西北大学学报(自然科学版)
2025 年 2 月,第 55 卷第 1 期,Feb. ,2025,Vol. 55,No. 1
Journal of Northwest University (Natural Science Edition)

·文化遗产数字化保护·

面向非遗美术图像分类的提示学习方法

张秦瑜[1,2],刘鑫达[1,2],鲁倬铭[3],周明全[1,2,4]

(1. 西北大学 文化遗产数字化国家地方联合工程研究中心,陕西 西安 710127;2. 西北大学 信息科学与技术学院,陕西 西安 710127;3. 加州戴维斯大学 文理学院,美国 戴维斯 95616;
4. 北京师范大学 教育部虚拟现实应用工程研究中心,北京 100875)

摘要 针对中国非物质文化遗产美术作品分类中处理效率低、数据复杂等问题,提出了一种基于预训练视觉语言大模型的上下文提示微调策略,以提升小样本情况下的分类性能并应对当前任务的挑战。该方法通过引入可学习的上下文优化提示(软提示),使模型能够在少量样本条件下快速适应下游分类任务,从而有效缩短训练时间并提升收敛速度。具体而言,利用注意力机制,将由软提示生成的文本特征与预训练视觉语言模型的原始特征相结合,并通过对比损失优化嵌入表示。这一机制减少了不同特征之间的嵌入差异,避免了模型对已知类别的过度拟合,提升了在未见类别上的泛化能力。此外,保留原始特征信息帮助模型避免训练过程中遗忘基础知识,确保即便在小样本条件下,模型仍能保持较高的分类准确率。实验结果表明,所提出方法在非遗美术图像分类任务中的准确率提升了 1.79%,泛化识别能力提升了 10.4%,同时具备较低的计算成本。

关键词 非物质文化遗产;图像分类;上下文优化;注意力机制

中图分类号:TP391 **DOI**:10.16152/j.cnki.xdxbzr.2025-01-009

Research on a prompt learning method for intangible cultural heritage art image classification

ZHANG Qinyu[1,2], LIU Xinda[1,2], LU Zhuoming[3], ZHOU Mingquan[1,2,4]

(1. National and Local Joint Engineering Research Center for Cultural Heritage Digitization, Northwest University, Xi'an 710127, China;
2. School of Information Science and Technology, Northwest University, Xi'an 710127, China;
3. College of Letters and Science, University of California, Davis CA 95616, USA;
4. Virtual Reality Research Center of Ministry of Education, Being Normal University, Bejing 100875, China)

Abstract To address the issues of prolonged processing time, low efficiency, and high data complexity in the classification of Chinese intangible cultural heritage (ICH) artworks, this paper proposes a context-based text prompt tuning strategy based on a pre-trained vision-language model. This approach introduces trainable context optimization soft prompts, enabling the model to quickly adapt to downstream classification tasks under

收稿日期:2024-11-20

基金项目:虚拟现实技术与系统全国重点实验室(北京航空航天大学)开放课题基金(VRLAB2024C02);文化和旅游部重点实验室项目(1222000812、cr2021K01);西安市科技计划社会发展科技创新示范项目(2024JH-CXSF-0014);国家自然科学基金(62271393)。

第一作者:张秦瑜,男,从事多模态大模型提示学习研究,zhangqinyu@stumail.nwu.edu.cn。

通信作者:刘鑫达,男,博士,讲师,从事虚拟现实、人工智能与数字文化遗产保护等研究,liuxinda@nwu.edu.cn。

limited sample conditions, thereby effectively reducing training time and improving convergence speed. Specifically, the proposed method integrates text features generated by the soft prompts with the original features of the pre-trained vision-language model through an attention mechanism, and optimizes the embedded representations via a contrastive loss function. This mechanism significantly reduces the embedding discrepancy between the two types of features, preventing the model from overfitting to visible base categories and enhancing its generalization ability to unseen classes. Moreover, the retention of original features helps mitigate catastrophic forgetting during training, ensuring high classification accuracy even under few-shot conditions. Experimental results demonstrate that the proposed method improves classification accuracy by 1.79%, enhances generalization by 10.4%, and maintains low computational cost.

Keywords　intangible cultural heritage; image classification; contextual optimization; attention mechanism

当前,非物质文化遗产(以下简称“非遗”)的保护与传承愈发受到重视。非遗不仅承载了深厚的历史和文化价值,同时也是民族身份认同和文化多样性的关键体现。为有效保护和传承这些宝贵的文化遗产,构建科学合理的分类体系至关重要[1]。王燕妮的研究指出,民俗类非遗在我国国家级非遗名录中占有重要地位,科学分类对非遗的申报、管理、传承和活化具有显著意义[2]。然而,现有的非遗分类体系仍存在不足,如缺乏对文化空间的分类,体系过于复杂,且依赖人工整理的效率较低[2]。

随着深度学习技术的发展,图像分类领域涌现出一系列高度自动化且高效的特征提取算法,为民俗非遗美术作品的分类提供了新的思路[3]。其中,卷积神经网络(convolutional neural networks, CNN)作为图像分类的核心技术之一,通过逐层卷积操作能够提取图像的多层次特征,在复杂视觉任务中表现卓越。诸如 LeNet[4]、AlexNet[5]、VGG[6]和 ResNet[7]等模型通过增加网络深度和引入创新结构模块(如 Inception 模块和残差连接)不断优化性能,EfficientNet[8]则通过复合缩放策略在保持高精度的同时降低了计算成本,使其适用于计算资源受限的环境。然而,这些传统方法在非遗美术图像分类任务中依然面临诸多挑战。首先,CNN 的有效性通常依赖于大规模标注数据集,但非遗美术图像数据稀缺且种类繁多,构建大规模数据集的成本较高。例如,剪纸和皮影虽然在视觉上有显著差异,但现有模型难以通过少量数据训练完成二者的准确区分。此外,数据集的有限性容易导致模型过拟合,从而限制其泛化能力,难以应对新类别增长的需求。同时,传统分类方法高度依赖人工标注,增加了成本与难度[9-10]。近年来,诸如 SimCLR[11]等对比学习方法通过无监督预训练利用无标签数据进行特征学习,并通过微调完成下游分类任务,减少了对大规模标注数据的依赖。尽管如此,这些方法在处理非遗美术图像等特定数据时,无法充分利用预训练过程中学到的图像通用特征,仍旧需要大量任务相关的图像数据进行微调。因此,它们在小样本分类任务中的适用性较为有限,难以满足非遗图像分类中图像数据获取困难的实际需求。

