视觉提示学习综述

廖 宁[1)] 曹 敏[2)] 严骏驰[1)]

[1)](上海交通大学人工智能教育部重点实验室 上海 200240)

[2)](苏州大学计算机科学与技术学院 江苏 苏州 215021)

摘 要 近年来,随着提示学习方法在自然语言处理领域被提出,其日益受到研究人员广泛关注.它通过将各类下游任务重构成预训练任务的形式,以参数高效和数据高效的方式将大规模预训练模型应用在各类自然语言相关下游任务中.其中以GPT系列为代表的模型通过提示学习在对话生成和多模态图文理解等任务上取得了巨大的成功.然而,这类模型及方法还不能解决视觉中的稠密任务.受此启发,一些研究人员逐渐将提示学习广泛应用到视觉相关的各类任务当中,如图像识别、目标检测、图像分割、领域适应、持续学习等.由于目前还没有提示学习应用在视觉相关领域中的综述,本文将对视觉单模态领域以及视觉语言多模态领域的提示学习方法展开全面论述和分析.作为回顾,我们首先简要介绍自然语言处理领域的预训练模型,并对提示学习的基本概念、下游应用形式以及提示模板类型进行阐述和分类.其次,我们分别介绍视觉单模态领域以及视觉语言多模态领域里提示学习方法适配的预训练模型和任务.再次,我们分别介绍视觉单模态领域以及视觉语言多模态领域的提示学习方法.在自然语言处理领域,提示学习方法以继承预训练形式实现多任务统一为主要目的;与此不同,在视觉相关领域,提示学习方法侧重于面向特定下游任务进行设计.为此,我们将从方法设计上进行简单分类,然后从应用任务角度详细介绍视觉单模态提示学习和视觉语言多模态提示学习方法.最后,我们对比分析了自然语言处理领域和视觉相关领域提示学习研究的进展,并对未来研究路线给出了展望.

关键词 大规模预训练模型;自然语言处理;视觉单模态提示学习;视觉语言多模态提示学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.00790

Visual Prompt Learning: A Survey

LIAO Ning[1)] CAO Min[2)] YAN Jun-Chi[1)]

[1)](*Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai Jiao Tong University, Shanghai* 200240)

[2)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215021)

Abstract With the rapid development of deep learning models and the increasing parameter size, fine-tuning the entire model in various downstream applications with different objectives is prohibitive. To solve this significant issue, prompt learning has been primarily proposed in the field of natural language processing (NLP), and has been widely studied in recent years. By reformulating various downstream tasks as the same form of the pre-training one, prompt learning successfully leverages large-scale pre-trained language models in various downstream applications with great efficiency from both the parameter and data perspectives. Among them, models pre-trained by masked language modeling (MLM) represented by BERT have achieved great success in tasks requiring word-level output such as text classification, named entity recognition by "cloze prompt"; models pre-trained via autoregressive/casual language modeling (A/CLM) such

收稿日期:2023-05-30;在线发布日期:2024-01-15. 本课题得到国家自然科学基金优秀青年科学基金项目(No. 62222607)、上海市级科技重大专项(No. 2021SHZDZX0102)、国家自然科学基金(No. 62002252)资助.**廖 宁**,博士研究生,主要研究领域为多模态大模型、提示/指令学习、开放集识别. E-mail:liaoning@sjtu.edu.cn. **曹 敏**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为视觉语言学习、异常检测. **严骏驰(通信作者)**,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为机器学习及与组合优化、量子计算的交叉. E-mail:yanjunchi@sjtu.edu.cn

as GPT have been widely applied in tasks requiring text-level output using "prefix prompt", the tasks include dialogue generation, question answering, summarization, etc. Witnessing the success of prompt learning in NLP area, language models have also been applied in multimodal vision-language understanding problems through prompt learning. However, they still could not solve dense tasks in vision-related area. In addition, the expensive and complex process of fine-tuning the entire vision model in practical applications also occurs in vision-related area. Inspired by the great success of prompt learning in NLP, it has been gradually applied to various vision-related tasks, including image classification, object detection, image segmentation, domain adaptation, continual learning, etc. Seeing the lack of a comprehensive survey of prompt learning in vision area, therefore, this paper aims at conducting a comprehensive introduction and analysis on the prompt learning methods in unimodal vision area and multimodal vision-language area. First, we briefly introduce the pre-training models, the basic concepts of prompt learning, the forms of downstream applications, and the types of prompts in NLP as the preliminary. Second, we deliver the pre-training models that adopted in unimodal vision and multimodal vision-language prompt learning methods, respectively. Then, we give a comprehensive introduction to the prompt learning methods in vision-related areas. It is worth mentioning that prompt learning methods in NLP are designed for inheriting the pre-training tasks in all downstream applications. Differently, current prompt learning methods in unimodal vision and multimodal vision-language fields are designed for specific downstream applications. Therefore, we will conduct a brief introduction from the method design, and then give the details of unimodal visual prompt learning and multimodal vision-language prompt learning methods from the perspective of application tasks. On the one side, unimodal visual prompt learning methods are mainly designed by concatenating learnable prompt tokens, adding optimizable pixel-wise perturbations, learning prompt networks, combining multiple prompt modules, constructing the label mapping, neural architecture search, etc. On the other side, the popular design of multimodal vision-language prompt learning methods includes textual prompt learning, vision-guided textual prompt learning, text or knowledge-guided textual prompt learning, vision-language joint prompt learning, distribution-based prompt learning, multitask-shared prompt learning, gradient-guided prompt learning, etc. Finally, we make an in-depth analysis and comparison between the prompt learning methods in NLP and vision-related fields, and propose a prospect and summary for future research.

Keywords　large-scale pre-trained model; natural language processing; unimodal visual prompt learning; multimodal vision-language prompt learning

1　引　言

近年来,随着如 GPT[1],BERT[2],T5[3]等大规模预训练语言模型的相继提出,"预训练－微调"范式极大地推动了自然语言处理领域的发展.在这个范式中,首先对以 Transformer[4]为主干的模型在广泛无标注的语料数据集上通过语言建模[1-2,5]等任务进行自监督预训练,然后在下游应用中针对不同的任务设计不同的优化目标和添加新的网络模块,通过对模型和添加网络模块的全部参数进行微调来实现部署和应用.由于预训练数据体量大、模型参数多,预训练模型具有极强的文本综合表征和理解能力,使得"预训练－微调"范式在各类下游任务上都展现出了卓越的性能.然而,这种范式存在以下几个问题:(1)在不同任务上都需要优化和调整模型的全部参数,造成了巨大的计算开销以及部署成本的增加;(2)需要针对不同任务进行不同的优化目标设计,不可避免地造成了预训练与下游任务之间的差异,限制了对预训练知识的充分利用;(3)收集专属各类下游任务的训练集对模型进行微调成为此范式必不缺少的一个环节,不适用于数据资源匮乏的实际应用场景.

为此，LAMA[6]、GPT-3[7]等大规模语言模型相继被提出，这些大规模语言模型采用一种“预训练－提示－预测”的新范式，一定程度地解决了“预训练－微调”范式中存在的问题，再次推动了自然语言处理领域的发展，其中的提示学习也成为近几年的研究热点. 不同于“预训练－微调”范式需要将预训练模型通过不同的目标设计适配应用到各类下游任务中，“预训练－提示－预测”范式通过将下游任务重构成预训练任务的形式，使得各类下游任务能够以预训练预测的方式被解决，这些下游任务包括事实调查[6,8]、文本分类[9-10]、自然语言推理[11]、命名体识别[12]、常识推理[13-14]、问答[15]等. 例如，在根据影评“这部电影很好看”对电影情感类别判断的例子中，不需要专门收集下游数据和额外增加一个需训练优化的分类层到预训练模型，只需要将影评与提示模版“这部电影的类型是___”串接起来作为模型的输入，直接借助于预训练阶段的语言建模任务就可以在空白处预测出电影的类别. 总的来说，“预训练－提示－预测”范式展现出了以下优势：(1)预训练模型的全部参数都可以保持不变，极大降低了下游应用的计算和部署成本；(2)通过任务重构保证了下游任务与预训练任务的一致性，可以更加充分地利用预训练模型的知识；(3)额外收集下游训练集在这种范式下不是必要的，除了可以和微调的方式一样应用在数据充足的场景下，提示学习还可以在零样本或者少样本场景下使用.

在视觉单模态以及视觉语言多模态领域，“预训练－微调”范式被广泛采用[16-17]，也同样存在计算成本高、部署复杂等难题. 受提示学习高效利用大规模预训练语言模型的启发，很多学者将提示学习引入到视觉单模态和视觉语言多模态领域来解决各类相关下游任务.

目前的视觉单模态提示学习方法包括串接可优化向量序列[18-20]、添加像素级可优化扰动[21-23]、学习提示网络层[24-26]、面向特定成分的组合提示学习[27-28]、建立标签映射[29-31]、任务重构[31]、网络结构搜索[32]等. 这些方法适用的下游任务包括数据均衡视觉分类[18,21,24,31]、持续学习[19,33-34]、领域泛化、领域适应[20,28,35]、细粒度目标检索[36]、对抗鲁棒学习[23]、语义分割[37]、长尾识别[38]、开放集学习[39]等.

在视觉语言多模态领域，提示学习方法包括纯文本提示学习[40-42]、视觉信息引导的文本提示学习[43-44]、文本或外部知识引导的文本提示学习[45-46]、文本和视觉联合提示学习[47-48]、面向特定成分的组合提示学习[49-50]、基于分布的提示学习[51-52]、多任务共享的提示学习[53]、梯度引导的提示学习[54]、无监督提示学习[55]、建立颜色与标签关系[56]、视觉映射到语言空间[57]等. 这些视觉语言多模态提示学习方法被应用于各类下游任务，包括数据均衡视觉分类[40,43,47,51,53]、基础到新类别泛化[45-46,48,52]、领域泛化[40,43,48,58]、领域适应[59-60]、视觉问答[61-62]、图片描述[63-64]、图文检索[65]、视觉蕴含[61]、视觉推理[66]、多标签分类[67]、开放集识别[31,68]、去偏差提示学习[69-70]、组合零样本学习[71-72]、图像分割[73-74]等.

针对自然语言处理领域中的提示学习方法[6,75]已经有相关综述[76]展开了全面的介绍. 而在视觉相关领域中，目前只有针对预训练技术的综述[77-78]，而缺少视觉领域提示学习方法的综述. 为此，本文对单模态视觉以及多模态视觉语言领域中的提示学习方法展开全面介绍.

