基于图像内容理解的判别性类别提示学习

王楠井[1],刘阿建[2],梁凤梅[1*],张小梅[2],万 军[2],谢 珺[1],雷 震[2]
(1. 太原理工大学电子信息工程学院,山西晋中 030600;2. 中国科学院自动化研究所,北京 100190)

摘 要: 近年来,通过图像与文本的联合表示,基于对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)的方法将文本信息作为分类器的权值,在通用图像识别任务中展现出卓越性能. 但是现有方法仅单独构建类别文本提示,比如上下文优化(Context Optimization,CoOp)和条件上下文优化(Conditional Context Optimization,CoCoOp)等,没有考虑图像的内容语义信息与类别的重要性,限制了模型对图像类别的理解与判别. 为了解决上述问题,本文在CLIP的基础上提出了一种新方法:基于图像内容理解的判别性类别提示学习(Discriminative Category Prompt Learning based on image content understanding,DCPL),借助图像中丰富的内容特征来学习文本提示,提高文本提示对类别的判别性. 具体来说,DCPL包含提示生成(Prompt Generation,PG)模块和文本监督(Text Supervision,TS)模块. PG模块将图像特征和初始化的查询向量作为输入,通过自注意力机制和交叉注意力机制使输出的文本提示中包含充分的图像语义信息;TS模块将固定的类别提示模板作为监督,为可学习文本提示在类别层面和logits层面注入类别信息,增强了类别的重要性. 最后,DCPL在ImageNet、Caltech101和Oxford-Pets等11个公开分类数据集上的16-shots平均准确率达到了81.84%,较以往最优方法Cross-Modal的平均准确率提升了0.98个百分点.

关键词: 视觉-语言模型;图像识别;提示调优;注意力机制;文本监督(TS);适配器微调;transformer
基金项目: 虚拟现实技术与系统全国重点实验室开放课题(No.VRLAB2023A06);山西省科技合作交流专项(No.202104041101030)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)02-0493-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240522

Discriminative Category Prompt Learning Based on Image Content Understanding

WANG Nan-jing[1], LIU A-jian[2], LIANG Feng-mei[1*], ZHANG Xiao-mei[2], WAN Jun[2], XIE Jun[1], LEI Zhen[2]
(1. *College of Electronic Information Engineering, Taiyuan University of Technology, Jinzhong, Shanxi* 030600, *China*;
2. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: In recent years, the contrastive language-image pre-training (CLIP)-based method takes the text information as the weight of the classifier through the joint representation of image and text and shows excellent performance in the general image recognition task. However, the existing methods only construct text prompts of categories, such as context optimization (CoOp) and conditional context optimization (CoCoOp), without considering the importance of image content semantic information and categories, which limits the model's understanding and discrimination of image categories. To solve the above problems, this article proposes a new method based on CLIP: discriminative category prompt learning based on image content understanding (DCPL), which uses rich content features in images to learn text prompts and introduces manual templates to improve the discrimination of text prompts on categories. Specifically, DCPL includes a prompt generation module and a text supervision module: The prompt generation module takes image features and initialized query vectors as inputs, and makes the output text prompt contain sufficient image semantic information through the self/cross-attention mechanism; The text supervision module uses the fixed category prompt template as the supervision to inject category information into the category level and logits level for the learnable text prompt, increasing the importance of categories. Finally, the average accuracy of 16 shots of DCPL on 11 public classified datasets, such as ImageNet, Caltech101, Oxford pets, etc., is 81.84%, The average accuracy has increased by 0.98 percentage points compared with that

收稿日期:2024-06-05;修回日期:2024-10-30;责任编辑:覃怀银
*通讯作者:梁凤梅

of the previous optimal method, Cross-Modal.

Key words: visual-language model; image recognition; prompt tuning; attention mechanism; text supervision; adapter tuning; transformer

Foundation Item(s): Open Research Fund Project of National Key Laboratory of Virtual Reality Technology and Systems (No.VRLAB2023A06); Shanxi Science and Technology Cooperation and Exchange Project (No.202104041101030)

1 引言

早期图像识别方法依赖传统特征算子,如像素强度、边缘检测、角点检测等以提取有效特征进行分类识别. 随着机器学习的发展,卷积神经网络(Convolutional Neural Network,CNN)成为研究重点,其核心是通过卷积核扫描图像检测特定特征,随着网络层次加深,特征从简单边缘纹理变为复杂模式. AlexNet[1]、视觉几何组网络(Visual Geometry Group Network,VGGNet)[2]、GoogLeNet[3]和残差网络(Residual Network,ResNet)[4]等基于CNN的方法在通用图像分类和细粒度图像分类上效果较好. 虽然CNN在图像识别中取得巨大成功,但视觉单模态识别有时难以处理复杂或类别模糊的图像,引入文本等模态数据可提供更多信息,使图像识别更精确全面.

为了通过多模态融合来提升通用图像识别任务的效果,视觉语言模型[5](Visual Language Models,VLM)应运而生,标志着计算机视觉与自然语言处理(Natural Language Processing,NLP)领域的融合. VLM的代表性方法对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)[6]利用对比学习将4亿个图像-文本对齐,其显著的进步是通过计算图像特征和手工文本提示特征之间的相似度来实现开放词汇识别的能力,且无需额外训练,其在物体检测[7]、图像生成[8]和视觉问题回答[9]等领域有着优秀的表现. 尽管如此,在特定场景或类型的图像文本配对上CLIP可能表现不佳,特别是在与训练数据显著不同的新领域上,其识别准确度低于现有CNN方法,如何更好地将CLIP迁移至下游任务成为新的研究重点.