在此背景下,视觉-语言预训练模型为非遗美术图像分类带来了新的突破。通过在大规模图像-文本配对数据集上进行预训练,这些模型能够学习到丰富的视觉和语言特征表示,从而在零样本或少样本条件下实现高效的下游任务识别[12-13]。如 OpenAI 开发的 CLIP 模型利用4 亿图像-文本对进行训练,已在多个图像分类任务中展示了出色的表现,并实现了零样本迁移[14]。相比于传统对比学习方法,这类视觉-语言模型在处理小样本分类任务时表现出了更强的泛化能力和适应性。为进一步提升视觉-语言大模型在下游任务中的表现,研究者们借鉴自然语言处理领域提示微调方法的思路,通过引入可学习的文本提示信息优化模型,以更好地适应封闭数据集上的特定任务[15]。上下文优化提示方法(CoOp)通过将提示中的上下文词转换为可学习文本向量,进而对预训练模型进行高效微调[16],与视觉特征对齐。该方法在少样本分类任务中可以得到较高的准确率,但易导致模型过度拟合训练过的类别,进而影响在未见新类别上的泛化能力[17]。为此,周等提出了条件上下文优化提示方法(CoCoOp),该方法通过在可学习提示中引入图像特征信息,增强了模型对未知类别的泛化能力[17]。尽管如此,CoCoOp 中提示词的共享性干扰了模型对不同类别特征的区分能力,进而在已知类别的识别精度

上存在一定损失[18]。

针对上述挑战,本文提出了一种基于注意力机制的视觉-语言提示微调方法,并将其应用于陕西非遗美术图像分类任务。陕西的非遗美术作品类型丰富多样,包括剪纸、皮影、泥塑、刺绣等。这些作品不仅在视觉特征上差异显著,还蕴含着丰富的文化背景和语义信息,传统的图像分类方法难以有效区分。本文基于预训练的视觉-语言大模型(如 CLIP),利用其已学习的通用文本特征,在小样本环境下实现高效识别。CLIP 模型具备强大的图像特征提取能力,无需依赖大规模标注数据进行训练,只需少量数据(每类 1 至 16 张图片)对图像特征与文本特征进行微调对齐,即可适应下游任务,克服了传统数据驱动算法对大规模标注图像依赖的局限性。此外,预训练模型的泛化能力更强,有效防止数据偏差或数据质量问题导致的过拟合风险。相比传统方法,本方法在小样本训练中显著降低了对计算资源和时间的需求,为非遗美术图像分类任务提供了天然的优势,尤其在数据有限的情况下,依然能够获得较高的识别精度与泛化性能。

具体而言,本文方法通过结合可学习的上下文提示与人工设计的文本提示,并引入注意力机制,显著增强了模型对不同类别和复杂场景的适应性。尽管剪纸和壁画等类别在视觉特征上存在较大差异,但通过视觉-语言对齐和提示微调策略,本方法能够更好地捕捉其潜在的语义关联。与现有方法相比,本文方法不仅在准确率上实现了显著提升,还增强了模型对未知类别的泛化能力,降低了泛化性能与基础准确率之间的权衡。实验结果表明,本文提出的方法在陕西非遗美术图像分类任务中的准确率提升了 1.79%,泛化能力提升了 10.4%。

1 相关方法

1.1 视觉语言预训练模型

近期研究显示,向预训练模型引入大规模的图像-文本数据对是构建强大视觉-语言模型的关键,这些模型能够在下游任务中实现零样本和少样本学习。CLIP 模型[14]便是其中的代表,它采用了对比损失来训练,基于从互联网上收集的 4 亿个图像-文本对,通过这种方式,CLIP 成功地训练了其视觉编码器和文本编码器。同期的研究,例如 ALIGN[19],也采用了相似的策略,它利用了 18 亿个带噪声的图像-文本对,并应用对比损失来预训练模型。这些视觉-语言模型均采用了双编码器架构,包括一个图像编码器和一个文本编码器,这种架构使得模型能够通过文本和图像特征的对齐,在多种视觉分类任务中展现出卓越的基于提示的零样本性能。

本研究提出的模型亦采用 CLIP 进行知识迁移。与当前的最新研究如 CoOp[16] 和 CoCoOp[17] 一样,这些方法都依赖于 CLIP 模型强大的视觉-语言联合表示能力,以实现对新任务的快速适应和学习。具体来说,本研究在 CLIP 的基础上,通过引入注意力机制优化的视觉-语言提示微调方法,进一步提升模型的泛化能力和分类性能。

1.2 提示微调

为提升预训练视觉-语言模型在特定下游任务中的适应性,研究者们常采用与任务紧密相关的文本标记进行提示微调,以此推断任务特定的文本知识。例如在 CLIP 模型中,手工设计的模板“a photo of a [CLASS]”被广泛用于构建零样本预测的文本嵌入。然而,这种手工提示存在局限性,因为它们未能充分考虑特定任务的知识细节,导致其在描述任务时的能力受限[20]。

针对这一挑战,CoOp 方法应运而生。CoOp 通过少量样本学习,用可学习的软提示替代了原先的手工提示,从而提高了模型对任务文本的推断能力。然而,CoOp 方法存在其固有缺陷:它生成的可学习提示是针对每个任务独特且固定的,这意味着 CoOp 在推断任务相关提示时,未能充分考虑不同图像之间的特征差异[17]。为了克服 CoOp 的这一局限,CoCoOp 方法被提出。CoCoOp 通过为每个图像生成特定的图像条件上下文,并将其与文本条件上下文相结合,进行提示调整。具体来说,CoCoOp 利用一个轻量级的神经网络,为每个图像学习生成一个向量,这个向量作为可学习的文本提示的一部分。这种方法显著提升了模型在未知类别上的泛化能力,但与此同时,也带来了对基类识别精度的一定损失。

总之,CoCoOp 方法通过动态调整提示,增强了模型对不同图像特征的敏感度,从而在一定程度上解决了 CoOp 方法在泛化性方面的不足。尽管如此,CoCoOp 在提升泛化能力的同时,同样牺牲了对基类的识别精度,无法实现模型性能的全面优化[21]。

1.3　注意力机制融合特征

在视觉-语言模型的研究领域,注意力机制已经成为一种至关重要的技术手段。它赋予了模型在处理多模态数据时能够动态地聚焦于相关信息的能力,从而显著提升了模型对数据的理解和表征能力。特别是在视觉-语言联合表示学习任务中,注意力机制通过学习输入特征之间的相关性权重,实现对关键信息的有效筛选和整合[22]。