我们将首先介绍自然语言处理领域的预训练模型和提示学习方法[6,75]的基本概念，并且结合大规模预训练语言模型的预训练任务简要介绍提示学习的应用形式和模版类型. 其次，我们将分别介绍视觉单模态与视觉语言多模态领域里的预训练模型. 再次，我们将针对提示学习在各类下游任务上的广泛应用，分别详细介绍视觉单模态以及视觉语言多模态领域中针对各类应用任务提示学习方法的设计以及特点的分析. 最后，我们给出在未来研究中视觉和多模态提示学习方法发展的方向，并总结全文. 全文组织结构如图 1 所示.

2 大规模语言模型及提示学习

提示学习方法首先在自然语言处理领域被提出[6,75-77]，旨在实现大模型在多个下游任务中的高效应用，具体而言，其通过在模型的输入层或者中间层嵌入任务相关的信息，从而引导大模型解决多个任务. 通过继承预训练形式实现了保持下游任务与预训练一致性的应用. 我们在本节首先简要介绍自然语言处理领域的预训练模型，其次介绍基于预训练模型的提示学习方法，其中包括基本概念、下游应用形式和提示模版种类.

2.1 预训练模型

预训练模型是指在大规模数据上通过具体代理任务进行训练后得到的模型，通常具有较强的特征提取和数据理解能力，可广泛应用于各类下游任务当中. 在自然语言处理领域，目前有四种典型的预训练模型[76]，分别是：

(1)自回归语言模型：在预训练阶段，通过从左至

右逐一预测词的方式进行语言表征学习.具体来说,给定一个文本序列,模型需要在每个位置根据此前左侧的可见文本序列预测当前位置的词.这类工作的代表模型有 GPT-3[7]、RETRO[78]、GPT-Neo[79]等.

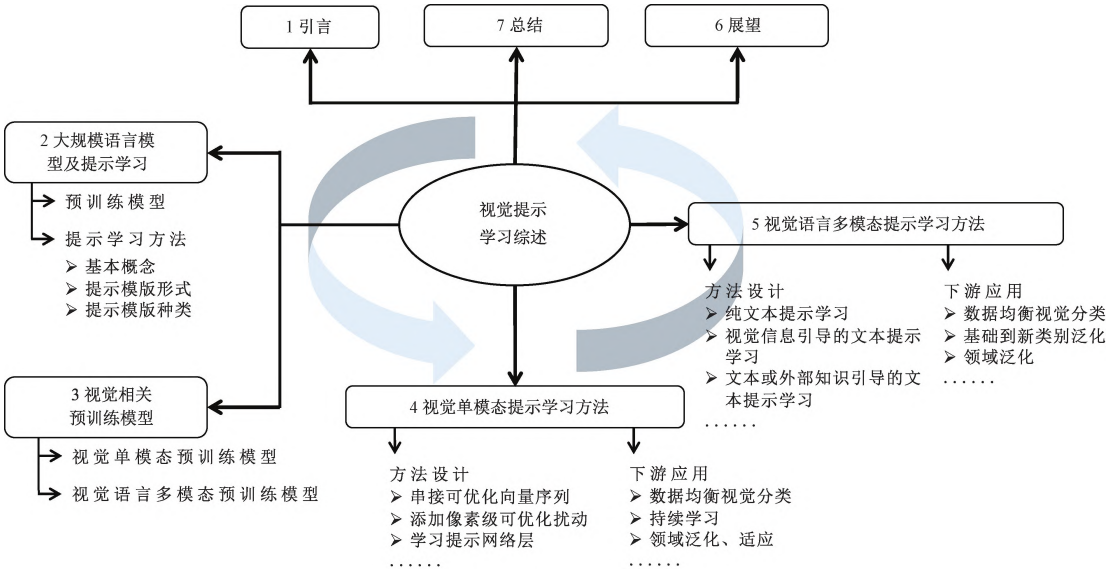


图 1　全文组织结构

(2)掩码语言模型:自回归语言模型只能从左至右对文本进行表征学习,限制了模型对文本双向建模的能力.为此,掩码语言模型在预训练阶段通过对文本中的词或片段进行随机掩码,根据前后文来预测掩码处的词,实现了对文本的双向理解表征.这类代表模型有 BERT[2]、ERNIE[80-81]等.

(3)前缀序列语言模型:面向如机器翻译或者文本总结等基于给定的条件文本进行新的文本生成任务,将给定文本作为前缀序列经过双向文本建模后,利用自回归的方式来预测后续的目标文本.这类工作的代表模型有 UniLM[82]、UniLMv2[83]、ERNIE-M[84]等.

(4)编码－解码语言模型:类似于前缀序列语言模型的处理,这类模型将给定的条件文本利用编码器进行双向文本建模,之后利用独立的解码器通过自回归的方式预测后续的文本.这类工作的代表模型有 MASS[85]、T5[3]、BART[5]等.

2.2　提示学习方法

2.2.1　基本概念

为了以参数高效、数据高效并且减少预训练与下游任务差距的方式来利用大规模语言模型进行下游任务,TemplateNER[12]和 KPT[86]等相继被提出,这些方法借助提示模版将各类下游任务重构成预训练任务的形式,之后通过预训练阶段的预测方式求解下游任务.对于文本分类[9-10]和命名体识别[12]等任务,提示学习将任务重构并进行预测的过程主要包括如图 2 所示的三个部分:

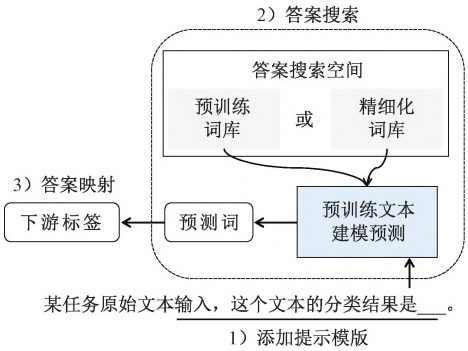


图 2　分类等任务提示学习流程

(1)添加提示模版:给定一个任务的文本输入,我们需要针对任务形式添加对应的提示模版.例如,针对分类任务,我们可以将模版设定为:“这个文本的分类结果是____”.之后将提示模版与待分类的文本串接起来作为掩码语言模型的输入.

(2)答案搜索:基于添加了提示模版的文本输入,大规模语言模型可以通过与预训练阶段的预测任务一致的方式预测空白处的词汇.例如,针对第一步的文本分类任务以及掩码语言模型,我们可以定义答案空间为预训练阶段的全体词汇库,或者是由先验知识引导确定的更加精细的小范围词汇库.通过掩码预测任务在词汇库中搜索掩码预测概率最高的词作为候选答案.

(3)答案映射:大规模语言模型预测出的结果有可能不会完全对应到下游任务的标签,因此提示学习需要设计答案映射这一步.例如,针对文本的分类任务而言,它的分类标签只有“正向情感”和“反向情

感”两类，我们需要将模版空白处预测的结果映射到这两类标签中. 如“好的”、“开心的”、“兴奋的”等词汇可以映射到“正向情感”标签；“坏的”、“消极的”、“难过的”等词汇可以映射到“负向情感”标签.

2.2.2 提示模版形式

对于不同的自然语言处理下游任务、不同类型的预训练模型，研究人员通常会选择不同形式的提示模版对任务进行重构，从而保证下游任务与预训练任务的一致性. 典型的提示模版形式有两种：

(1)完形填空模版：在一个文本中设置空白部分用于答案预测. 完形填空模版通常适配掩码语言模型，例如 LAMA[6]、Template[12]等大规模语言模型中使用这种完形填空模版作为提示模版，它的预测方式与预训练任务的方式完全一致. 从下游任务角度来看，文本分类[9-10]、自然语言推理[11]、命名体识别[12]、常识推理[13-14]、问答[15]等任务都可以通过输入文本串接完形填空模版后，以掩码预测的方式在空白处生成对应答案的相关词汇，之后通过答案映射获得最终结果，完成下游任务.

(2)前缀序列模版：将一个文本串接到原始输入上，用于引导之后的文本生成. 前缀序列模版适用于如 T5[3]等预训练模型，例如 Prefix-Tuning[87]、Prompt-Tuning[10]等方法中采用这种前缀序列模版，它的预测形式与自回归语言模型、前缀序列语言模型以及编码—解码语言模型的预训练任务高度一致. 作为一种可以根据前序文本生成新的文本的提示模版，前缀序列模版适用于文本生成[7,11]、信息提取[12]、文本生成评估[88]等任务.

2.2.3 提示模版种类

除了从形式上可以将提示模版划分为完形填空和前缀序列模版，我们还可以从提示模版本身参数是否会被优化角度对提示模版的种类进行分类：

(1)离散模版[89-90]的每个词都是有实际语义的，并且存在于预训练词库中的，这些词对应的词表征的模型参数在预训练阶段优化后被固定住，在下游应用中不会被进一步优化. 离散模版适用于零样本场景的任务中.

(2)连续模版[10,87]的每个词都是不具有实际语义的，并且不存在于预训练词库中的，这些词对应的词表征的模型参数在下游应用中可以针对特定任务和特定数据被优化. 这个优化过程称为提示调优. 在有对应下游数据可以用来辅助提示调优的场景下，连续模版能够展现出对特定任务和特定数据针对性的优势. 连续模版被之后的视觉以及多模态提示学习方法广泛使用.

3 视觉相关预训练模型

3.1 视觉单模态预训练模型

在视觉单模态领域里，对于大部分如图像识别等视觉任务，提示学习方法一般是基于判别式预训练模型进行设计的. 一类判别式预训练模型是基于大量有标签的数据如 ImageNet[91]利用交叉熵损失函数通过有监督分类对模型进行训练优化，这类模型包括 ResNet-18/50/101[92]、ResNet-28-2[93]、ResNeXt-101-32x8d[94]、ResNeXt(Instagram)[95]、Big Transfer (BiT-M)[96]、ViT[97]、RegNetX-32G[98]、DeiT-S/B[99]、Swin-S/B[100]. 另一类判别式预训练模型是 CLIP[101]的视觉编码器. CLIP 是基于包含大量图文对的数据集，通过对比交叉熵损失函数判断图文是否匹配来完成模型预训练. 此外，还有方法[31]受自然语言处理领域的方法启发，基于生成式预训练视觉模型采用提示学习来对任务进行重构. 代表的生成式预训练视觉模型有 BEiTv2[102]，其在预训练阶段对图片块随机掩码，通过预测掩码部分对应的视觉词来实现自监督预训练.