对CLIP进行提示调优是一种有效的迁移方式,该方法通过精心设计的提示语句引导模型的学习,可以在不微调基础模型的前提下使CLIP与下游任务相匹配. 在该方向上的典型方法是上下文优化(Context Optimization,CoOp)[10],与CLIP相比,CoOp将原有的固定提示模板替换为一组可学习的向量,在训练过程中,通过基于交叉熵的分类损失来优化这些可学习提示向量. 在CoOp的基础上,条件上下文优化(Conditional Context Optimization,CoCoOp)[11]通过学习一个轻量级神经网络,为每个图像生成输入条件向量,并与可学习提示相结合,使其更适应每个实例,减少了提示对类别偏移的敏感性. 与CoOp和CoCoOp不同,多模态提示学习(Multi-modal Prompt Learning,MaPLe)[12]使用固定的提示模板,在视觉和语言分支中学习上下文提示,通过耦合函数逐级将视觉提示与语言提示相关联,以改善视觉和语言表示之间的对齐,从而提高模型在下游任务中的泛化能力.

尽管提示调优为提高CLIP模型在下游任务上的识别准确度和泛化性带来了显著进展,但是现有方法仅单独构建类别的文本提示,如图1(*a*)所示,没有考虑图像实例所包含的丰富内容语义信息,以及类别在提示中的重要性,这样会限制模型对图像的理解和对类别的判别性. 针对上述问题,本文提出了基于图像内容理解的判别性类别提示学习(Discriminative Category Prompt Learning based on image content understanding,DCPL),如图1(*b*)所示,其在继承CLIP图像文本编码器的基础上添加了提示生成(Prompt Generation,PG)模块和文本监督(Text Supervision,TS)模块. 和已有的方法相比,DCPL更注重于从图像中提取内容语义信息,加强了可学习提示与图像的关联性,并在训练的上游阶段可学习提示就能从图像中获得丰富且有效的内容信息. 本文设计了基于注意力[13,14]机制的PG模块,直接通过输入图像特征和初始化查询向量,为文本提示注入丰富的图像内容信息,在避免手动调整提示的同时加深了模型对不同类别图像间微小差异的理解.

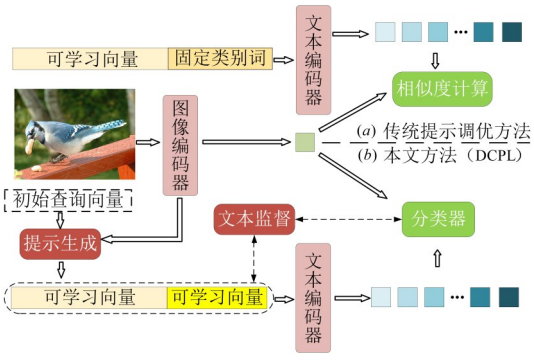


图1 传统提示调优方法与DCPL对比

进一步地,在提高可学习提示的类判别性这一方向上,本文采用了CLIP的固定提示模板,对可学习提示进行监督,借助固定提示模板中的类别信息,能够增强可学习提示的类判别性,有效防止训练结果出现过拟合现象. 本文从文本模态切入,引入了TS模块,分别在

类别与logits两个层面优化可学习提示,以使可学习提示能够充分汲取类别信息.

DCPL在未来有着令人期待的应用前景:在智能安防领域,可依据其对图像内容的深入理解,精确分析监控画面,识别安全威胁;在医疗领域,利用TS模块汲取类别信息,区分肿瘤影像的良性与恶性病变;在工业生产中,能够高效检测和分类产品外观,依靠可学习提示与图像的强关联性,及时发现瑕疵品,确保质量一致性;在文化艺术领域,准确分类文物和艺术作品图像,为文化遗产保护等提供支持.

2　相关工作

2.1　VLM

近年来,大语言模型(Large Language Model,LLM)的快速发展为构建视觉-语言模型提供了可行的解决方案. 视觉-语言模型将LLM与基础视觉模型相结合,在添加少量连接参数进行训练后,能够有效地理解图像和文本. CLIP的提出使视觉-语言模型上升到一个新的高度,其使用大规模图像-文本对进行对比训练,在视觉表征学习方面取得了显著的进展. 在此基础上,进一步探索如何利用CLIP来增强视觉识别任务. 为了更深层次地融合可学习提示特征和视觉特征,本文从BLIP-2[15](Bootstrapping Language-Image Pre-training)获得启发,引入一个基于注意力机制的PG模块,将视觉和语言模态直接连接,从而为提示注入丰富的图像特征信息.

2.2　提示调优

提示学习源于NLP领域. 在GPT-2(Generative Pre-trained Transformer 2.0)[16,17]中,预训练语言模型在未经微调的情况下,通过添加前缀描述的方式对下游任务进行了适配. 在VLM领域,CoOp率先提出了通过可学习的提示模板对模型进行调整的方法,为这一领域的研究开辟了崭新的方向. 进一步地,CoCoOp引入了轻量级神经网络为每个图像生成输入条件令牌,从而学习到了泛化性更好的提示. KgCoOp[18]通过一般知识来约束经过学习的提示嵌入以增强泛化能力,显著地提升了模型对图像与文本关联性的理解能力. MaPLe则在文本和图像编码器每一层的隐藏表示中附加了软提示,有效提升了模型在few-shot级别的图像识别任务中的性能. 在先前的研究中,多模态提示调优主要聚焦于优化类别描述,忽视了类别词本身在这些模板中的关键作用. 文献[19]创新性地提出了基于搜索的方法,在训练过程中自动筛选出更为合适的类别表述. 同时,还有研究[20]提出采用可训练向量作为软标签表述,来替代传统的类别词汇,有效避免了在庞大词典中搜索的困难. DCPL中的可学习文本提示通过TS模块从类别和logits层面入手,从词汇和句子两个角度为生成提示提供完善的类别信息.