注意力机制的工作原理通常涉及以下几个步骤:首先,通过线性变换,将输入的文本特征转换为查询(Query)、键(Key)和值(Value)的表示形式[23]。这一步骤通常通过全连接层实现,目的是在空间中重新表达输入特征。其次,计算 Query 和 Key 之间的点积,得到未归一化的注意力分数,这一分数反映了查询与键之间的相关性。接着,应用 SoftMax 函数[24]对这些分数进行归一化处理,确保得到的注意力权重之和为 1,从而为后续的加权求和做好准备。然后,利用这些归一化的权重与值(Value)进行加权求和,生成与上下文相关联的特征表示,这一步骤实现了输入特征与相关信息的融合,产生了更为丰富的特征表示。最后,通过特征融合策略,将上下文特征与原始文本特征结合,并通过线性层进一步变换,产生最终的特征表示。

注意力机制的这些特性,不仅增强了模型对多模态信息的整合能力,而且为视觉-语言模型在图像描述生成、视觉问答、零样本学习等任务中的应用开辟了新的道路。实际应用结果表明,基于注意力机制的模型在多个视觉-语言任务中均取得了显著的性能提升[25]。例如:在图像描述生成中,模型能够更准确地捕捉图像内容,生成更丰富、更准确的描述;在视觉问答任务中,模型能够更有效地识别与问题相关的图像区域,从而提高回答的准确性;在零样本学习任务中,通过有效的特征融合,模型对未见类别的识别能力得到了显著提升。这些成果证明了注意力机制在视觉-语言模型中的有效性[25]。

2　非遗美术图像分类方法

本文提出了一种改进的上下文优化提示算法(见图 1)。为全面解释该模型的工作机制,首先将详细描述 CoOp 方法中上下文优化提示的具体步骤和原理。接着,将深入定义注意力机制,解析其在模型中的应用与作用。最后,阐述本文所提出的优化策略及其实现细节,展示其如何有效提升模型的性能与泛化能力。

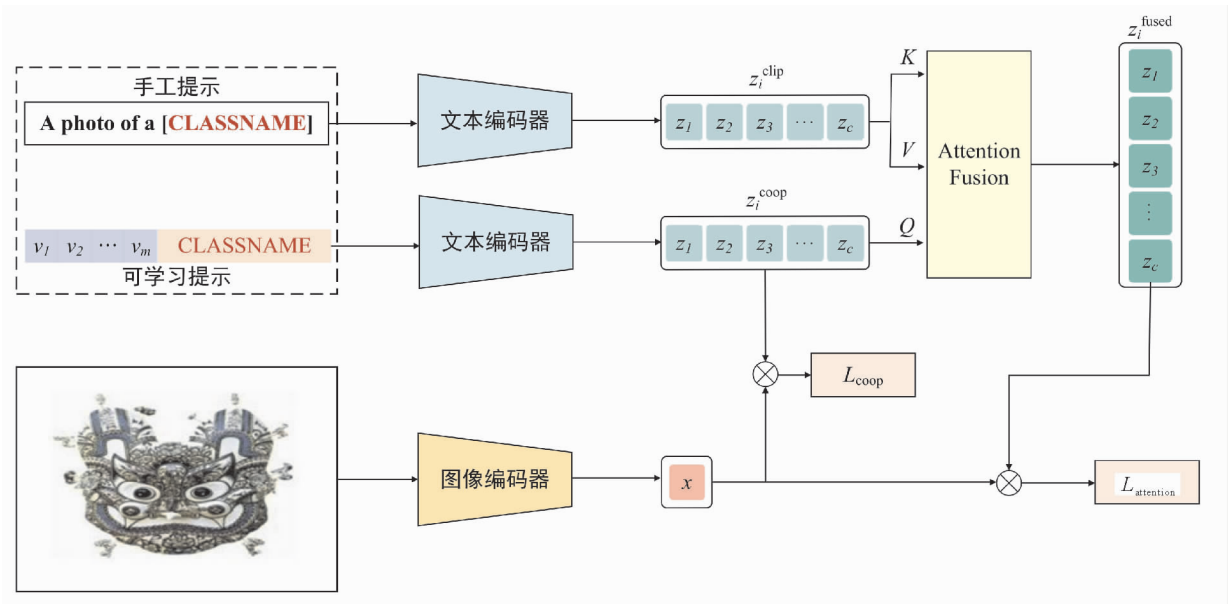

图 1　本文算法框架图

Fig. 1　The algorithm framework

2.1　上下文优化提示

在现有的视觉语言模型中,CLIP 是一个具有强大功能的预训练模型,它通过对 4 亿对图片文本对进行训练,具有强大的零样本图片识别能力。由于 CLIP 是基于文本图像对进行训练的,因此它包含了两种类型的编码器,其一是视觉编码器,另

一种是文本编码器。其中,视觉编码器通过将给定的图像映射到视觉嵌入空间,与通过文本编码器嵌入的文本进行相似度匹配以获得二者的相似度。提示微调是一种通过冻结视觉和文本编码器参数来适应下游任务的技术方法。在这一过程中,模型的核心结构保持不变,不进行大规模的参数更新,而是依赖于精心设计的提示词或可学习的提示词来引导模型的表现。具体而言,手工设计的提示词是由人工根据任务需求构造的短语或句子,旨在引导 CLIP 模型理解和适应特定的任务场景。而可学习的提示词则允许模型自动学习最适合当前任务的上下文,通过优化提示词的表示,使 CLIP 能够在不改变编码器的情况下高效适应新的下游任务。这种方法有效地减少了对大规模数据和训练资源的需求,同时保留了 CLIP 强大的通用性与迁移能力,在少样本学习方面表现出显著优势。

形式化来说,视觉编码器与文本编码器分别定义为 α 和 β,下游任务包含 N_c 个分类类别,CLIP 使用手工提示进行文本嵌入,即 $z^{\text{clip}} = \{z_i^{\text{clip}}\}_{i=1}^{N_c}$ 表示所有的文本嵌入,z_i^{clip} 就表示第 i 个文本嵌入。具体来说,假设第 i 个类别为[CLASSNAME],那么其对应的文本嵌入 z_i^{clip} 是由手工设置的提示词“a photo of a [CLASSNAME].”通过 Transformer 编码器 $\gamma(\cdot)$ 与文本编码器 $\beta(\cdot)$ 得到,其中 $\gamma(\cdot)$ 可以将提示词作为输入并输出向量化的文本标记。所以,第 i 个提示词定义如式(1)所示。

$$t_i^{\text{clip}} = \gamma(\text{a photo of a [CLASSNAME].}) \tag{1}$$

然后将 t_i^{clip} 进一步通过文本编码器 β 投影到文本特征嵌入 z_i^{clip} 上,$z_i^{\text{clip}} = \beta(t_i^{\text{clip}})$。