对于其他任务如图像修复或者图像生成[103]，对应的提示学习方法使用的预训练模型有 VQGAN[104]，BEiT[105]，MAE[106]. 对于图像合成任务[107]，预训练模型有 Taming Transformer[104]，MaskGIT[108].

3.2 视觉语言多模态预训练模型

在视觉语言多模态领域，最常用的预训练模型之一是在大规模图文对数据上通过对比学习训练得到的预训练模型 CLIP[101]. CLIP 是一个双塔结构，视觉编码器和文本编码器分别对输入的图片和文本进行特征编码和特征提取，特征之间的余弦相似度作为图文匹配度的衡量，匹配对的图文特征之间余弦相似度高，反之则低. 除了双塔结构的模型，多模态提示学习方法还会使用单塔结构的模型，代表模型有 ViLT[109]，其将文本和图片首先经过编码后统一输入到一个多模态编码器中得到多模态特征，通过图文匹配以及掩码语言建模任务实现预训练. 以上两个模型都是以判别式为主的方式进行预训练，它们只能学习到粗粒度的图文表征，而忽略了预训练任务与下游任务的适配. 相比判别式预训练模型，生成式预训练模型则更容易实现预训练和下游任务的一致性，典型的模型有 Wang 等人[110]提出的

OFA 模型.该模型将视觉定位、图片描述、图文匹配、视觉问答和目标检测等任务都重新建模成文本序列预测问题,实现了与预训练阶段自回归文本预测任务的统一.

除去以上利用视觉文本数据从头预训练的模型,多模态提示学习方法使用的模型还包括将单独训练好的视觉以及文本预训练模型进行组合的模式.其中包括 CLIP 的视觉编码器与语言模型 GPT-J[111] 的组合,视觉模型 ViT[97] 与语言模型 BERT[2] 的组合,视觉模型 NF-ResNet-50[112] 与语言模型 GPT-2[75] 的组合,BLIP[113] 与语言模型 OPT[114] 的组合.

4　视觉单模态提示学习方法

在自然语言处理领域,提示学习方法设计的目标是通过任务重构保持所有下游任务与预训练任务形式一致.而在视觉领域,提示学习方法设计的目的则是通过参数高效的方式利用预训练模型解决下游任务,因此是面向特定任务而进行设计.在本节,我们首先从方法设计角度对视觉单模态提示学习方法进行简单分类介绍,然后从下游应用角度对各方法进行详细介绍,本节框架如表 1 所示.

表 1　视觉单模态提示学习框架

方法设计	串接可优化向量序列	VPT[18], CSVPT[20], DoPrompt[28], DePT[35], LPT[38], Sohn 等人[107]
	添加像素级可优化扰动	VP[21], EVP[22], C-AVP[23], OpenPrompt[39]
	学习提示网络层	CSVPT[20], Pro-Tuning[24], PGN[25], LION[26], SPM[37]
	面向特定成分的组合提示学习	CSVPT[20], DAM-VP[27], LPT[38]
	建立标签映射	SEMAP[29], ILM-VP[30]
	任务重构	VPTM[31]
	网络结构搜索	NOAH[32]
	创建提示池与键值查询	L2P[19], DualPrompt[33], CODA-Prompt[34]
	上下文样例模版创建	Bar 等人[103], Zhang 等人[126]
下游应用	数据均衡视觉分类	VPT[18], VP[21], EVP[22], Pro-Tuning[24], PGN[25], LION[26], DAM-VP[27], SEMAP[29], ILM-VP[30], NOAH[32]
	持续学习	L2P[19], DualPrompt[33], CODA-Prompt[34]
	领域泛化、适应	CSVPT[20], DoPrompt[28], DePT[35]
	细粒度目标检索	FRPT[36]
	对抗鲁棒学习	C-AVP[23]
	语义分割	SPM[37], Bar 等人[103], Zhang 等人[126]
	长尾识别	LPT[38]
	开放集学习	OpenPrompt[39]
	图像合成	Sohn 等人[107]

4.1　方法设计

4.1.1　串接可优化向量序列

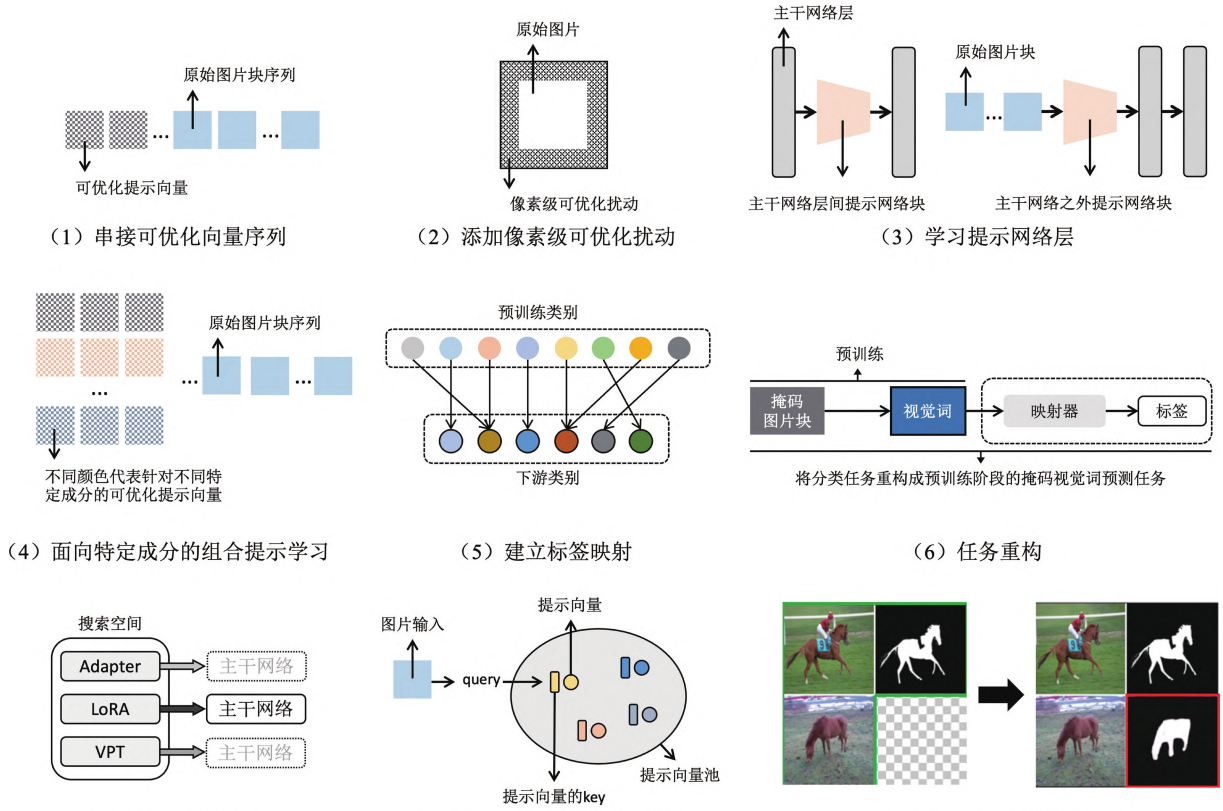
受自然语言领域处理 Prefix-Tuning[87] 等连续提示学习方法启发,如图 3(1)所示,这类方法[18,20,28,35,38,107] 基于 Transformer 结构,通过在原始输入序列或者 Transformer 结构的每一层特征序列上串接额外的可优化向量序列作为提示.然后在下游任务的微调阶段,主干模型冻结,只优化提示向量以及适配下游任务的新添加模块参数实现调优.

4.1.2　添加像素级可优化扰动

串接可优化向量序列的方法通常只适用于 Transformer 结构的预训练模型,不同于这类方法,添加像素级可优化扰动的提示学习方法可以适用于各类结构的视觉预训练模型.如图 3(2)所示,这些方法[20-22,30] 不依赖模型结构,直接在输入图像的像素空间上添加可优化的随机扰动块或者是矩形框与原图进行相加.在下游任务的微调阶段,通过优化扰动部分的参数完成提示调优.

4.1.3　学习提示网络层

这类方法是在主干模型上设计特定的附加网络作为提示模块,一般有两种类型,一类是如图 3(3)左侧所示的在主干网络层间添加的插入型提示模块,Pro-Tuning[24] 和 SPM[37] 在卷积网络或者 Transformer 的每一层之间插入阶段性的提示网络层,将上一层的特征输出作为该网络层的输入,经过提示网络变换得到的输出作为下一级的输入;LION[26] 在主干模型的输入端前与输出端后插入特定的网络层,将原图进行特征提取并串接到对应位置.另一类是如图 3(3)右侧所示的在主干网络之外的生成型提示模块,PGN[25] 设计提示网络层用于对图片特征生成提示序列,然后串接到原始图片序列上进行微调;CSVPT[20] 基于综合分类表征以及图片序列来设计提示生成网络,得到样本特定以及领域共享的提示序列,之后串接到原始图片序列上.



（1）串接可优化向量序列 （2）添加像素级可优化扰动 （3）学习提示网络层

（4）面向特定成分的组合提示学习 （5）建立标签映射 （6）任务重构

（7）网络结构搜索 （8）创建提示池与键值查询 （9）上下文样例模版创建

图 3 视觉单模态提示学习方法简图

4.1.4 面向特定成分的组合提示学习

当一个任务的所有类别或者领域数据都采用同样的提示模版时，模型存在表征能力差、泛化能力弱的问题. 对此，如图 3(4)所示的这类方法针对不同类别/域数据设计不同的提示模版. DAM-VP[27]将类别特征进行聚类得到多个聚类中心，对每个聚类组都设计对应的像素级提示，在测试阶段通过计算测试数据与各聚类中心的距离来选择最适合的提示模版. CSVPT[20]面向领域泛化任务设计了领域共享的提示向量以及样本特定的提示向量，解决提示学习在其他领域上泛化差的难题. LPT[38]在串接可优化向量序列的提示学习基础之上，面向长尾识别任务设计了类别共享的提示向量，以学习泛化性能优良的特征表示，并且基于类别之间的特征相似度对类别进行分组，设计组特定的提示模版，以保证方法的细粒度识别能力.

