵损失函数表达式如式（8）所示，通过最大化交叉熵，直接优化分类准确性，确保模型对各类别的预测概率分布逼近真实标签分布。

$$L_{\text{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{=1}^{C}y_{i,c}\log(p_{i,c}) \tag{8}$$

式中：N 为样本数量；C 为类别数；$y_{i,c}$ 为样本 i 的真实类别 c 的 One-Hot 编码；$p_{i,c}$ 为模型样本 i 属于类别 c 的预测概率。

对比损失函数表达式为：

$$L_{\text{Contrastive}}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N}\exp(s_{i,j}/\tau)} \tag{9}$$

式中：$s_{i,j}=\cos(\boldsymbol{v}_{\text{img}}^i,\boldsymbol{v}_{\text{text}}^j)$表示图像特征$\boldsymbol{v}_{\text{img}}^i$与文本特征$\boldsymbol{v}_{\text{text}}^j$的余弦相似度；$\tau$ 表示温度参数，设为 0.07。

通过计算特征空间中的余弦相似性，学习样本间的相似性，拉近同类样本特征，推远异类样本特征，从而提升特征的判别性，增强模型对输入数据的表征能力，尤其适用于小样本或数据稀疏场景。

总损失函数为：

$$L_{\text{Total}}=-L_{\text{CE}}+\lambda L_{\text{Contrastive}} \tag{10}$$

式中：λ=1，即两项损失权重相等，可以平衡分类任务与特征学习。

2.2 实验结果

将本文提出的基于多模态元学习的方法（MAML-CLIP）与 MAML-ResNet18 模型、CLIP-Zero 模型在 Mini-ImageNet 数据集上进行对比，其中，MAML-ResNet18 模型使用标准的 ResNet18 作为骨干网络，采用元学习算法；CLIP-Zero 模型直接使用预训练的 CLIP。

采用准确率、宏 F_1 值和 AUC（One-vs-One）作为评价指标来评估模型的分类性能，结果如表 1 所示。

表 1 不同模型在 Mini-ImageNet 数据集上的性能对比

单位：%

模型	5-way-1-shot			5-way-5-shot		
	ACC	F_1	AUC-OVO	ACC	F_1	AUC-OVO
MAML-ResNet18	48.7	45.9	72.3	65.4	63.2	83.6
CLIP-Zero	58.2	56.3	78.5	62.1	60.8	81.7
MAML-CLIP	**70.9**	**70.8**	**92.4**	**77.8**	**77.7**	**94.9**

从表 1 可以看出，本文提出的 MAML-CLIP 方法在 5-way-1-shot 和 5-way-5-shot 的任务中，准确率均高于 MAML-ResNet18 模型和 CLIP-Zero 模型，表明多模态特征的引入、对 CLIP 模型进行部分微调策略和元学习的结合能够有效提升小样本图像分类性能。具体表现为：

在模型性能方面，MAML-CLIP 模型在 1-shot 和 5-shot 场景中都呈现出非常出色的性能表现。其中，1-shot 时准确率达到 70.9%，5-shot 时准确率为 77.8%，显著优于其他两

种模型。表明将 CLIP 强大的预训练表征能力同 MAML 的元学习能力结合起来，能够有效提高小样本学习的成效。CLIP-Zero 在 1-shot 任务中准确率是 58.2%，超过了 MAML-ResNet18 模型的 48.7%，证明了大规模预训练模型即使不做微调，其零样本能力也能胜过传统元学习方法，体现出预训练模型在小样本学习中的重要意义。当样本数量从 1-shot 增长到 5-shot 时，MAML-ResNet18 的性能提高最为明显，准确率从 72.3% 提高至 83.6%，增长了 11.3%，CLIP-Zero 的提高有限，仅有 3.9% 的增加。表明传统元学习方法对样本数量的依赖程度更高，预训练模型在样本数量极少的情况下就能展现出较好的性能。

此外，为了直观展示本文方法在特征提取方面的优势，采用 t-SNE 算法 [10] 对模型生成的特征进行可视化，分别从 MinilmageNet 数据集中随机选取 5 类图像和 20 类图像，可视化结果如图 3~4 所示。可视化结果表明，右图 MAML-CLIP 在低维特征空间中实现了较好的类内聚合和类间分离，而左图 MAML-ResNet 表现出较高的类间混淆。这一差异凸显了跨模态预训练对元学习特征泛化的促进作用，说明引入多模态预训练模型 CLIP 作为元学习的特征提取器，能大幅提升模型对新颖类别的快速适应能力。