给定图像 I 及其标签 y,通过视觉编码器 $\alpha(\cdot)$ 提取视觉特征嵌入 $x = \alpha(I)$。之后,计算视觉嵌入 x 和文本嵌入 w_i^{clip} 之间的预测概率进行预测结果如式(2)所示。

$$P_{\text{clip}}(y \mid x) = \frac{\exp\left(\frac{d(x, z_y^{\text{clip}})}{\partial}\right)}{\sum_{i=1}^{N} \exp\left(\frac{d(x, z_i^{\text{clip}})}{\partial}\right)} \tag{2}$$

其中,$d(\cdot)$ 表示余弦相似度[26],∂ 是一个可学的温度参数。虽然 CLIP 可以很容易被应用与零样本下游任务,但是由于它固定的手工提示,使得 CLIP 在很多下游任务中表现不佳。为了解决这个问题,CoOp 通过自动学习一组连续的文本向量来生成下游任务需要的文本嵌入[27]。具体来说 CoOp 首先定义了 M 个文本向量 $\boldsymbol{V} = \{v_1, v_2, \cdots, v_M\}$ 作为可学习的提示。然后将类别词嵌入 c_i 与可学习提示 $\boldsymbol{V}$ 连接得到完整的向量化文本标记 t_i^{coop} 形式化描述为

$$t_i^{\text{coop}} = \{v_1, v_2, \cdots, v_M, c_i\} \tag{3}$$

其中,i 表示第 i 个类别。所以最终得到的文本特征嵌入为 $z_i^{\text{coop}} = \beta(t_i^{\text{coop}})$。通过引入下游任务少量样本,CoOp 可以最小化图片特征 x 和类别文本嵌入 z_y^{coop} 之间的负对数似然[28]来优化可学习的向量 $\boldsymbol{V}$ 过程如公式(4)所示:

$$P_{\text{coop}}(y \mid x) = \frac{\exp\left(\frac{d(x, z_y^{\text{coop}})}{\partial}\right)}{\sum_{i=1}^{n} \exp\left(\frac{d(x, z_i^{\text{coop}})}{\partial}\right)} \tag{4}$$

需要注意的是,在训练过程中视觉编码器与文本编码器都处于冻结状态,CoOp 只推断出合适的与任务相关的提示 t_i^{coop} 来增强其泛化能力和区分能力。

2.2 注意力机制

在现有的视觉语言模型研究领域中,注意力机制已成为提升模型性能的关键技术。注意力机制能够动态地聚焦于多模态数据中的相关信息,从而显著提高模型对数据的理解和表征能力[29]。特别是在视觉语联合表示学习任务中,注意力机制通过学习输入特征之间的相关性权重,实现了对关键信息的有效筛选和整合[30]。

本文提出一种基于注意力的文本特征融合方法。通过设计 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 值的线性变换,构建了注意力网络,并使用 Softmax 函数生成相应的标准化注意力权重。

该模块首先对输入的文本特征进行线性变换,生成查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)和值($\boldsymbol{V}$)的表示形式。形式化地,$\boldsymbol{Q} = \boldsymbol{W}_Q \cdot z_y; \boldsymbol{K} = \boldsymbol{W}_K \cdot z_y; \boldsymbol{V} = \boldsymbol{W}_V \cdot z_y$ 其中 $\boldsymbol{W}$ 表示权重参数,z_y 表示文本特征嵌入。接着计算查询 $\boldsymbol{Q}$ 和键 $\boldsymbol{K}$ 之间的点积,进行缩放,得到注意力分数为

$$\text{Attention} = \text{Softmax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{5}$$

利用归一化的注意力权重与值 $\boldsymbol{V}$ 进行加权求和,生成上下文相关的特征 $\boldsymbol{Z}_f$ 表示为

$$\boldsymbol{Z}_f = \text{Attention} \cdot \boldsymbol{V} \tag{6}$$

最后,将上下文特征与原始文本特征相结合,并通过线性层进行进一步的特征变换,以产生融合后

的特征 z_{fused} 表示为

$$z_{\text{fused}} = \boldsymbol{W}_f \cdot [z_f, z_y] \tag{7}$$

其中,$\boldsymbol{W}_f$ 代表一个用于特征变换的线性变换矩阵。

2.3　基于注意力机制的上下文优化提示

尽管现存的基于 CoOp 的提示微调算法能够有效使预训练的 CLIP 模型适应下游任务,但由于训练过程中只使用了少部分标注图片,该算法容易对看见的类别产生过拟合。现有的改进方法 CoCoOp 通过在文本嵌入中引入图像特征信息来增强对未知类别的识别能力,但这种方式也会干扰模型对不同类别的特征区分能力,从而降低基类识别能力。本文通过深入分析 CLIP 和 CoOp 在未知类别上的准确性,发现可学习的文本嵌入会在一定程度上对基础类别过拟合,从而远离未知类别。

为了解决这一问题,本文引入注意力机制来融合 CLIP 和 CoOp 的文本嵌入,以增强可学习提示和固定提示之间的相似性,减少对通用文本知识的遗忘,从而提高对未见领域的通用性。与 CoCoOp 在词嵌入阶段引入图像信息不同,本方法通过在文本编码器后进行处理,不会破坏模型对细粒度信息的捕获,因此在基类上的表现不会因为额外信息的引入而下降。总结来说,本文提出了一种基于注意力机制的上下文优化新型提示微调方法,以推断出在已见类别上具有高区分度、在未见类别上具有高通用性的可学习提示。具体流程如图 1 所示。

首先,手工提示词通过冻结的文本解码器获得第 1 组文本嵌入,该过程与 CLIP 一致,然后让可学习的提示词通过冻结的文本解码器获得第 2 组文本嵌入;二者分别作为注意力机制模块的查询(Q)、键(K)和值(V)输入,最终得到融合的文本嵌入。输入图片 I 经过冻结的图像解码器获得图像特征嵌入,然后将可学习提示得到的文本嵌入与图像嵌入相结合,计算余弦相似度 L_{coop},以获取可见类别的信息,从而获得较好的准确率。接着,通过融合的文本嵌入与图像嵌入结合,计算二者的余弦相似度 $L_{\text{attention}}$,以防止模型对基类过拟合。