3　判别性类别提示学习

如图2所示,DCPL框架主要包含如下模块:基于CLIP的图像与文本编码器、PG模块和TS模块. 本节将对DCPL的各个模块以及总体流程进行详细介绍.

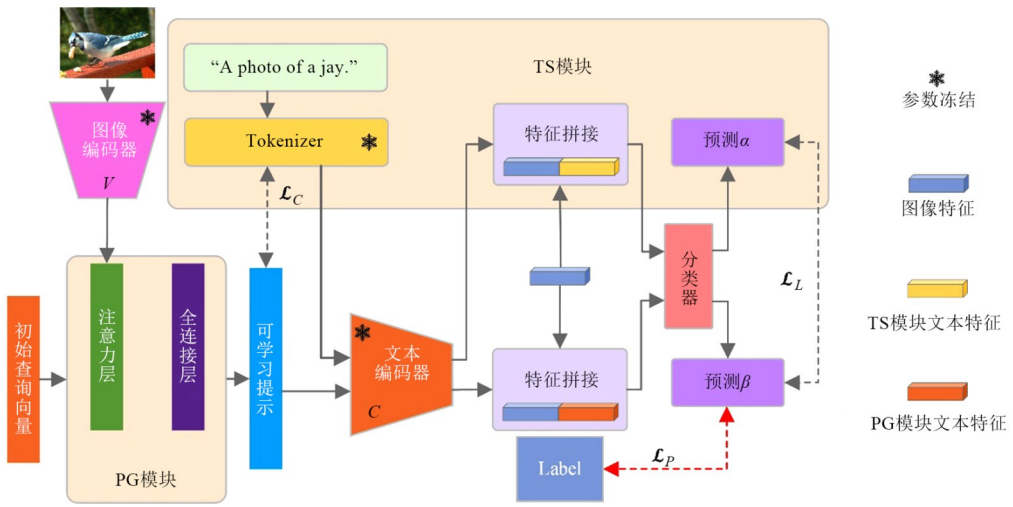


图2　DCPL总体流程示意图

3.1　CLIP模型

CLIP模型在2021年由OpenAI提出,它从自然语言中学习有效的视觉表示方法,其工作原理是在同一个嵌入空间中将图像和文本的特征对齐,通过对比学习的方式实现跨模态理解. 从总体上看,CLIP包含两大部分,分别是图像编码器V和文本编码器C,在训练的

过程中，两个编码器分别提取对应的文本和图像特征，在最大化配对文本和图像特征的余弦相似度的同时，最小化不配对图像和文本特征之间的余弦相似度，这样的训练方式使得CLIP模型能够学习到跨模态的语义关系，确保匹配的图像和文本在特征空间中更为接近，而不匹配的图像和文本更为远离．具体而言，假设一个数据集D包含N个类别，将其表示为$D\{\boldsymbol{I}_i,\boldsymbol{T}_i\}_{i=0}^{N-1}$，其中$\boldsymbol{I}_i$为图像，$\boldsymbol{T}_i$为$\boldsymbol{I}_i$对应文本描述的编码．在CLIP中，文本描述一般被设计为“a photo of a [Class$_i$]．”，i=0，1，…，$N-1$，其中Class$_i$是数据集中每个类别对应的名称．在训练过程中，图像通过V得到512维度的图像特征$\boldsymbol{v}_i=V(\boldsymbol{I}_i)$，文本描述通过$C$得到维度同样为512的文本特征$\boldsymbol{t}_i=C(\boldsymbol{T}_i)$．训练时，匹配的图像和文本特征对被视为正对$\{\boldsymbol{v}_i,\boldsymbol{t}_i\}$，若不配对则被视为负对$\{\boldsymbol{v}_i,\boldsymbol{t}_j|i\neq j\}$，CLIP最大化正对的余弦相似度，并且将负对的余弦相似度最小化．在进行零样本推理时，对类别的预测概率计算式为

$$z(\hat{y}|\boldsymbol{I})=\frac{\exp\left(\cos\left(\boldsymbol{v},\boldsymbol{t}_{\hat{y}}\right)/\tau\right)}{\sum_{i=0}^{N-1}\exp\left(\cos\left(\boldsymbol{v},\boldsymbol{t}_i\right)/\tau\right)} \tag{1}$$

其中，$\hat{y}$为图像与文本特征相似度最高的类别；$\cos(\cdot,\cdot)$为余弦相似度计算；τ为从CLIP获取的温度参数．

3.2 PG模块的设计

为了让CLIP模型快速迁移到下游任务，设计了基于注意力机制的PG模块，如图3所示．注意力机制是一种模仿人类注意力分配机制的机器学习技术，使模型可以学习如何在输入数据中选择性地关注重要的部分，其核心思想是让模型在处理输入数据时，根据每个输入的重要性来分配不同的权重，使模型可以更加关注与当前任务相关的信息，忽略一些无关或不重要的信息．在PG模块中随机生成K个大小为[1，512]的可学习查询向量$\boldsymbol{q}$，其数值符合均值为0，且标准差为0.02的正态分布，它们共同构成一组大小为$[B,K,512]$的初始查询向量$\boldsymbol{Q}$，其中B为批处理大小．将$\boldsymbol{Q}$和由CLIP图像编码器V获得的图像特征$\boldsymbol{v}$送入PG模块中，通过如式(2)所示步骤生成用于输入文本编码器的可学习提示$\boldsymbol{p}$．

$$\begin{cases}\boldsymbol{Q}_a=\boldsymbol{Q}+\mathrm{SA}\left(\mathrm{LN}\left(\boldsymbol{Q}\right)\right)\\ \boldsymbol{Q}_b=\boldsymbol{Q}_a+\mathrm{CA}\left(\mathrm{LN}\left(\boldsymbol{Q}_a\right),\mathrm{LN}\left(\boldsymbol{v}\right)\right)\\ \boldsymbol{p}=\boldsymbol{Q}_b+\mathrm{FFN}\left(\mathrm{LN}\left(\boldsymbol{Q}_b\right)\right)\end{cases} \tag{2}$$