4.1.5 建立标签映射

视觉预训练模型一般都是基于大量有标签数据通过优化分类损失来指导模型预训练. 基于这类模型的提示学习方法在下游应用时除了需要优化关于提示部分的参数，还需要额外设计针对下游数据集的待优化分类头. 为了更大程度复用预训练模型而进一步提升参数高效性，如图 3(5)所示的这类方法在微调阶段复用预训练模型的分类头，设计从预训练类别到下游标签类别的映射规则来实现提示学习. ILM-VP[30]设计了一种迭代优化的方式来实现提示向量的优化. 在每一步训练结束后，对于某个下游类别的图像，其被预测次数最多的预训练类别则作为当前优化步的预训练到下游类别的映射，基于该映射关系进一步优化提示向量，如此交替迭代实现提示调优. SEMAP[29]提出了一种基于语义相似度的从下游标签到预训练类别的一对一以及一对多的映射规则.

4.1.6 任务重构

为了保证预训练任务与下游任务的一致性，如图 3(6)所示，这类方法通过将下游任务重构成预训练任务的形式实现提示学习. 具体的，基于掩码视觉词预训练模型 BEiTv2[102]，VPTM[31]将下游视觉分类任务改造成掩码视觉词预测的形式，首先预测出掩码位置的视觉词，之后基于原型映射到下游标签，从而完成图像识别.

4.1.7 网络结构搜索

如图 3(7)所示，这类方法将现有的参数有效(parameter-efficient)方法看作可拼接的模块，对不

同下游任务/数据集随机选择一种参数有效方法作为提示进行调优,然后选择性能最好的一个作为该任务/数据集上的最终提示.比如 NOAH[32]集成了参数有效方法 Adapter[115]、LoRA[116]、VPT[18],作为可拼接的提示模块,对每个下游任务以及数据集上通过随机选择以上一个子模块作为提示进行调优,选择性能最好的一个作为特定任务及数据上的最终提示.

4.1.8　创建提示池与键值查询匹配

为了设计能够适配下游任务数据的多样性的提示学习方法,如图 3(8)所示的这类方法[19,33-34]创建了提示向量池,并且对每个提示向量配置了对应的索引键(key).每个输入图片通过查询函数得到对应的 query,将 query 在提示向量池中查询对应的 key 来选择匹配该样本的提示向量,之后将选择的匹配提示向量串接到图片特征序列上.

4.1.9　上下文样例模版构建

针对语义分割、边缘检测、图片上色等结果需要通过可视化直观展现的任务,这类方法[101]参考自然语言处理领域的上下文学习(in-context learning[117-118]),如图 3(9)所示,将原图及任务对应的结果图作为样例提示,将待测试原图和用于输出预测结果的空白占位图片块,与样例串接作为一个完整的上下文学习模版输入到模型中.预训练模型根据样例提示输出待预测图片的结果.

4.2　下游应用

4.2.1　数据均衡视觉分类

针对数据均衡场景下的图片识别任务,VPT[18]基于有监督预训练的 ViT 模型,在输入图片块序列上串接可学习的提示向量,在综合分类表征上添加面向下游数据的分类头.通过冻结主干模型,仅优化提示向量参数以及分类头参数实现分类.Pro-Tuning[24]在主干模型的每一层插入提示网络,上一层的输出经过提示网络转换的输出与原始特征输出相加作为下一层的输入.通过优化提示网络以及面向下游数据的分类头参数实现分类.PGN[25]首先定义了一个向量词库来存储可学习的特征向量.PGN 基于输入图片生成多个由向量词库中特征向量组合得到的提示向量,并且串接到输入图片块序列上作为模型的最终输入.

VP[21]在输入图像的四周添加矩形框形式的像素级可优化扰动,通过固定的从有监督预训练类别到下游类别的有序映射来优化扰动参数实现提示调优.由于 VP 是直接在原始图片上通过相加的方式添加扰动,会丢失图片原始信息.EVP[22]基于 VP 首先将压缩后的原始图片进行数据增强,之后在图像外周添加可学习扰动,利用梯度正则化策略实现提示调用.这种数据增强的方式通过增加数据的多样性从而增强提示扰动的泛化性.NOAH[32]通过整合现有的参数高效方法到主干模型上形成超网,然后随机选择子模块在下游数据上进行调优,最后选择性能最好的子模块作为最终的提示模块.

由于图像数据集具有多样性,在一个数据集上学习得到的提示模版容易在其他数据集上呈现较差的性能表现.为了解决这个问题,基于 VP 的方法 DAM-VP[27]设计了一个基于聚类的提示选择策略和一个基于元学习的初始化策略.基于聚类的提示选择策略采用聚类算法将下游数据的特征聚类成多个子集,引导每个子集学习其对应的提示模版,在测试阶段,测试数据根据与各类中心的距离被拆分成多个对应的子集,从而使用对应的提示模版.基于元学习的初始化策略将通过聚类算法划分出的多个子集看作是多个独立的任务数据,通过在各个子集中采样一部分数据从而得到混合的小批次数据,然后利用该混合数据依次对每个子集的提示模版进行微调,上一个子集微调好的提示模版作为下一个子集对应提示模版的初始化,从而实现了不同子集之间信息的共享,并且可以高效地应用在新的数据集上.

为了进一步提升提示学习效率,LION[26]设计了轻量级的网络块用于生成针对每张图片的提示模版,并将提示模版分别添加在预训练模型的输入前端和输出后端.此外,为了保证训练的鲁棒性和稳定性,LION 借鉴 Lottery Ticket Hypothesis[119]只对一些重要的参数进行优化来避免过拟合.

以上方法都注重提示模版的设计而忽略了提示学习中的标签映射规则.为此,ILM-VP[30]基于 VP 设计了一种双边优化的从预训练到下游标签的最优映射.其在调优的每一步先由预训练模型做预测,训练集中每个类别的样本预测成预训练类别最多的那个,作为当前的映射,然后优化视觉提示参数.之后基于优化后的视觉提示进一步寻找每个训练类别预测最多的预训练类别,最终实现复用预训练分类头的最优标签映射.SEMAP[29]基于有监督分类预训练模型提出了两种基于语义关系的映射.第一种是一对一的映射规则 SEMAP-1,其利用 CLIP[101]的文本编码器计算预训练类别与下游类别的语义相似度,相似度最高的两个构成相互映射关系;第二种是基于 SEMAP-1 的多对一映射 SEMAP-A,除了考虑相似度最高的类别,SEMAP-A 还会对与最相似

预训练类别相似度的差值在设定阈值范围内的其他类别建立映射关系，把多个类别的概率相加得到最终下游类别上的概率.值得注意的是，这里的阈值并不是固定的，而是根据相似度排序从一个初始阈值依次乘上衰减系数进行更新.

4.2.2 持续学习

面临动态的学习场景，视觉模型需要在已经学习过之前类别的前提下去学习新的类别.大部分视觉在这种动态的实际场景下容易出现灾难性遗忘[120]的现象.因此，解决持续学习[121-122]成为视觉领域重要的下游任务之一.

现有的持续学习方法通常要求把之前学习过的数据存储下来，以便在将来的新任务上采样一部分来进行重复学习.这种存储数据的模式受到了数据私密性以及存储价格和空间的限制.为了摆脱这种限制，结合提示学习方法，L2P[19]首先创建了一种既有相似任务共享性又有任务独立性的提示向量池用来存储编码的知识，每个提示向量还设置了对应的键值(key).每个图片输入通过询问(query)函数得到专属该图片的查询值，利用该值可以从提示向量池中找到与其余弦相似度最高的键值(key)，从而选择对应的提示向量串接到原始图片输入序列上得到最终模型输入.在下游数据上通过目标函数优化更新提示向量参数.

由于 L2P 只学习了一个提示向量池，忽略了区分所有任务的共同特征与每个任务独有的特征.为此，DualPrompt[33]分别设计了两个互不联合的提示向量空间，分别是任务无关的提示(G -Prompt)和任务特定的提示(E-Prompt).基于 ViT 模型，两种提示所添加的网络层无重叠，每个层的输出会与提示向量串接起来经过转化函数得到下一层的输入.值得一提的是，任务无关的提示是可直接串接到特征上作为通用知识的引导，而任务特定的提示都设置了对应的可学习的键值(key)，需要将原始输入经过预训练模型提取的最后一层综合分类表征作为询问(query)并且计算余弦相似度来选择最终匹配各样本的提示.

DualPrompt 不管任务难易都只从固定任务的特定提示向量空间来选择提示，具有一定的局限性.为此，CODA-Prompt[34]设计了一种与任务数据复杂程度相关的提示，也就是一组提示成分，在下游任务上对这些成分进行加权求和得到专属不同数据上的提示.通过这种方式，新的任务也可以基于这些已有的提示成分生成新的提示向量，并且复用此前任务的知识.

4.2.3 领域泛化和适应

为了使得在源域上训练好的视觉模型使其能够在与源数据分布不一样的目标域上成功应用，领域泛化[123]和适应任务被提出.而由于训练和测试数据的分布差异太大，现有的方法难以学习到领域通用的或者学习领域共享与领域特定内容拆解开的方法，导致其不能够在各种基准上超越基线方法.

为了解决以上难题，结合提示学习，CSVPT[20]提出了领域共享的提示以及样本特定的提示来对预训练模型的输入进行修改.领域共享的提示用来学习任务的上下文背景，学习好后就被固定而不被优化，容易在源域上过拟合.为了进一步泛化到新的领域，该方法还设计了提示生成模块，利用预训练模型的综合分类表征以及图片块的嵌入表征来生成样本特定的提示，从而实现对每个数据分布信息的利用.领域共享的提示以及样本特定的提示相加后与原始图片序列串接在一起作为预训练模型的最终输入.DoPrompt[28]设计了 Domain Prompt Learning(DPL)和 Prompt Adapter Learning(PAL)模块.DPL 模块针对每个源域都设置对应的提示向量，然后串接到图片块序列后用来学习领域特定的知识，基于源域数据进行交叉熵损失优化来更新领域提示的参数.PAL 模块把多个源域的提示向量进行线性加权求和，从而生成针对每一个目标图片的自适应提示.这种方式不仅可以区分来自不同领域的特征，还可以产生有利于正确预测的提示，与目标域越接近的源域提示对应的权重会越大.DePT[35]基于有标签的源域数据来优化 VPT[18]形式的提示向量，之后在适应阶段进一步微调提示向量以及分类头的参数.其次，对于只给定未标记目标域数据的学习目标，DePT 通过创建的记忆库细化伪标记机制引导源域初始化模型.为了进一步缓解自训练过程中错误的积累，该方法还为提示学习设计了一个分层自监督正则化项，从而引导更强的表征学习.