形式上,手工提示词的文本嵌入和可学习的文本嵌入分别定义为 $z_i^{\text{clip}} = \beta(t_i^{\text{clip}})$ 和 $z_i^{\text{coop}} = \beta(t_i^{\text{coop}})$,其中 t_i^{clip} 表示手工提示词向量化文本标记,t_i^{coop} 表示可学习的向量化文本标记,i 表示第 i 个类别。则经过注意力模块的融合文本嵌入 z_{fused} 表示为

$$z_{\text{fused}} = \text{Attention}(z^{\text{coop}}, z^{\text{clip}}, z^{\text{clip}}) \tag{8}$$

用于拟合基类的标准对比损失 L_{coop} 表示为

$$L_{\text{coop}} = -\sum_{x \in X} \log \frac{\exp\left(\frac{d(x, z_y^{\text{coop}})}{\partial}\right)}{\sum_{i=1}^{n} \exp\left(\frac{d(x, z_i^{\text{coop}})}{\partial}\right)} \tag{9}$$

同样地,为了有效防止模型在基类数据上产生过拟合,模型引入了对比损失,其作用机制如式(10)所示。通过这种损失函数的引导,模型能够更好地捕捉类间差异,避免在基类上过度拟合。

$$L_{\text{attention}} = -\sum_{x \in X} \log \frac{\exp\left(\frac{d(x, z_{\text{fused}})}{\partial}\right)}{\sum_{i=1}^{n} \exp\left(\frac{d(x, z_{\text{fused}})}{\partial}\right)} \tag{10}$$

通过结合标准交叉熵损失 L_{coop},最终的目标损失函数如式(11)所示。

$$L = (1 - \theta) \cdot L_{\text{coop}} + \theta \cdot L_{\text{attention}} \tag{11}$$

其中 θ 用于平衡 L 在最终目标任务中的效果。

3　实验与分析

本文遵循以下两组实验设置来评估算法优劣性:①少样本的图像分类;②在数据集中从基础类别到不可见新类别的泛化。所有实验都是基于预训练的 CLIP 模型进行的。

数据集:本文在 5 个图像细粒度分类数据集上进行了从基础类别到新类别的泛化能力测试。其中 4 个通用公共数据集 Flowers102[31]、Stanford Dogs[32]、Fruit92[33]、Veg200[33]用于细粒度视觉分类,这些数据集广泛应用于细粒度视觉分类研究,涵盖了从植物(如 Flowers102、Fruit92、Veg200)到动物(如 Stanford Dogs)的多个领域,具有显著的类别区分细微性和样本分布不均性特点。这些特性为模型在处理类别间差异较小、视觉特征高度相似的数据时提供了充分验证的基础。此外,这些数据集为研究基础类到新类别的泛化性能提供了多样化的实验场景,有助于探索模型在应对跨域或新类别任务时的表现。特别地,为了衡量该模型在非遗美术图像分类任务上的能力,我们收集并整理了共 12 000 张陕西省非物质文化遗产美术作品图像数据集如图 2 所示,

该分类方式是依据陕西非物质文化遗产保护中心发布的陕西民俗艺术官方分类标准而建立的,其中包含了凤翔木版年画、延川布堆画、农民画(安塞民间绘画,澄城手绘门帘)、麦秸画(黄陵麦秸画、吴起糜粘画工艺)、泥塑(凤翔泥塑、吴起泥塑工艺)、面塑(黄陵面花、华州面花、澄城面花)、绥德石雕、民间木雕(阎良核雕技艺、佳县庙宇木雕雕刻技艺)、西秦刺绣、剪纸(定边剪纸艺术、延川剪纸、黄陵剪纸、洛川剪纸)、皮影(华县皮影、礼泉皮影)、宝鸡陈仓区宝鸡社火脸谱、乾州布玩具、合阳提线木偶戏、凤翔草编共计 15 个类别 25 种品类陕西本土非物质文化遗产美术作品分类数据。与通用细粒度数据集相比,该数据集具有强异质性、高视觉相似性、文化背景复杂性等特点,能够真实反映实际应用场景中复杂分类任务的挑战性。例如,某些类别之间在材质、纹理和色彩上高度相似,仅在细节上有所差异。此外,数据集中包含多个品类和地区工艺的变体,进一步提升了任务的难度。通过在此非遗美术作品分类数据集上的实验,不仅可以验证所提出模型在处理精细差异和视觉特征复杂性时的表现,还能够展示其在高异质性实际场景中的应用潜力。这种设计同时突出了细粒度分类任务与非遗美术图像分类任务在需求上的一致性,为实验设置的合理性提供了有力支撑。

图 2 陕西非遗美术作品数据集(部分)

Fig. 2 Shaanxi intangible cultural heritage artworks dataset (Partial)

训练细节:算法的实现基于 CoOp 和 CLIP 模型,并在具有 Vit-B/16[14] 的视觉骨干网络上进行实验。受 CoOp 启发,上下文文本向量的长度始终固定为4,并使用模板"a photo of a [ ]"初始化上下文向量。最终性能是在三个随机种子上平均得出的,以进行公平比较。此外,实验遵循 CoOp 的训练计划和数据增强设置。超参数 θ 被设置为 0.1。所有实验都是在一张 RTX 4090 显卡上进行。

基线模型:实验使用 3 种基于 CoOp 的方法与本文提出的算法进行比较。

· CLIP 使用手工制作的模板“a photo of a [ ]”来生成用于知识迁移的提示。

· CoOp 用下游数据集推断出的一组可学习提示替换了手工制作的提示。

· CoCoOp 通过结合每张图片的图像上下文和 CoOp 中的可学习提示来生成条件性图像提示。

· AttentionCoOp 该模型即本文使用的方法,通过注意力机制融合可学习文本嵌入与手工文本嵌入来生成文本提示。

3.1 少样本学习实验

和近期的 CoOp 工作的实验相似,实验遵循 CLIP 中采用的少样本评估协议,分别使用 1,2,4,

8 和 16 个样本进行训练,并在完整的测试集上部署模型。报告了 3 次运行的平均结果以供比较。详细的结果如图 3 所示。图 3 给出了 4 个基线模型在 5 个数据集上的表现以及 5 个数据集的平均表现实验结果显示,改进的模型在 5 个数据集上,数据集训练样本数量为 1,2,4,8,16 时均获得了最高的准确率。其原因在于,通过在反向传播阶段前引入注意力机制融合的文本嵌入不但可以保持可学习文本嵌入学习到的分类细节信息,还能额外补充 CLIP 模型捕获的信息,使得该模型在不同的样本数量下均有着优秀的表现。