其中，LN(·)为层归一化操作；SA(·)为自注意力操作，使$\boldsymbol{Q}$内部包含的可学习查询之间建立依赖关系；CA(·，·)为交叉注意力操作，使$\boldsymbol{Q}_a$与$\boldsymbol{v}$建立可学习查询与图像特征之间的依赖关系；最后的FFN(·)由两个全连接层组成，$\boldsymbol{Q}$在训练过程中通过梯度反向传播更新参数．通过PG模块，$\boldsymbol{p}$在理解语言模态信息的同时还可以捕捉到视觉-语言模态之间的交互依赖关系，从而学习到图像中丰富的视觉信息，能够检测到不同类别图像之间的微小差异．

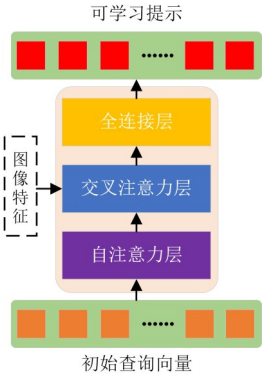


图3 PG模块内部结构

3.3 TS模块的设计

由于DCPL在通过图像实例生成可学习提示的同时冻结了CLIP中的参数，因此，在少样本训练方式下可能会使可学习提示过拟合到特定的图片上，偏离了真正的类别，从而导致类判别能力变差．为此，本文设计了TS模块来优化可学习文本提示，它从两个层面对PG模块的输出进行监督，分别是类别层面和logits层面．如式(3)所示，TS模块将手工选择的固定文本提示P_i'作为教师提示进行监督，然后使用预训练文本编码器的分词器(Tokenizer)将其进行编码操作，得到编码后的手工模板$\boldsymbol{p}_i'$：

$$\begin{cases}\boldsymbol{p}_i'=\mathrm{Tokenizer}\left(P_i'\right)\\ \boldsymbol{p}_i'=\left(s_1,s_2,\cdots,s_k\right)C_i\\ \boldsymbol{p}_i=\left[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_k\right]\left[\boldsymbol{X}_i\right]\end{cases} \tag{3}$$

其中，$\left(s_1,s_2,\cdots,s_k\right)$为编码后的手工提示模板；$\left[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_k\right]$为对类别进行描述的可学习向量；$\boldsymbol{X}_i$为与$\boldsymbol{p}_i'$的类别标签$C_i$相同位置对应的可学习类别向量；$k$为手工固定模板类别描述部分的词汇个数．在类别层面上，TS模块将$\boldsymbol{p}_i'$映射到与$\boldsymbol{p}_i$相同的512维度特征空间上，在手工模板与可学习提示之间建立起约束关系，用如式(4)方式计算类别损失：

$$\mathcal{L}_C=\mathrm{MSE}\left(\boldsymbol{X}_i,\mathrm{Em}\left(C_i\right)\right) \tag{4}$$

其中，MSE为均方误差损失函数；Em(·)为映射操作．

Logits层面上，虽然在训练过程中冻结了V和C的所有参数，且特征空间中的图像表示保持不变，但可学习提示的参数在反向传播时可能会改变CLIP特征空间中的文本表示，导致模型的类判别能力降低．因此，为

了减小DCPL在训练过程中在文本特征空间上与CLIP的偏移,设计了一个基于知识蒸馏的logits损失如式(5)所示,$\boldsymbol{v}$为图像特征,$\boldsymbol{t}'$和$\boldsymbol{t}$分别为对应的固定提示特征和可学习提示特征,Fc(·)是一个包含全连接层的分类器,$\boldsymbol{\alpha}$为基于固定文本提示得到的输出结果,$\boldsymbol{\beta}$为基于可学习提示得到的输出结果,[·,·]为特征拼接操作,T为温度参数,在本文实验中将其设置为2.

$$\begin{cases}\boldsymbol{\alpha}=\mathrm{Fc}\left(\left[\boldsymbol{v},\boldsymbol{t}'\right]\right)\\ \boldsymbol{\beta}=\mathrm{Fc}\left(\left[\boldsymbol{v},\boldsymbol{t}\right]\right)\\ \mathcal{L}_L=T^2\sum_{i=0}^{N-1}\boldsymbol{\alpha}_i\Big/T\ln\left(\boldsymbol{\beta}_i/T\right)\end{cases}\tag{5}$$

3.4　总体流程设计

如图2所示,DCPL继承了CLIP的文本编码器C和图像编码器V,并冻结其参数,在此基础上引入了两个模块,分别是PG模块和TS模块,PG模块将图像特征和初始化查询向量作为输入,生成可学习提示;TS模块将手工模板提示比如"a photo of a jay."作为教师提示,对生成的可学习提示在类别和logits两个层面进行监督.