4.2.4 细粒度目标检索

在视觉检索任务中通常需要在一批包含多个类别的图像中检索并返回与检索词同类别的图像，但是这对于在视觉上存在相似目标的检索具有一定的挑战性.为了解决此问题，细粒度目标检索任务[124-125]应运而生，它通过学习具有强判别性和泛化性的特征实现对相似的目标进行识别.

现有方法虽然能够学习到具有良好判别性的特征，但是其特征的泛化性较差，在上一个阶段学习好的模型需要在下一个阶段继续微调模型才可能得到

判别性强的特征.这种连续微调的方式不仅存在效率低下的问题,还会导致模型可能收敛到一个次优解,尤其是在数据有限的条件下,更为可能.因此,FRPT[36]提出 Discriminative Perturbation Prompt(DPP)模块对图片内容进行解析,经过非均匀抽样操作来放大有助于类别预测的内容;他们还提出了特征自适应预测头模块,通过类别引导的实例归一化去除主模型提取的物种差异,对特征进行优化,使优化后的特征只包含子类别之间的差异.两个模块都是直接插入到预训练模型的中间层来实现的,通过优化交叉熵损失来实现对模块的更新而保留预训练模型参数不变.

4.2.5　对抗鲁棒学习

VP[21]成功帮助改善了模型的泛化能力,但是只学习了统一的像素级的提示扰动,缺少了鲁棒的对抗样本攻击干扰的能力.基于此,Chen 等人[23]提出了 C-AVP 来生成类别特定的视觉提示.C-AVP 将每一类图像关联到一个对抗性视觉提示,并考虑这些视觉提示之间的耦合关系来增强鲁棒性,进一步扩展 VP 的设计空间.

4.2.6　语义分割

为了设计能够应用于各类预训练模型的结合提示学习的语义分割方法,Liu 等人[37]提出了 SPM 框架.他们将主干模型划分为多个阶段,基于下游数据集进行分阶段的提示调优.具体地,将划分后主干模型上一阶段输出的特征图以及语义图作为输入,从而来迭代学习合理的语义感知的视觉提示,之后在最后一层输出的特征图通过分割头来生成语义分割图.这种方式可以融合丰富的中间层语义图信息,以渐进循环的方式学习任意两个阶段之间有效的视觉提示.Zhang 等人[126]参考自然语言处理领域中的 in-context learning,针对语义分割任务,给定成对的图片以及分割结果作为提示样例形成上下文背景,将待预测图片输入预训练模型后得到对应的分割结果.由于良好的提示样例对于分割结果的作用很重要,作者提出了两种针对每个输入图片的检索提示样例的方法.第一种是无监督的提示检索方法,直接提取图片特征然后比较待预测图片特征和训练集中图片特征的余弦相似度,选择出相似度高的样例作为提示;另一种是有监督的提示检索方法,假设源数据包含标签,直接可以根据分割损失函数优化选择最合适的提示样例.Bar 等人[103]把每个任务定义成矩阵排列形式的图片组合,该组合的每行包含有特定任务的输入输出图片提示样例,最后一行则由待预测图片以及空白区域构成.作者从 arxiv 平台的关于计算机视觉文章里收集了 88k 关于各类视觉任务的图片,以及对应的任务结果,然后基于这些数据训练了一个大规模模型来根据任意网格的部分进行对应预测输出.

4.2.7　长尾识别

对类别样本数量分布不均匀的数据集进行训练,经常会导致模型对多样本类别过拟合、少样本类别被忽略的问题.为了解决这个问题,长尾识别任务被提出,通常有三个做法:(1)对长尾数据分布进行重采样实现数据平衡分布[127-128];(2)对训练损失函数进行权重设置,少样本的类别权重大,反之亦然[129-131];(3)特殊设计的解耦训练[127]、知识蒸馏[132]或集合学习[133].现有的方法通常需要从头训练或者微调模型,成本昂贵,并且在长尾数据上微调模型,会造成对某些特定类别数据的过拟合学习,从而损害模型的泛化能力.

结合提示学习,Dong 等人[38]提出了 LPT 方法来解决长尾识别问题,其中包含两个主要部分.第一部分是全类别共享的提示,用来学习类别共享的特征;另一部分是将类别进行分组后设置的组特定的提示,用来集合组内相似的特征来使得模型具有细粒度的分辨能力.LPT 的训练包括了两个阶段,第一阶段优化全类别共享的提示以及分类器,第二阶段优化新添加的组特定的提示并且在第一阶段的基础上进一步微调分类器.为了节省计算开销,全类别共享的提示只在 Transformer 的前面少数层使用,组特定的提示则在后面的剩余层使用.

4.2.8　开放集学习

大部分的视觉提示学习方法都是针对闭集场景的设计,不适用于在半监督开放场景.针对该场景,Li 等人[39]提出了 OpenPrompt 方法.为了在无监督的条件下检测出分布外样本,该方法将所有样本的表征投射到提示相关的联合特征空间上来扩大分布内样本和分布外样本的差距.此外,作者将检测出来的分布外的样本输入到预训练模型中学习分布外样本特定的提示,通过对比学习的方式将分布内数据和分布外数据的距离进一步拉远.

4.2.9　图像合成

为了生成与训练数据相似的且可信的图片,Sohn 等人[107]利用生成式视觉 Transformer 来通过迁移学习的方式进行图像合成.该方法包括两部分,第一个是提示生成器,可以通过类别或单个样本的条件变量来引导预训练模型到目标分布的迁移,实

现可控的图片合成.第二个部分通过组合和插值提示来加强生成图像的多样性.

5 视觉语言多模态提示学习方法

与视觉单模态提示学习方法类似,视觉语言多模态提示学习方法面向不同的下游应用进行了特定于任务的设计.本节我们将先从方法设计上对现有的视觉语言多模态提示学习方法进行分类,之后从各类下游应用任务角度详细介绍具体方法,本节框架如表 2 所示.

表 2 视觉语言多模态提示学习框架

方法设计	纯文本提示学习	CoOp[40], TaskRes[41], CoHOZ[68], TPT[134], DetPro[135], PROMPTDET[136], OrdinalCLIP[137]
	视觉信息引导单文本提示学习	CoCoOp[43], Img2Prompt[44], StyLIP[58], PL-UIC[64], DPL[138], MAPL[139], LVP-M3[140]
	文本或外部知识引导的文本提示学习	LASP[45], KgCoOp[46]
	文本和视觉联合提示学习	UPT[47], MaPLe[48], CAVPT[141], MetaPrompt[142], P3OVD[143], Yang 等人[144]
	面向特定成分的组合提示学习	R-Tuning[42], PTP[49], Lee 等人[50], DAPL[59], Wang 等人[63], CPL[65], Dual-CoOp[67], CSP[71], PromptCompVL[72], TaI[145], S-Prompts[153]
	基于分布的提示学习	ProDA[51], Derakhshani 等人[52],ZegOT[74], PLOT[146], PBPrompt[148]
	多任务共享的提示学习	MVLPT[53], SoftCPT[149]
	梯度引导的提示学习	ProGrad[54], SubPT[150]
	无监督提示学习	UPL[55]
	建立颜色与标签关系	CPT[56]
	视觉映射到语言空间	FROZEN[57]
下游应用	数据均衡视觉分类	CoOp[40], TaskRes[41], CoCoOp[43], UPT[47], PTP[49], Lee 等人[50], ProDA[51], Derakhshani 等人[52], MVLPT[53], ProGrad[54], UPL[55], CPL[65], CAVPT[141],PLOT[146], SoftCPT[149]
	基础到新类别泛化	CoCoOp[43], LASP[45], KgCoOp[46], MaPLe[48], Derakhshani 等人[52], ProGrad[54], PBPrompt[148]
	领域泛化	CoOp[40], TaskRes[41], CoCoOp[43], LASP[45], KgCoOp[46], MaPLe[48], Derakhshani 等人[52], ProGrad[54], StyLIP[58], PLOT[146], PBPrompt[148]
	领域适应	DAPL[59], PØDA[60], DPL[138], MetaPrompt[142], S-Prompts[153]
	视觉问答	Img2Prompt[44], FROZEN[57], UniVL[61], FEWVLM[62], CPL[65], MAPL[139], Yang 等人[144]
	图片描述	FROZEN[57], UniVL[61], FEWVLM[62], Wang 等人[63], PL-UIC[64], MAPL[139], Yang 等人[144]
	图文检索	UniVL[61], CPL[65]
	视觉蕴含	UniVL[61], Yang 等人[144]
	视觉推理	IPVR[66], TPT[134]
	多标签分类	DualCoOp[67], TaI[145]
	开放集识别	R-Tuning[42], CoHOZ[68], ZOC[162]
	去偏差提示学习	Chuang 等人[69], Berg 等人[70], SubPT[150], Menon 等人[163]
	组合零样本学习	CSP[71], PromptCompVL[72], ZPE[165]
	图像分割	CLIPSeg[73], ZegOT[74]
	目标检测	DetPro[135], PROMPTDET[136], P3OVD[143]
	多模态分类	Lee 等人[50]
	增强预训练	PTP[49], PEVL[171]
	视觉定位	CPT[56]
	多语言多模态机器翻译	LVP-M3[140]
	序数回归	OrdinalCLIP[137]
	图像编辑	Hertz 等人[173]
	生成模型分布控制	PromptGen[174]
	3D 识别	PointCLIPv2[177], Hegde 等人[178]
	视频相关任务	ALPRO[179], Re-Pro[180], Ju 等人[181], VoP[182], PZVMR[183], Yang 等人[184]

5.1　方法设计

5.1.1　纯文本提示学习

纯文本提示学习技术最先在如 CLIP[101] 这种双塔结构的多模态预训练模型上得以应用. 最开始的设计则是如“a photo of a _____.”的人工提示模版,通过在空白处添加类别词并将提示句输入到文本编码器中提取文本特征,同时将图片输入到视觉编码器中提取特征,之后便可以实现提示文本与图片的相似度计算,从而应用于零样本视觉分类任务. 由于人工模版通常需要大量的尝试,并且不能针对下游数据集进行特定的优化,受自然语言领域处理里的 Prefix-Tuning[87] 等连续提示学习方法启发,纯文本连续提示学习方法[40-41,68,134-137] 被提出. 如图 4(1)所示,这类方法将提示模版设置为一系列可以在连续空间进行优化的提示向量,在下游数据集上面向特定任务根据优化损失实现提示调优.