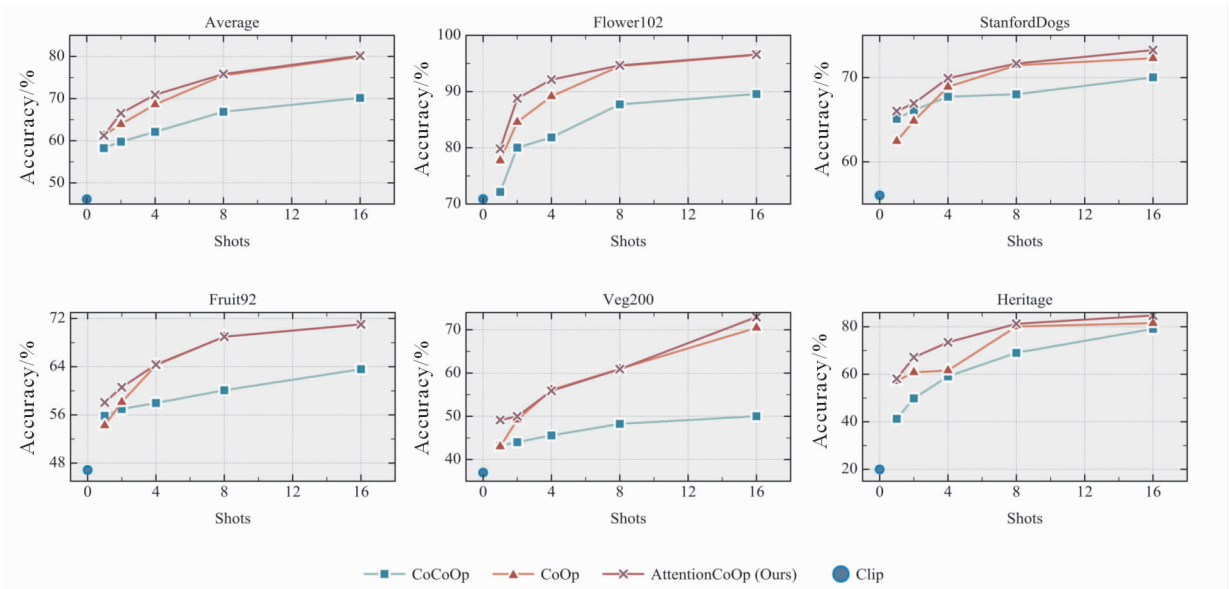

图 3　少样本学习在 5 个数据集上的准确率

Fig. 3　Accuracy of few-shot learning on 5 datasets

3.2　基类到新类的泛化实验

与先前的工作 CoOp 和 CoCoOp 类似,每个数据集类别被平均分成两组:基础类别和新类别。与零样本设置相似,新类别与基础类别不重叠。为了验证本文方法的泛化能力,所有比较的方法和本文提出的方法都使用基础类别进行提示调整,并在新类别上进行评估。详细的结果展示在表 1 中。表 1 给出了基于 ViT-B/16[14] 骨干网络和 16-shot 样本的所有 5 个数据集的详细性能。

表 1　基类到新类泛化实验中各个模型的准确率比较

Tab. 1　Accuracy comparison of different models in the base-to-new class generalization experiment

	Base	New	H
CLIP	60. 87	56. 46	57. 40
CoOp	85. 60	47. 40	61. 01
CoCoOp	80. 23	57. 33	66. 87
Ours	**86. 31**	**57. 59**	**69. 08**

(a) Average over 5 datasets

	Base	New	H
CLIP	72. 08	77. 80	74. 83
CoOp	**97. 63**	69. 55	81. 23
CoCoOp	94. 87	71. 75	81. 71
Ours	97. 49	**78. 60**	**87. 03**

(b) Flowers102

	Base	New	H
CLIP	61. 90	63. 00	62. 45
CoOp	75. 80	58. 03	65. 74
CoCoOp	74. 12	**67. 87**	70. 86
Ours	**76. 33**	66. 52	**71. 09**

(c) Stanford Dogs

	Base	New	H
CLIP	53. 00	61. 10	56. 76
CoOp	85. 39	46. 96	60. 60
CoCoOp	74. 58	**60. 05**	66. 53
Ours	**85. 66**	58. 99	**69. 86**

(d) Fruit92

	Base	New	H
CLIP	44. 90	43. 80	44. 34
CoOp	78. 19	30. 20	43. 57
CoCoOp	67. 19	**45. 79**	**54. 46**
Ours	**79. 28**	41. 20	54. 22

(e) Veg200

	Base	New	H
CLIP	72. 45	36. 60	48. 63
CoOp	91. 01	32. 25	47. 62
CoCoOp	90. 38	41. 17	56. 57
Ours	**92. 80**	**42. 65**	**58. 44**

(f) Heritage

结果如表 1 所示,本文提出的 AttentionCoOp 在所有数据集上的调和平均值 H 均获得了比现有方法更高的平均性能,这展示了该模型在基础类别和新类别识别性能平衡方面的优越性。在现有的 CoOp 与 CoCoOp 算法中,本文方法无论是基类还是新类都有超过 CoOp 的识别表现。这也证明了 AttentionCoOp 在保留基类识别准确率的同时,可以有效提高未见类别的泛化能力。尽管在一部分数据集上,CoCoOp 在新类的识别率略高于本文方法,但可以看出,CoCoOp 在基类上的识别率大幅度损失。其原因在于 CoCoOp 通过结合每张图片的图像上下文和 CoOp 中的可学习提示来生成条件性图像提示。这种方法使得 CoCoOp 在新类别上相对于 CoOp 有所改进,但也干扰了其在基类的表现。

3.3 超参数 θ 实验

在本文的研究中,超参数 θ 的关键作用如公式(11)在于通过提示调整机制,将可学习的文本特征向通用文本特征逼近,从而直接影响最终损失函数在最终任务的表现,这种调整过程受到参数 θ 的严格约束。为深入分析参数对模型性能的影响,本文评估了不同参数值下模型的表现,结果如图 4 所示。