最后,模型输出形状均为[B,512]的文本特征$\boldsymbol{t}$与图像特征$\boldsymbol{v}$,在进行简单的拼接操作后,变为形状为[B,1 024]的融合特征,通过分类器得到最终的预测结果,将预测结果与对应的标签进行分类损失计算:

$$\mathcal{L}_P=\mathrm{CE}\left(\boldsymbol{\beta}_i,\boldsymbol{y}\right)\tag{6}$$

其中,$\boldsymbol{y}$为输入图像的对应标签;$\mathrm{CE}(\cdot,\cdot)$为交叉熵损失函数.

综上所述,DCPL在继承CLIP预训练图像-文本编码器的基础上,添加了负责生成可学习提示的PG模块和负责优化可学习提示的TS模块,训练过程中的总损失函数为

$$\mathcal{L}=\mathcal{L}_P+\mathcal{L}_C+\lambda*\mathcal{L}_L\tag{7}$$

其中,λ为$\mathcal{L}_L$的权重超参数,在本文实验中将其设置为0.005.

在训练结束后的推理阶段,DCPL不再需要手工模板的监督,PG模块会通过输入的图像自动生成合适的提示,并将其输入到文本编码器中.由可学习提示得到的文本特征在与图像特征合并后被传输到分类器中,以输出最后的预测结果.

4　实验

4.1　数据集及实验环境

本文采用了11个公开图像分类数据集进行评估,表1介绍了每个数据集的类别数量,包括训练集、验证集和测试集的样本数量以及在训练时使用的手工模板,这些数据集涵盖了不同的场景和规模.通用分类方向的数据集有ImageNet[21]、Caltech101[22]和SUN397[23],专业分类方向有EuroSAT[24]、UCF101[25]和DTD(Describable Textures Dataset)[26],细粒度分类方向包含Food101[27]、StanfordCars[28]、FGVCAircraft[29]、OxfordPets[30]和Flowers102[31],这些数据集不仅包含对通用对象、场景、动作和细粒度类别的分类任务,还包含了识别纹理与卫星图像等较为专业的任务.本节的实验使用Pytorch框架实现,使用的计算机处理器为Intel(R)Xeon(R)Gold5220CPU,内存为196 GB,GPU为两块显存12 GB的RTX2080ti显卡.采用的编程语言是Python,程序均在Ubuntu20.04系统下运行.

4.2　实验设置

在训练过程中,对数据集的每个类别随机抽取16张图片作为训练集,并在完整的测试集上进行准确率测试.将ViT-B/16设置为图像编码器,初始化提示查询向量的个数设置为32,输入图片的大小为224×224.采用AdamW优化器[32],学习率为5×10^{-5},权重衰减为1×10^{-4},初始学习率设为1×10^{-5},并采用cosine的方式来降低学习率,批处理大小为8,迭代次数为12 800次.在实验过程中,对每个数据集分别使用表1中相对应的手工模板作为TS,这些手工模板从Tip-Adapter[33]中获得.

另外,除了直接冻结预训练模型的权重,一些微调方法比如Convpass[34](Convolutional bypasses)也是一种提高模型性能的有效方式.如图4所示的Convpass适配器由三个卷积层组成,分别是一个1×1卷积层、一个3×3卷积层和一个1×1卷积层,通过引入卷积操作在ViT中加入了视觉归纳偏差,从而更有效地适应下游视觉任务,在数据有限的情况下提升效果尤为明显.在训练过程中,transformer层的参数全部被冻结,只有Convpass模块中的参数通过反向传播进行更新.经过实验研究发现,将Convpass应用在transformer残差注意力块的多层感知器(MultiLayer Perceptron,MLP)层,能取得较好的提升效果,为模型性能的优化提供了新的思路和方法.

4.3　实验结果

主要从对比试验、域泛化实验和消融实验三方面来评估DCPL的性能,测试指标为图像识别准确率.

4.3.1　与传统提示调优方法的对比实验

为验证本文方法的性能,进行了详尽的比较研究,如表2~表4所示,涵盖了当前的多种先进方法,包括CoOp、CoCoOp、MaPLe、Tip-Adapter和较新的方法Cross-Modal[35],其中,最佳结果和次佳结果分别以加粗字体和下划线字体表示.本文方法在测试中展现了较好的性能,通用分类任务方面的准确度均值为81.82%,其中,在ImageNet、SUN397上的准确度较以往最优方法Cross-Modal分别提升了0.39个百分点、0.17个百分点;专业分类任务方面的均值为81.52%,其中,在DTD上的

表1 数据集详细信息

数据集	类别/个	训练/张	验证/张	测试/张	手工模板
Caltech101	100	4 128	1 649	2 465	a photo of a [class].
OxfordPets	37	2 944	736	3 669	a photo of a [class], a type of pet.
StanfordCars	196	6 509	1 635	8 041	a photo of a [class].
Flowers102	102	4 093	1 633	2 463	a photo of a [class], a type of flower.
Food101	101	50 500	20 200	30 300	a photo of a [class], a type of food.
FGVCAircraft	100	3 334	3 333	3 333	a photo of a [class], a type of aircraft.
SUN397	397	15 880	3 970	19 850	a photo of a [class].
DTD	47	2 820	1 128	1 692	[class] texture.
EuroSAT	10	13 500	5 400	8 100	a centered satellite photo of [class].
UCF101	101	7 639	1 898	3 783	a photo of a person doing [class].
ImageNet	1 000	1 281 167	-	50 000	itap of a [class]. a bad photo of the [class]. a origami [class]. a photo of the large [class]. a [class] in a video game. art of the [class]. a photo of the small [class].