5.1.2　视觉信息引导的文本提示学习

在多模态场景下,只针对文本进行特定于任务的提示学习容易导致泛化性差、图文特征不对齐等问题. 为了解决这些问题,可以将视觉信息引入到文本空间作为文本提示学习的引导,如图 4(2)所示. CoCoOp[43]、DPL[138]、StyLIP[58] 设计网络学习特定于图片样本的表征,并且整合到纯文本的连续提示向量上实现灵活的、泛化性强的提示学习. MAPL[139]、PL-UIC[64]、LVP-M3[140] 将图片特征经过映射网络传递到文本空间辅助语言模型对视觉的理解. Img2Prompt[44] 利用现有的图片描述模型将针对图片样本生成的描述输入到语言模型中加强模态之间的理解.

5.1.3　文本或外部知识引导的文本提示学习

在纯文本的连续提示调优过程中,特定于下游数据集的提示向量容易产生过拟合的问题. 受人工提示文本具有强泛化性特点的启发,如图 4(3)所示,LASP[45] 和 KgCoOp[46] 在除了图文匹配的优化损失之外,还设计了对应的文本与文本或外部知识之间的相似度损失,使得提示向量与一系列人工提示文本或者外部知识保持一定的相似度,从而保留提示调优的泛化能力.

5.1.4　文本和视觉联合提示学习

以上的多模态提示学习都只限制于在文本上进行设计,由于文本特征和视觉信息内部的差异,这种单模态的方式限制了两个模态在下游任务上做灵活的适配,容易陷入次优解. 为此,许多方法提出在文本和视觉部分都进行提示学习,如图 4(4)所示. UPT[47] 设计了视觉和文本统一的提示向量. MaPLe[48]、CAVPT[141]、MetaPrompt[142]、P3OVD[143] 在文本和视觉分别设计了各自的提示向量,并且 MaPLe 将视觉信息通过耦合函数传递到文本空间进一步加强模态之间的交互. Yang 等人[144] 基于 OFA 模型设计了模态一致的提示向量串接到输入序列上. 这些双模态的提示向量在下游数据集上针对任务相关的损失函数进行提示调优,实现了模态联合的提示学习.

5.1.5　面向特定成分的组合提示学习

对于一个任务只设计一组提示可能会造成对类别数量多、视觉多样性高、领域来源广、特征属性丰富的数据表征能力不够的问题. 为了解决此问题,如图 4(5)所示,PTP[49] 对视觉特征相似的图片设置相同的提示,对视觉特征不同的图片设置不同的提示. Lee 等人[50] 面向模态可能缺失的真实场景,对模态齐全、只有图片、只有文本三种类型输入分别设置对应的提示. 在使用时根据模态的缺失选择对应的提示向量加入到模型中. 其中包括在输入层可以串接到输入序列上的提示向量以及在 transformer 注意力层中串接在 key 和 value 上的提示向量,通过控制 query 的长度不变而使得输出序列的长度不变. 在下游应用过程中通过优化提示向量参数以及综合分类表征 CLS 上外接的分类头的参数实现提示调优. DualCoOp[67] 将多分类问题建模成正负二分类问题,并对正与负特征分别设置对应的提示. TaI[145] 设置了全局和局部的提示分别学习整图与区域的信息. R-Tuning[42] 将大规模数据集按类别分组,对每个组分别设置不同的提示向量. Wang 等人[63] 为不同领域的数据设置不同的提示. 为了实现对多个任务的提示学习,CPL[65] 分别设置了任务通用的以及任务特定的提示. 针对领域适应任务,DAPL[59] 分别设置了领域通用、领域特定以及类别对应的提示. CSP[71]、PromptCompVL[72] 针对组合零样本学习任务对属性以及类别信息分别设置了不同的提示.

5.1.6　基于分布的提示学习

除了将特征相似的数据通过聚类后对不同类设置不同提示,还有一类方法将属性等特征通过建模成满足某种分布来实现提示学习,如图 4(6)所示. ProDA[51] 将图片以及可学习的提示文本分别输入到视觉以及文本编码器提取对应的特征,通过将两个模态对应的特征建模成高斯分布进行匹配实现视觉分类. Derakhshani 等人[52] 将图片特征建模成高

斯分布，并且从该分布中随机采样一个向量作为视觉信息的可泛化表征，并与纯文本提示向量相加从而保证提示向量具有视觉信息引导的多样性. PLOT[146]和 ZegOT[74]将视觉特征和提示对应的文本特征看成两个离散的分布，通过最优运输[147]来实现跨模态的匹配. Liu 等人[148]提出 PBPrompt，将每个标签建模成一个变分分布从而将不确定性引入到标签空间，之后从该分布中随机采样得到对应的提示. 类似地，PBPrompt 最后通过最优运输来实现视觉与文本分布的匹配.

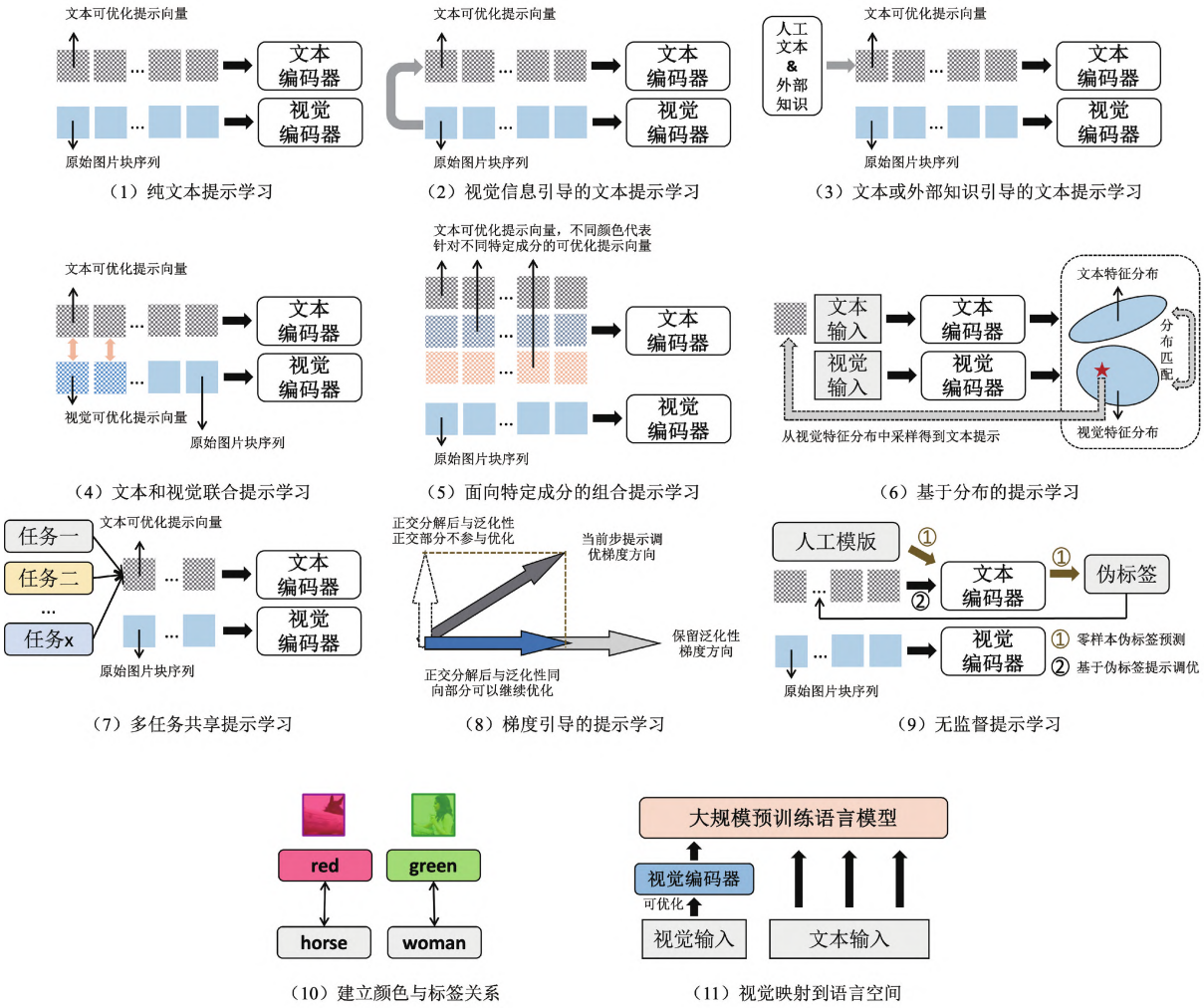


图 4 视觉语言多模态提示学习简图

5.1.7 多任务共享的提示学习

以上方法都是针对不同任务设计不同的提示，忽略了任务之间的相关性，限制了信息共享. 为了更好地利用不同任务之间的关联信息，如图 4(7)所示，Shen 等人[53]提出 MVLPT 方法将多个源任务合并联合优化一组提示，从而实现多任务共享，之后将共享提示作为目标任务提示的初始化. SoftCPT[149]提出了一个针对多个任务的元网络，在该网络中，每个任务的名称与任务元提示串接，对应的数据标签与标签元提示串接后输入到文本编码器后进行特征融合，即可得到最终的文本提示，之后在下游任务上进行提示调优.

5.1.8 梯度引导的提示学习

由于在连续空间进行调优会导致提示朝着下游数据产生过拟合现象，为了解决这个问题，对提示调优过程中的梯度变化进行分析并且实现梯度引导的提示学习成为一个解决方案，如图 4(8)所示. ProGrad[54]将调优过程中每一步的梯度方向正交分解为代表通用知识的方向以及其垂直方向，如果梯度方向与通用知识方向夹角为锐角，则在该步更新提示参数，否则不更新. Ma 等人[150]发现提示调优初始阶段能够保留泛化性，而靠后的阶段会逐渐损失泛化性. 为此，其提出 SubPT，定义靠前阶段的梯度方向为泛化性强的主特征方向，在靠后阶段将梯度

正交投影到主特征方向上进行提示调优.

5.1.9　无监督提示学习

现有的提示学习方法都依赖于下游有标签的数据进行提示调优,为了在无标签数据的场景下实现提示学习,如图 4(9)所示,Huang 等人[55]提出了无监督的提示学习方法 UPL.其利用人工提示模版对下游数据进行零样本预测,从而给无标签的数据打上伪标签,之后参考 CoOp[40]基于伪标签在连续空间上进行提示调优.