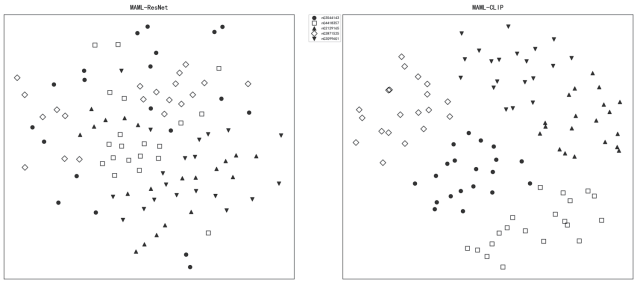

图 3 5 类图像 t-SNE 可视化结果

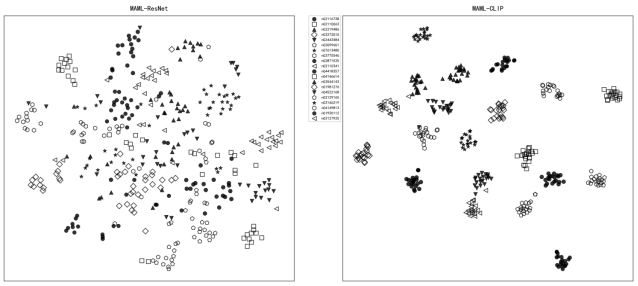

图 4 20 类图像 t-SNE 可视化结果

3 结论

本文提出了一种融合文本特征与图像特征的多模态元学习的小样本图像分类方法，结果表明模型在数据稀缺场景下的性能有较大提升。但是仍存在一些局限性，现有模型的性能提升受限于文本编码器对类别名称和描述的语义理解能力，如果类别名称歧义较大或过于生僻，文本特征无法精准捕捉到原始语义，影响特征准确性。计算开销较大的 CLIP 模型与元学习的双重优化过程相结合后，对计算资源要求较高。针对以上问题，未来工作可以致力于探索更高效更轻量的多模态特征融合网络架构，降低计算成本；探索如何引入更为丰富或者具备结构化特征的语义信息，以此来强化文本表征能力；将本文框架扩展至更多模态（如音频、视频）的小样本学习任务中，进一步提升模型的通用性。

参考文献：

[1]Li Feifei, Fergus R, Perona P. A bayesian approach to unsupervised one-shot learning of object categories[C/OL]// Proceedings Ninth IEEE International Conference on Computer VisionPiscataway:IEEE,2003[2026-04-27]. https://ieeexplore.ieee.org/document/1238476.DOI:10.1109/ICCV.2003.1238476.

[2] 李凡长 , 刘洋 , 吴鹏翔 , 等 . 元学习研究综述 [J]. 计算机学报 , 2021, 44(2):422-446.

[3]Radford A, Kim J W, Hallacy C, et al. CLIP: Connecting Text and Images[PP/OL].(2021-03-01)[2026-04-27].https://arxiv.org/abs/2103.00020.

[4]Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017,60(6):84-90.

[5]Radford A, Kim J W, Hallacy C, et al.Learning transferable visual models from natural language supervision[PP/OL]. (2021-02-26)[2026-04-27].https://doi.org/10.48550/arXiv.2103.00020.

[6]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.NewYork: ACM, 2017: 6000 - 6010.

[7]Hendrycks D, Gimpel K. Gaussian error linear units (GELUs) [PP/OL].(2023-06-06)[2026-04-27].https://doi.org/10.48550/arXiv.1606.08415.

[8] 史燕燕 , 史殿习 , 乔子腾 , 等 . 小样本目标检测研究综述 [J]. 计算机学报 ,2023,46(8):1753-1780.

[9]Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts[PP/OL]. (2017-05-03)[2026-04-27].https://doi.org/10.48550/arXiv.1608.03983.

[10]Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9:2579-2605.

【作者简介】

王雁（1994—），女，山西孝义人，硕士，助教，研究方向：计算机视觉与信号处理。

（收稿日期：2025-09-19 修回日期：2026-05-07）