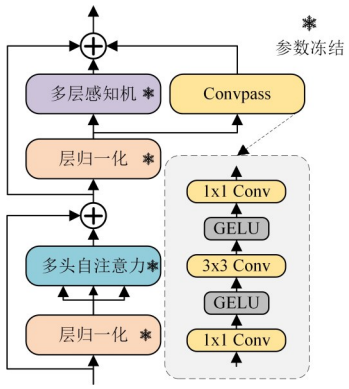


图4 Convpass适配器

准确率较以往最优方法Cross-Modal提升了0.33个百分点；细粒度分类任务方面的均值为82.05%，其中，在FGVCAircraft、Flowers102和StanfordCars上的准确率较以往最优方法Cross-Modal分别提升了4.60、0.86、0.72个百分点. 最后，DCPL在11个数据集上的均值为81.84%，较以往最优方法Cross-Modal的均值提升0.98个百分点. 通过在三种类型数据集上的测试发现，DCPL在细粒度分类任务上的提升效果最为明显，说明PM模块在训练过程中，从图像实例中捕获到了不同类别之间的微小特征差异，使其生成的可学习提示具有区分细小差异类别的能力. 在DCPL的transformer残差注意力块上添加对少量参数进行微调的Convpass适配器后，改进方法DCPL-F以较大的余量超过了传统方法，在11个数据集上的均值为82.48%，较DCPL的均值提升0.64个百分点，证明了适配器对视觉-语言模型微调的有效性.

表2 通用分类数据集上的识别准确度实验 单位:%

方法	准确率			
	ImageNet	Caltech101	SUN397	均值
CLIP	66.84	93.04	62.43	74.10
CoOp	71.49	95.71	74.37	80.52
CoCoOp	71.32	95.36	74.09	80.26
Tip-Adapter	70.69	95.11	72.06	79.29
MaPLe	71.90	95.30	76.00	81.07
Cross-Modal	72.74	96.20	76.18	81.71
DCPL	73.13	95.97	76.35	81.82
DCPL-F	**73.81**	**96.40**	**76.61**	**82.27**

表3 专业分类数据集上的识别准确度实验 单位:%

方法	准确率			
	EuroSAT	UCF101	DTD	均值
CLIP	46.03	67.24	44.96	52.74
CoOp	81.24	82.47	67.93	77.21
CoCoOp	79.81	81.37	68.52	76.57
Tip-Adapter	78.35	78.46	66.12	74.31
MaPLe	88.47	82.26	69.77	80.17
Cross-Modal	83.71	84.00	72.46	80.06
DCPL	88.01	83.76	72.79	81.52
DCPL-F	**91.75**	**84.84**	**72.93**	**83.17**

4.3.2 域泛化性能实验

为了对本文方法的域泛化性能进行评估，保存了在ImageNet上的训练权重并在ImageNet-V2[36]和ImageNet-Sketch[37]上进行测试. ImageNet-V2数据集的图像样本是从ImageNet中重新采样和验证而来，目的

表4　细粒度分类数据集上的识别准确度实验　　单位:%

方法	准确率					
	StanfordCars	FGVCAircraft	OxfordPets	Flowers102	Food101	均值
CLIP	65.33	25.06	89.22	71.29	86.11	67.40
CoOp	77.06	39.11	92.85	96.02	86.47	78.30
CoCoOp	76.88	38.74	92.85	95.63	**86.97**	78.21
Tip-Adapter	75.27	39.79	91.88	94.53	86.43	77.58
MaPLe	80.13	42.56	93.10	96.70	86.53	79.80
Cross-Modal	83.31	44.51	92.73	97.16	86.47	80.84
DCPL	84.03	**49.11**	92.42	**98.02**	86.66	82.05
DCPL-F	**85.53**	48.32	**93.27**	96.90	86.91	**82.19**

是提供更具挑战性和真实性的数据来评估图像分类模型的性能;ImageNet-Sketch数据集包含50 000张素描风格的图像,覆盖了ImageNet中的1 000个类别,主要用于研究图像分类任务中手绘风格的识别,与ImageNet相比其识别难度更高. 这三个数据集具有相同的类别,但它们的域分布有所区别,因此,可以很好地验证DCPL的域迁移能力.

由表5可知,DCPL在ImageNet-V2上的准确率领先传统方法,在ImageNet-Sketch上的准确率较以往方法而言也有较好的成绩,在三个数据集上的平均准确率为62.09%,较Cross-Modal的平均准确率领先0.27个百分点. 通过将ImageNet扩展到分布外的数据集,证明了DCPL优越的域泛化性能,也证明DCPL在处理域分布变化方面有着巨大的潜力.

表5　域泛化性能对比实验　　单位:%

方法	源	目标		均值
	ImageNet	-V2	-Sketch	
CLIP	68.57	60.84	46.15	58.52
CoOp	71.49	64.32	47.92	61.24
CoCoOp	71.32	64.12	48.75	61.40
MaPLe	71.90	64.06	49.22	61.73
Cross-Modal	72.78	64.76	47.93	61.82
DCPL	73.13	65.28	47.87	62.09

4.3.3　消融实验

在本次研究中,对DCPL进行了模块消融实验,旨在探究不同模块对模型性能的影响. 如表6所示,实验设置了三种模型变体,分别是完整的DCPL模型、去除PG模块的DCPL-NO PG以及去除PG和TS模块的DCPL-NO PG&TS,并在4个不同的数据集上进行了测试,以准确率作为评估指标. 实验结果表明,完整的DCPL模型在四个数据集上均取得较高准确率.