5.1.10　建立颜色与标签关系

在自然语言处理领域,大部分模型都通过掩码语言建模进行预训练。为了实现基于这种形式的跨模态提示学习,从掩码部分预测出视觉区域目标的类别,如图 4(10)所示,CPT[56]将图片中的目标按类别使用不同的颜色块覆盖,并且建立目标类别与颜色的映射.在提示学习中通过在提示语句设置的空白处预测出对应目标所覆盖的颜色,之后根据映射关系实现最终目标类别的预测.

5.1.11　视觉映射到语言空间

为了让大规模语言预训练模型能够理解视觉信息,如图 4(11)所示,Tsimpoukelli 等人[57]提出 FROZEN 模型.该模型将图片通过视觉编码器提取的特征映射到语言空间,形成视觉信息提示.在下游数据上进行优化的过程中,保持语言模型的参数不变,只有视觉编码器的参数需要从头训练.

5.2　下游应用

5.2.1　数据均衡视觉分类

在解决数据均衡视觉分类问题时,基于 CLIP[101]设计人工模版需要足够的专家知识并且会花费大量的时间不断尝试,为此,CoOp[40]在文本侧将提示设计为一组可在连续空间优化的向量参数.在下游提示微调过程中按照 CLIP 预训练的方式来进行对比学习从而优化提示向量的参数.为了增强 CoOp 的泛化能力,CoCoOp[43]设计了一个轻量级的元网络用来生成针对每一个输入特定的条件化的表征,之后与纯文本的连续提示向量相加得到专属该样本的提示,并参与优化和推理,实现更加灵活的提示学习.为了在纯文本提示优化上进一步保留预训练模型的知识,TaskRes[41]在预训练模型分类器从人工文本提取出来的类别表征上额外添加一个可学习的向量序列,面向下游任务进行优化,实现一种残差相加的提示调优方法.

类似地,为了在提示学习过程中保留预训练模型学习到的知识,ProGrad[54]将通用的预训练知识的方向定义为零样本识别和少样本微调模型预测之间的 KL 散度的梯度的方向,将数据标签和少样本微调模型预测之间的交叉熵的梯度的方向定义为领域特定的方向.领域特定的方向可以分解成与通用知识正交或者平行的方向,如果平行的方向同向,也就是领域特定知识方向与通用方向是锐角,则更新参数,否则不更新.如此就可以避免和通用知识的冲突,实现提示微调.

由于视觉特征以及文本特征的内在差异,以上只在文本部分进行的提示学习方法不能获得取得在多模态上一致的性能提升,为此 UPT[47]学习了一组统一的模态通用的提示.并且不同于直接将提示串接到输入上的方式,UPT 使用了轻量级的 Transformer 来将统一的提示进行转化,通过自注意力机制来帮助两个模态之间的交互从而最大化补足相互之间的影响.其输出被划分为两个部分,并分别作为视觉以及文本的提示串接到对应的输入序列或者是 Transformer 中间层.在提示调优过程中,提示向量以及轻量 Transformer 的参数都会被优化.除了在视觉和文本侧分别添加对应模态的提示,Xing 等人[141]还加强了对下游任务中类别的关注.其将视觉综合分类表征作为 key 和 value,类别词通过人工模版以及文本编码器提取出来的特征作为 query,通过人工模版以及视觉综合分类表征之间的交叉注意力机制生成 CAVPT,作为串接到输入序列的提示向量.另外为了确保 CAVPT 的性能,作者引入了 k-way 分类器,将映射到提示向量的线性层前的内容进行交叉熵监督分类.实验证明这种提示学习方法能够取得较好的分类精度.

Lu 等人[51]认为现有的多模态提示学习面临两方面的挑战:一方面需要针对不同的任务进行不同的设计,耗时费力;另一方面是视觉内容具有多样性特点.为了解决这两个挑战,Lu 等人提出了 ProDA,通过将提示经过文本编码器后输出端的特征进行高斯建模以进行有效的训练.该方法不仅能够从少量训练样本中学习到低偏差的提示,同时还学习了多样性提示的分布以应对多变的视觉信息.类似的,Derakhshani 等人[52]认为不考虑分布的提示学习容易造成泛化能力上的弊端,为此将提示分成两个部分,一部分是可直接设定的连续提示向量,另一部分是特定于输入样本的向量.给定一个图片,其经过视觉编码器提取的特征经过元网络学习得到高斯分布的均值和方差,从此分布中随机采样得到一个向量作为该图片的可泛化性表征,与直接设定的连

续提示向量进行相加，得到最终输入到文本编码器的提示. Chen 等人[146]认为每张图片都包含多种属性，据此提出 PLOT 方法，该方法设置了多个不同的提示分别关注不同的图片属性，将局部视觉特征和多个提示对应的文本特征分别看作两个离散分布的采样点，基于最优运输模型[147]来实现细粒度跨模态的匹配. 具体的，作者采用了交替优化的方式：第一步先固定模型和提示参数，通过 sinkhorn 算法[151]来优化最优运输问题的解；第二步固定最优运输模型的参数，通过反向梯度传播来优化提示向量的参数，实现不同的提示匹配不同的局部视觉特征. PTP[49]首先定义了 k 组原型，每组原型包含一个定义在视觉特征空间的图片原型以及提示原型. 图片原型定义在特征空间. 对于特征相似的图片，其共享对应的一组提示，而特征不同的图片应该使用不同的提示.

为了进一步实现多任务多数据集之间的提示信息共享，促进更快更高效的提示调优，MVLPT[53]设置了一组多任务共享的提示，首先在多个源任务上联合优化该组共享提示，第二步将共享提示作为目标任务提示的初始化，针对一组或者一个目标任务进一步优化提示向量的参数. SoftCPT[149]对不同任务以及类别都定义了对应的元提示，将任务名和类别名分别串接到对应的元提示上后通过文本编码器提取对应的任务特征和类别名特征. 提取到的任务以及类别特征输入到一个轻量级网络中进行串接或者平均操作来生成最终的任务相关的提示向量. 元提示以及轻量级网络可以在多个任务和数据集上复用，实现信息共享.

以上方法均针对图片和文本两个模态数据齐全的理想情况进行设计. 此外，为了适配于实际应用场景，Lee 等人[50]提出了一种具有模态感知的提示学习方法. 该方法分别对模态齐全、只有图片模态以及只有文本模态数据三类场景设置了对应的提示向量. 在下游提示调优过程中通过对输入层和注意力层的提示向量以及分类器的参数进行优化，实现了模态自适应的提示学习. 在下游实验中通过对数据进行预处理模拟模态缺失的实际场景进行验证，取得了超越基线方法的分类效果.

为了实现在无标签场景下的视觉分类，Huang 等人[47]提出了无监督的提示学习方法 UPT. 其首先用人工模版对训练集进行零样本预测，从而为每张图片打上伪标签，之后对每个类别选择得分较高的少数样本来构建训练集，然后基于此实现类似 CoOp[40]的提示调优.

5.2.2 基础到新类别泛化

为了解决在训练集之外的类别图片分类问题，增强提示学习的泛化性成为关键. 除了以上介绍的 CoCoOp[43]、ProGrad[54]以及 Derakhshani 等人[52]提出的方法外，MaPLe[48]在视觉和文本分支的前面几层添加可学习的提示向量，并且提出了可感知模态分支的提示调优方法，通过设计的视觉文本耦合函数将文本的提示映射到视觉的提示空间实现模态信息共享，从而促进两个模态之间的梯度协同.

考虑到人工模版比连续提示学习具有更强的泛化性，LASP[45]添加连续提示与人工提示文本之间的交叉熵作为新的损失函数，促使连续提示在特征空间与手工提示相近，从而保留了连续提示在预训练阶段学习掌握到的泛化性强的通用知识，并可以将其应用到下游任务中. 这种文本到文本的损失函数可以认为是一种正则项或者基于语言的数据增强，从而帮助学习更加具有区分性的类别中心. 与 LASP 类似，KgCoOp[46]在 CoOp[40]提示调优过程的基础上，添加了向人工提示特征拉近的损失函数，从而保留提示的泛化性.

此外，Liu 等人[148]基于贝叶斯框架提出了 PBPrompt，将每个标签建模成一个变分分布，从而将不确定性引入标签空间. 从该分布中随机采样得到属于对应的提示，以此避免单个提示难以描述属性复杂度高的类别的问题. 特别地，为了解决提示学习泛化性差的问题，作者将提示向量特征和图片特征看作是共享语义空间的两个分布，基于最优运输进行视觉和文本的匹配，实现最终的提示调优.

5.2.3 领域泛化

在实际应用中，训练集数据所属领域与测试集数据所属领域可能不一致，为了解决在这种领域差异场景下的视觉分类问题，如前介绍的提示学习方法都具有很强的领域泛化能力，其中包括 CoOp[40]、CoCoOp[43]、MaPLe[48]、KgCoOp[46]、TaskRes[41]、ProGrad[54]、PLOT[146]、PBPrompt[148]和 Derakhshani 等人[52]提出的变分提示学习方法.

此外，Bose 等人[58]提出了将内容与风格信息解耦合的提示学习方法 StyLIP. 其首先利用视觉编码器多层的特征来获取图片对应的内容以及风格信息，从而避免高层特征语义丰富但是视觉信息少和底层特征语义少但是视觉信息丰富的缺点. 在每一层输出特征上添加风格映射器提取领域信息并学习风格特定的提示向量作为文本编码器的输入. 另外，

还需要将降维处理后的视觉特征输入到内容映射器中提取内容特征,并与提示文本的特征进行融合,最后与图像特征进行相似度计算.在下游数据上通过优化风格映射器、内容映射器以及提示参数实现提示微调.

5.2.4　领域适应

基于大规模预训练模型和提示学习,Zhang 等人[138]提出了领域适应方法 DPL.其中设计的提示生成器将 CLIP[101]视觉编码器提取的图像特征作为全连接层的输入之后生成对应的提示向量序列.利用生成的提示向量序列进行 CoOp 形式的提示调优对提示生成器的参数进行优化.为了避免对目标域图片的依赖,Fahes 等人[60]提出了 PODA,将对目标域的描述作为提示,引导源域数据特征的一种仿射变换,使得源域数据的特征贴近目标域实现零样本的领域适应.具体做法是将源域特征进行 AdaIN[152]数据增强,通过优化图片特征的均值和向量使得图片特征靠近目标域描述提示对应的文本特征来实现适应.