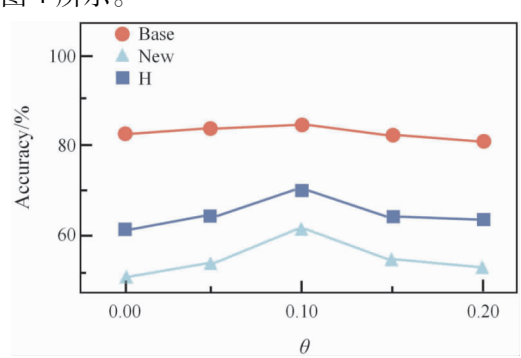

图 4 不同 θ 值对模型性能的影响

Fig. 4 The impact of different θ values on model performance

具体而言,我们考察了模型在 5 个数据集的基础类(Base)上的平均识别准确率、未知类别(New)的平均泛化准确率及其调和平均值(H),并将所有结果取平均值以分析参数对模型性能的整体影响。实验结果表明,当参数 θ 退化至 0 时,模型的泛化性能降至最低。这表明在没有正则化项的情况下,模型退化为原始的 CoOp 方法,难以有效捕捉文本特征的通用性。此外,这种性能下降可能与单一文本离散提示生成的嵌入特征空间表示过于稀疏有关,从而限制了模型的泛化能力。随着正则化项 θ 的引入,模型训练过程受到融合特征的约束,能够有效防止基础知识的遗忘,从而显著提升泛化能力。实验结果显示,模型性能随着 θ 值的增加呈现出先升后降的趋势。当 θ 取值为 0.1 时,模型在基础类、未知类别以及调和平均的 3 项关键性能指标上均达到最优表现。然而,随着 θ 值进一步增大,尽管正则化项强化了特征的通用性,但过强的正则化可能削弱模型对基础类细粒度特征的表达能力,导致性能有所下降。上述结果进一步验证了合理设置超参数在提升模型泛化能力方面的重要性。

3.4 消融实验

为了深入验证所提出的基于注意力机制的特征融合模块在增强模型整体性能方面的实际效果,我们设计并执行了一系列详尽的消融实验。这些实验主要针对 5 个具有代表性的细粒度图像分类数据集展开,旨在考察模型从基类到新类的泛化能力。具体来说,本实验着重比较了模型在引入注意力机制前后的识别准确率的性能变化,这种比较主要体现在基类(Base)、新类(New)以及二者的调和平均值(H)这 3 个关键性能指标上。以这些指标的变化值(即引入注意力机制后的性能指标值与未引入时的性能指标值之差)作为衡量模型性能提升或下降的标准。若变化值为正,则意味着引入注意力机制后,模型的性能得到了显著提升;反之,若变化值为负,则表明引入注意力机制后,模型的性能出现了下降。实验结果详见图 5 所示。

实验结果表明,在基类数据〔图 5(a)〕上,除 Flowers102 数据集外,其余 4 个数据集的模型分类准确率均出现了不同程度的提升。其中,Flowers102 数据集准确率仅略微下降了 0.14%,这一变化可能与该数据集内部样本特征分布的独特性或注意力机制对其基类特征表示的适配性有关。值得注意的是,在非遗美术图像分类数据集上,基类的准确率提升了 1.79%,显示出该模块在处理异质性较强的数据集时的显著优势。在不可见类别(New)的泛化性能方面〔图 5(b)〕,所有数据集的模型准确率均有所提升,进一步验证了注意力机制在增强模型特征提取和跨类别推广能力上的优势。尤其是在非遗美术图像分类数据集上,新类准确率的提升幅度高达 10.2%,为 5 个数据集中增幅最高的结果。这表明注意力机制能够有效

聚焦于关键特征,从而提高模型在复杂、细粒度分类任务中的泛化能力。此外,在综合性能指标调和均值 H 上〔图 5(c)〕,相较于未引入注意力机制的基线模型,加入注意力机制后模型在所有数据集上均取得了显著的性能提升。这一结果进一步说明,通过融合注意力机制,模型不仅能够更好地保留基类的特征表示,同时在新类的泛化性能方面也得到了全面优化。

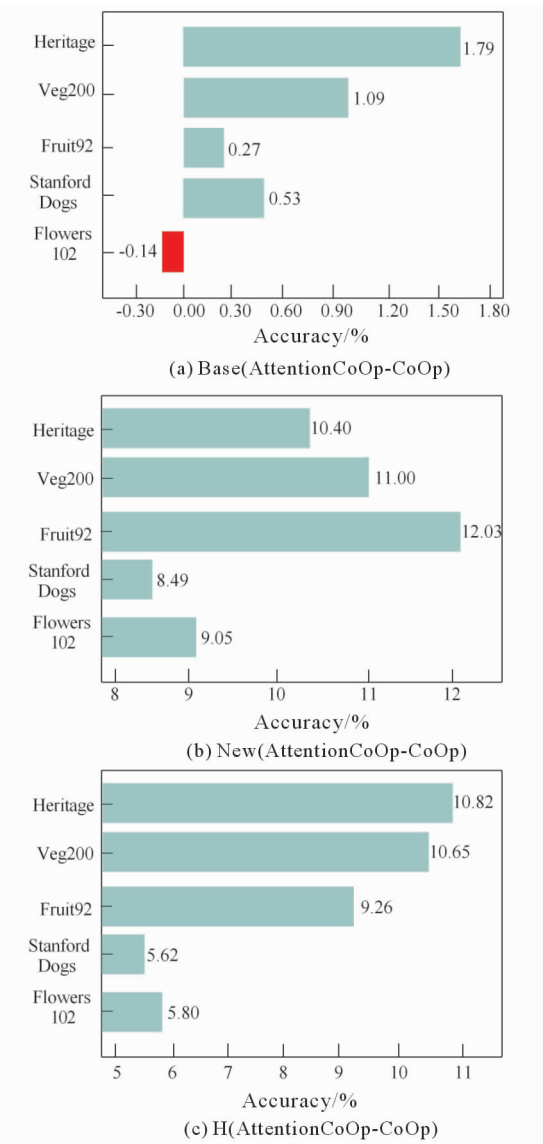

图 5　注意力机制对模型性能的影响

Fig. 5　The impact of attention mechanisms on mode performance

4　结论

本文提出的注意力机制驱动的上下文提示微调技术,有效解决了非遗美术图像分类中的数据获取难题和人工成本问题。在仅使用有限数据集和较低计算资源的情况下,该方法展现出了优异的分类表现。该方法通过结合可学习的文本嵌入与通用文本信息,仅需每个类别不超过 16 张图片的数据量,便能实现高识别精度。此外,它在维持基础类别识别精度的同时,显著提高了对新类别(未在训练中出现)的泛化能力,克服了 CoOp 和 CoCoOp 方法的缺陷。通过对多个基准测试的广泛评估,本方法被证实是一种既高效又精确的提示调整策略,其泛化能力尤为突出。考虑到其对资源的低需求和高运算效率,该策略在数字文化遗产保护领域具有极大的应用前景。

参考文献

[1] 黄永林. 中国非遗传承保护的四重价值[J]. 人民论坛·学术前沿,2024(1):76-83.
HUANG Y L. The protection and inheritance of Chinese intangible cultural heritage: Its quadruple values [J]. People's Forum: Academic Frontier, 2024(1): 76-83.