表6　DCPL模块消融实验　　单位:%

方法	准确率			
	Caltech101	FGVCAircraft	Flowers102	StanfordCars
DCPL	**95.97**	**49.11**	**98.02**	**84.03**
DCPL-NO PG	95.81	44.80	97.30	82.83
DCPL-NO PG&TS	95.73	44.27	97.24	82.73

当去除PG模块后,各数据集上的准确率有所下降,其中,在FGVCAircraft数据集上下降较为明显,这是因为缺少了PG模块,其模糊的类别名称如“707-320”“737-300”等给模型的识别和分类带来了困难. 进一步去除PG和TS模块后,准确率继续降低,这充分说明PG模块和TS模块在DCPL模型中起着重要作用,它们共同为模型性能的提升贡献力量. 同时,在不同的数据集上两个模块的影响程度有所差异,这反映出不同数据集的特点和任务需求会导致模块重要性的变化.

随后,以StanfordCars数据集为基础,进行了DCPL模型的t-SNE可视化实验. 如图5所示,实验中设置了不同的模型变体,包括完整的DCPL模型、去除TS模块的DCPL-No TS以及去除PG模块和TS模块的DCPL-No PG&TS. 从可视化结果来看,完整的DCPL模型能够较好地对数据进行特征提取和分类,特征分布较为清晰.

而当去除TS模块和PG模块后,模型的特征分布出现了一定程度的变化,可能会影响模型的性能. 这进一步验证了上文模块消融实验中得出的结论,即TS模块和PG模块在DCPL模型中起着重要作用.

为了验证TS模块在类别层面和logits层面上对生成的可学习提示进行约束的有效性,本文在所有数据集上对这两个层面的损失进行了消融研究,实验结果为11个数据集结果的均值. 表7中“类别”和“Logits”分别表示是否启用TS模块的类别监督和logits监督,可以

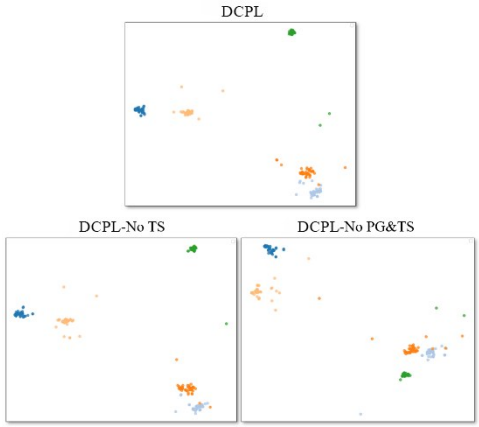


图5 t-SNE可视化实验

发现两个层面上的监督对模型都起到了提升作用，但是由于类别层面与logits层面的特征会互相影响，在两个层面共同监督下的效果并不会以各自的提升线性叠加.

表7 TS消融实验

类别	Logits	均值/%
		81.48
√		81.71
	√	81.72
√	√	**81.84**

如何确定有效的适配器添加位置也是值得研究的地方，本文分别在transformer层的不同位置添加Convpass适配器重新训练模型，并在包含1 000个类别的ImageNet数据集上进行测试. 由表8可知，将适配器放置于MLP层或注意力层对模型的性能都有较好的提升，但是将Convpass添加在MLP层是最佳的选择. 关于$\mathcal{L}_L$的权重超参数λ的选择，分别用0.05、0.01、0.005和0.001这4个参数在ImageNet数据集上进行测试，结果如表9所示，0.005是一个最优的选择.

表8 适配器添加位置对比实验

MLP层	注意力层	准确率/%
		73.13
√		**73.81**
	√	73.70
√	√	73.66

表9 超参数选择实验

超参数	准确率/%
0.050	72.97
0.010	72.00
0.005	**73.13**
0.001	72.21

5 结论

与传统的小样本学习方法相比，基于CLIP等VLM的小样本提示学习在下游任务中有着较好的泛化性，若想继续提高这一类方法的效率，从提示模板的标签表示入手是一个合适的方向. 本文提出了一种高效提示学习方法DCPL，通过PG模块直接对齐可学习提示与图像特征，同时引入了TS模块，从类别与logits两个层面分别指导提示的优化. 本文方法基于样本实例生成可学习提示，降低了模型对预定义类别名称的依赖，通过包含类别与logits的双层损失来进行约束，提高了DCPL的类判别性. 在通用分类、细粒度分类和专业分类的数据集上进行的广泛实验可以充分证明本文所提出模型框架的有效性. 在未来的工作中，如何更高效地在各个特征层面对可学习提示进行文本监督将是研究的重点.

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2024-06-05]. https://arxiv.org/abs/1409.1556v6.

[3] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1-9.

[4] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[5] 殷炯, 张哲东, 高宇涵, 等. 视觉语言预训练综述[J]. 软件学报, 2023, 34(5): 2000-2023.
YIN J, ZHANG Z D, GAO Y H, et al. Survey on vision-language pre-training[J]. Journal of Software, 2023, 34(5): 2000-2023. (in Chinese)

[6] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. New York: PMLR, 2021, 139: 8748-8763.

[7] GU X Y, LIN T Y, KUO W C, et al. Open-vocabulary object detection via vision and language knowledge distillation[EB/OL]. (2021-04-28)[2024-06-05]. https://arxiv.org/abs/2104.13921v3.

[8] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language under-

standing[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022, 35: 36479-36494.

[9] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: A visual language model for few-shot learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[10] 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.

[11] ZHOU K Y, YANG J K, LOY C C, et al. Conditional prompt learning for vision-language models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16795-16804.

[12] 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 (CVPR). Piscataway: IEEE, 2023: 19113-19122.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, New York: Curran Associates Inc., 2017: 6000-6010.

[14] CHEN C F R, FAN Q F, PANDA R. CrossViT: Cross-attention multi-scale vision transformer for image classification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 347-356.

[15] LI J N, LI D X, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. New York: PMLR, 2023: 19730-19742.

[16] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[17] 廖宁, 曹敏, 严骏驰 . 视觉提示学习综述[J]. 计算机学报, 2024, 47(4): 790-820.
LIAO N, CAO M, YAN J C. Visual prompt learning: A survey[J]. Chinese Journal of Computers, 2024, 47(4): 790-820. (in Chinese)

[18] YAO H T, ZHANG R, XU C S. Visual-language prompt tuning with knowledge-guided context optimization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 6757-6767.

[19] GAO T Y, FISCH A, CHEN D Q. Making pre-trained language models better few-shot learners[EB/OL]. (2020-12-31)[2024-06-05]. https://arxiv.org/abs/2012.15723v2.

[20] CUI G Q, HU S D, DING N, et al. Prototypical verbalizer for prompt-based few-shot tuning[EB/OL]. (2022-3-18)[2024-06-05]. https://arxiv.org/abs/2203.09770v1.

[21] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

[22] LI F F, FERGUS R, PERONA P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories[C]//2004 Conference on Computer Vision and Pattern Recognition Workshop. Piscataway: IEEE, 2005: 178.

[23] XIAO J X, HAYS J, EHINGER K A, et al. SUN database: Large-scale scene recognition from abbey to zoo[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2010: 3485-3492.

[24] HELBER P, BISCHKE B, DENGEL A, et al. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(7): 2217-2226.

[25] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[EB/OL]. (2012-12-03)[2024-06-05].https://arxiv.org/abs/1212.0402v1.

[26] 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.

[27] BOSSARD L, GUILLAUMIN M, VAN GOOL L. Food-101 - mining discriminative components with random forests[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014: 446-461.

[28] KRAUSE J, STARK M, JIA D, et al. 3D object representations for fine-grained categorization[C]//2013 IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE, 2013: 554-561.

[29] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. (2013-06-21)[2024-06-05]. https://arxiv.org/abs/1306.5151v1.

[30] PARKHI O M, VEDALDI A, ZISSERMAN A, et al.

Cats and dogs[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3498-3505.

[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. Piscataway: IEEE, 2008: 722-729.

[32] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. (2017-11-14)[2024-06-05]. https://arxiv.org/abs/1711.05101v3.

[33] ZHANG R R, FANG R Y, ZHANG W, et al. Tip-adapter: Training-free CLIP-adapter for better vision-language modeling[EB/OL]. (2021-11-06)[2024-06-05]. https://arxiv.org/abs/2111.03930v2.

[34] JIE S B, DENG Z H. Convolutional bypasses are better vision transformer adapters[EB/OL]. (2022-07-14)[2024-06-05]. https://arxiv.org/abs/2207.07039v3.

[35] LIN Z Q, YU S, KUANG Z Y, et al. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 19325-19337.

[36] RECHT B, ROELOFS R, SCHMIDT L, et al. Do imagenet classifiers generalize to imagenet? [C]//International Conference on Machine Learning. New York: PMLR, 2019: 5389-5400.

[37] WANG H H, GE S W, LIPTON Z, et al. Learning robust global representations by penalizing local predictive power[C]//33rd Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2019: 10506-10518.

作者简介

王楠井 男,2000年6月出生于山西省吕梁市. 现为太原理工大学电子信息工程学院硕士研究生. 主要研究方向为计算机视觉、图像识别与处理.

E-mail: 2632835656@qq.com

刘阿建 男,1992年6月出生于山西省运城市. 中国科学院自动化研究所助理研究员,“澳门青年学者计划”“基金委青年科学基金”“北京市青年人才托举工程”“博士后面上资助项目”获得者. 主要研究方向为计算机视觉、图像识别与处理. 中国电子学会会员编号:E190159314M.

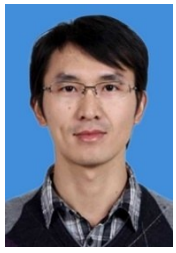
E-mail: ajianliu92@gmail.com

梁凤梅 女,1969年5月出生于山西省太原市. 现为太原理工大学副教授、硕士生导师. 主要研究方向为计算机视觉、智能信息处理. 获得省科技进步二等奖1项、省科技进步三等奖2项.

E-mail: fm_liang@163.com

张小梅 女,1995年2月出生于山东省菏泽市. 中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员. 主要研究方向为模式识别和计算机视觉.

E-mail: xiaomei.zhang@nlpria.ac.cn



万 军 男,1984年10月出生于江西省高安市. 中国科学院自动化研究所多模态人工智能系统全国重点实验室研究员. 主要研究方向为智能视频分析与交互技术. 中国电子学会会员编号:E190014368M.

E-mail: joewan10@gmail.com

谢 珺 女,1979年6月出生于山西省太原市. 现为太原理工大学副教授、硕士生导师. 主要研究方向为智能信号处理. 中国电子学会会员编号:E190091975M.

E-mail: xiejun@tyut.edu.cn

雷 震 男,1983年6月出生于浙江省嵊州市. IEEE Fellow,中国科学院自动化研究所多模态人工智能系统全国重点实验室研究员,中国科学院大学岗位教授,中国科学院香港创新院人工智能与机器人创新中心教授,博士生导师. 主要研究方向为计算机视觉、模式识别、人脸识别、目标检测与识别、智能视频分析、三维人脸/人体重建等.

E-mail: zhen.lei@ia.ac.cn