[2] 王燕妮. 中国民俗类非物质文化遗产分类研究[J]. 湖北民族学院学报(哲学社会科学版),2017,35(2):115-120.
WANG Y N. A study on the classification of chinese folklore intangible cultural heritage [J]. Journal of Hubei Minzu University (Philosophy and Social Sciences Edition), 2017, 35(2):115-120.

[3] 季长清,高志勇,秦静,等. 基于卷积神经网络的图像分类算法综述[J]. 计算机应用,2022,42(4):1044-1049.
JI C Q, GAO Z Y, QIN J, et al. Review of image classification algorithms based on convolutional neural network[J]. Computer Applications, 2022, 42(4): 1044-1049.

[4] LE CUN Y, BOSER B, DENKER J S, et al. Handwritten digit recognition with a back-propagation network[C]// Proceedings of the 3rd International Conference on Neural Information Processing Systems. ACM, 1989: 396-404.

[5] ISMAIL FAWAZ H, LUCAS B, FORESTIER G, et al. InceptionTime: Finding AlexNet for time series classification[J]. Data Mining and Knowledge Discovery, 2020, 34(6): 1936-1962.

[6] SENGUPTA A, YE Y T, WANG R, et al. Going deeper in spiking neural networks: VGG and residual architectures[J]. Frontiers in Neuroscience, 2019, 13: 95.

[7] ZHU Y, NEWSAM S. DenseNet for dense flow[C]// 2017 IEEE International Conference on Image Processing (ICIP). September 17-20, 2017. Beijing, China. IEEE, 2017: 790-794.

[8] KOONCE B. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization[M]. Berkeley, CA: Apress, 2021.

[9] 昝楠楠. 基于全局 CNN 与局部 LSTM 的国画图像分类算法[J]. 自动化技术与应用,2024,43(4):115-117.
SUI N N. Chinese painting image classification algorithm based on global CNN and local LSTM[J]. Automation Technology and Application, 2024, 43(4): 115-117.

[10] 生龙,马建飞,杨瑞欣,等. 基于特征交换的 CNN 图像分类算法研究[J]. 计算机工程,2020,46(9):268-273.
SHENG L, MA J F, YANG R X, et al. Research on CNN image classification algorithm based on feature exchange [J]. Computer Engineering, 2020, 46(9): 268-273.

[11] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]// Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 1597-1607.

[12] LIU X, ZHU Y, LIU L, et al. Feature-suppressed contrast for self-supervised food Pre-training[C]// Proceedings of the 31st ACM International Conference on Multimedia. 2023: 4359-4367.

[13] 朱若琳,蓝善祯,朱紫星. 视觉-语言多模态预训练模型前沿进展[J]. 中国传媒大学学报(自然科学版),2023,30(1):66-74.
ZHU R L, LAN S Z, ZHU Z X. Asurvey on vision-language multimodality pre-training [J]. Journal of Communication University of China (Natural Science Edition), 2023, 30(1):66-74.

[14] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International conference on machine learning. PMLR, 2021: 8748-8763.

[15] LEI Y, LI J, LI Z, et al. Prompt learning in computer vision: A survey[J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25(1): 42-63.

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

[17] ZHOU K Y, YANG J K, 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, 2022: 16816-16825.

[18] YAO H T, ZHANG R, XU C S. Visual-language prompttuning with knowledge-guided context optimization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6757-6767.

[19] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in neural information processing systems, 2021, 34: 9694-9705.

[20] GONDAL M W, GAST J, RUIZ I A, et al. Domain aligned CLIP for few-shot classification[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2024: 5709-5718.

[21] LONG S F, ZHAO Z, YUAN J K, et al. Task-oriented multi-modal mutual leaning for vision-language models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 21959-21969.

[22] PHAM C, NGUYEN V A, LE T, et al. Frequency attention for knowledge distillation[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2024: 2277-2286.

[23] CHORDIA S, PAWAR Y, KULKARNI S, et al. Attention is all you need to tell: Transformer-based image captioning[M]// Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2022. Singapore: Springer Nature Singapore, 2022: 607-617.

[24] LU J C, ZHANG J G, ZHU X T, et al. Softmax-free linear transformers[J]. International Journal of Computer Vision, 2024,132(8): 3355-3374.

[25] ZHANG J Y, HUANG J X, JIN S, et al. Vision-language models for vision tasks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024,46(8):5625-5644.

[26] XIA P P, ZHANG L, LI F Z. Learning similarity with cosine similarity ensemble[J]. Information sciences, 2015, 307: 39-52.

[27] SUN G Y, CHENG Y N, ZHANG Z X, et al. Text classification with improved word embedding and adaptive segmentation[J]. Expert Systems with Applications, 2024, 238: 121852.

[28] LASTRAS L A. Information theoretic lower bounds on

negative log likelihood[EB/OL]. 2019:1904. 06395. https://arxiv. org/abs/1904. 06395 vl.

[29] YU M X, WANG J, YOU R, et al. Multiple-local feature and attention fused person re-identification method [J]. Intelligent Data Analysis, 2024, 28(6): 1679-1695.

[30] GUO A Y, SHEN K, LIU J J. FE-FAIR: Feature-Enhanced Fused Attention for Image Super-Resolution [J]. Electronics, 2024, 13(6): 1075.

[31] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]// 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaheswar, India, IEEE, 2008: 722-729.

[32] ZHAO P S, XIE L X, ZHANG Y, et al. Universal-to-specific framework for complex action recognition[J]. IEEE Transactions on Multimedia, 2020, 23: 3441-3453.

[33] PINZÓN-ARENAS J O,JIMÉNEZ-MORENO R,PACHÓN-SUESCUN C G. ResSeg: Residual encoder-decoder convolutional neural network for food segmentation[J]. International Journal of Electrical & Computer Engineering (IJECE),2020,10(1):1017.

(编　辑　亢小玉)