由于现有的无监督领域适应方法通过对齐源域和目标域来学习领域通用特征的方式导致了语义特征的损失以及类别区分性的降低,Ge 等人[59]提出了结合提示学习的方法 DAPL.此方法的提示包括三个部分,分别是领域通用的提示、领域特定的提示以及类别标签.领域通用的提示用于表征通用的任务信息,可以在所有数据上共享通用;领域特定的提示代表各个域的特征,只在域内共享通用;标签信息用于区分不同类别.DAPL 通过对比学习的方式进行提示调优,一张图片只与在领域和类别信息都与其一致的文本构成正样本对,任何其他不满足这两个要求的文本都是负样本.

Zhao 等人[142]认为存在一个统一的元领域可以包含物体的所有属性,而下游数据的领域均是从该元领域中采样出来的属性分布.基于此假设,作者提出了 MetaPrompt,分别在文本以及图片侧都添加可学习的提示向量.为了避免提示对训练集过拟合,作者设计了一种非对称的对比损失函数,分别对文本和图像侧的提示向量进行单独优化而不是联合优化.并且对文本侧进行提示调优时,视觉图片不添加提示向量;对视觉侧进行提示调优时,文本不添加提示向量,只包含类别标签.之后把两个独立的对比损失函数相加进行优化.另外,为了学习到正则化的提示,作者将每一个数据批次拆分为领域不同的两个部分,用一部分训练优化,另一部分用来正则化调优的提示参数.

在领域增量场景中,不同领域的知识很难在同一模型中体现.此外,保存旧任务的数据会占用大量存储空间,并且存在隐私问题和新旧数据量不平衡的问题.为此,Wang 等人[153]提出了 S-Prompts,逐个对每个领域学习对应的提示.在增量训练时,预训练模型始终是固定的,通过提示调优可以将预训练模型调整迁移到不同的领域中.在这样设定下,不同领域的知识被编码到仅有少量参数的提示向量中,不仅避免了存储旧样本占用大量空间的问题,同时还可以极大地减少灾难性遗忘.测试时为了确定样本应该使用哪个领域的提示,作者针对预训练模型对每个领域的训练数据提取的特征进行 K-Means 聚类得到特征中心.在推理时,直接使用 K-NN 查询测试样本应该使用哪个领域的提示.由于领域增量任务的特征往往差别很大,这种简单的做法可以在领域增量学习中获得良好的性能.

5.2.5　视觉问答

结合大规模多模态预训练模型,许多针对视觉回答任务的提示学习方法被提出.UniVL[61]结合提示学习进行视觉语言预训练,该模型包括独立的视觉编码器和文本编码器,以及将两个独立模态编码器提取的特征进行融合的多模编码器.预训练阶段结合图文对比损失、掩码语言建模损失、图文匹配损失进行训练学习.由于双向注意力掩码有助于判别性任务,单向注意力掩码有助于生成任务,为了保证模型具有判别和生成的综合能力,作者将两种掩码以一定比例混合在预训练阶段进行使用.此外,作者还直接在预训练阶段就设置了提示向量基于训练数据在连续空间进行优化.在下游应用时,所有下游任务首先都被重构成文本生成的预训练形式的任务,之后将人工模版的提示替换成在预训练阶段优化的提示向量.

为了实现低消耗的下游应用,Jin 等人[62]提出 FEWVLM.他们首先使用在 Visual Genome[154]上训练好的 Faster R-CNN[155]检测得到输入图片的 36 个目标区域,这些区域的特征与文本特征一起输入到编码器中,然后采用前缀序列语言建模以及掩码语言建模任务进行预训练.在下游应用中,作者将输入文本添加到人工模版以及随机噪声组合形成的提示模版中,根据预训练的目标函数来进一步优化模型参数,作者发现提示模版对于零样本预测性能的影响很大,但是对少样本预测的影响相对较小,在给定更多数据的条件下,常规的

人工模版与噪声模版对于模型的学习速度影响一样. Yang 等人[144]继承 OFA[110]，通过将下游视觉问答任务重构成生成式文本预测任务，从而实现在连续空间的提示调优.

为了将视觉信息投影到文本空间，辅助语言模型对视觉的理解，Tsimpoukelli 等人[57]提出了 FROZEN，其依赖预训练好的语言模型作为文本编码器，在保证文本编码器参数不变的前提下，使用图片描述数据集从头训练视觉编码器，将每一张图片通过视觉编码器提取得到的成序列的连续表征输入到语言模型中，从而优化视觉编码器参数. 类似的，Mañas 等人[139]提出 MAPL，通过将视觉特征映射到语言模型的特征空间来辅助语言模型理解视觉信息，进行下游生成任务. 特别地，在视觉问答任务中将视觉特征与文本模版“Please answer the question. Question:{question} Answer:”串接后输入到语言模型生成答案.

从减少模态之间差异的角度出发，Guo 等人[44]提出了 Img2Prompt 方法，将图片输入到现有的图片描述模型中来生成描述文本，描述文本可以作为对图片的描述并将其输入到语言模型中从而减少图片和文本之间的模态差异. 基于从描述文本中提取的可能的答案词汇可以生成对应的问题得到问答对，作为大规模语言模型上下文学习的引导. 之后还针对问题相关的图片区域生成对应的描述. 最后将人工指令，描述以及问答对串接起来，并在末端加上当前视觉问答任务的问题，得到输入到大规模语言模型中的提示，通过自回归预测的方式输出视觉问答任务的结果.

由于现有的提示方法容易学习到虚假的表征，造成了在未见过的数据上呈现了较差的表征编码能力，即模型的泛化性较差. 为了解决这个问题，He 等人[65]提出了 CPL，将提示学习分为任务通用和任务相关两部分. 任务通用提示是端到端进行学习优化，任务相关提示是通过标签空间的映射来建立的. 视觉问答的提示则是通过 T5[3]语言模型来生成的. 为了生成视觉上的反事实样本而又避免视觉信息的冗余，作者提出基于文本的负采样策略，从一个批次数据里发现语义相似度最高的文本，之后通过 CLIP 的视觉编码器来获得对应的反事实负样本图片，紧接着在得到的图片中通过识别语义相似的正样本和负样本之间的最小非虚假特征变化来生成反事实样本. 基于生成的反事实样本和已有的事实样本，作者通过对比学习实现提示调优.

5.2.6 图片描述

与视觉问答类似，图片描述问题需要语言模型能够综合理解图片内容，并且输出对应的描述. MAPL[139]、UniVL[61]、FEWVLM[62]以及 Yang 等人[144]基于 OFA 提出的方法都可以很好地解决图片描述问题，Wang 等人[63]认为这些方法缺乏了对描述生成的可控性. 一个模型一旦被训练好后，其生成描述的风格就固定了并且难以被修改. 为了实现可控风格的描述生成，作者采用基于单向回归语言模型，并采用了带噪声的图文对数据进行预训练，通过语言建模，图文对比和匹配任务联合优化视觉编码器以及跨模态融合模型. 其中图文对比是通过简单的点乘两个模态特征来实现，图文匹配则是通过交叉注意力机制实现. 之后设计了多个提示作为锚点来区分来自不同领域的训练数据，在不同领域的数据集上分别优化对应的提示. 推理的时候用不同的提示就可以生成不同风格类型的描述文本.

Zhu 等人[64]认为基于对抗学习的图片描述方法[156-157]虽然能够构建图片和文本之间的关系，但由于这种关系是从零训练的，通常在泛化强的场景下性能较差. 为此作者提出了结合 CLIP[101]的提示学习的方法 PL-UIC[64]，整体上继承了以往基于对抗学习的模型框架，另外还设计了两个模块从 CLIP 中挖掘图文关联知识辅助图片描述任务：(1)设计线性层作为提示生成器用于将图片特征通过 CLIP 转换到文本空间中得到提示；(2)图片描述精细化模块利用 CLIP 对伪标注数据进行过滤，根据 CLIP 的输出结果去除图文相关性不高的图文对，从而使用较高质量的图文数据进行有监督训练. 进行一轮训练后，模型的能力得到提升，能够生成更高质量的伪数据，并可以继续使用 CLIP 过滤数据.

5.2.7 图文检索

图文检索任务是视觉语言多模态中的重要任务之一，可以分为两个子任务：图对文检索以及文对图检索. 具体来说，该任务需要根据给定的文本描述(或图片)从一组待选图片(或文本)中检索出对应的匹配内容. 基于 UniVL[61]和 CPL[65]的综合图文理解能力，这两个方法在图文检索任务中也展现出了出色的性能. UniVL[61]通过计算图文对比损失和图文匹配损失来评估图片和文本之间的相似度，在推理阶段基于图文相似度选择靠前的 k 组图文对作为最终检索结果. CPL[65]方法针对多个图文多模态任务首先学习了任务通用的提示，之后针对图文检索任务将图片的描述作为提示文本来学习任务专有的

提示向量.基于任务共享和任务专有的联合提示学习机制,该方法在图文检索任务上取得了超越 CLIP 方法零样本预测的性能.

5.2.8　视觉蕴含(Visual Entailment)

在视觉蕴含任务中,模型需要预测图像在语义上是否包含文本,这同样要求设计的方法具有综合的图文理解能力.以上介绍的 UniVL[61]和 Yang 等人[144]基于 OFA 提出的方法除了能够解决视觉问答和图片描述这两个问题,还可以在视觉蕴含任务上展现出优秀的性能.

5.2.9　视觉推理(Visual Reasoning)

视觉推理[66]旨在基于图文相关的目标问题,驱使模型首先识别图像内容,并且通过索引查询开放世界的知识从而执行逻辑推理,最终得出答案.Chen 等人[66]模仿人类的思考过程提出了 IPVR 方法,将整个视觉推理过程拆解为三个部分.第一个是 See 部分,它利用现有的场景解析器[158]检测出图片中所有的目标,并且用其对应的类别名代表这些目标,同时还生成图片的整体表征;第二个是 Think 部分,它利用图片描述模型[113]对检测出的所有区域中与问题相关的目标生成对应描述,并且利用大规模语言模型[114]将对应的描述和问题作为提示输入进行生成式的预测得到答案;第三个是 Confirm 部分,用于判定生成的答案与描述和问题以及图片输入是否具有合理性和连贯性.将生成的答案、描述、问题以及逻辑反复作为语言模型输入后,若得到的结果一致则认为满足合理性.为了实现在零样本场景下的视觉推理,Shu 等人[134]提出了 TPT 方法,判断询问图片(query)中是否包含支持图像集(support)中的概念.若存在,直接在测试样本上学习最优的类别词并且微调提示参数.

5.2.10　多标签分类

大部分提示学习方法都是在整个图片上做单标签分类,而图像中通常会包含多个物体,为了全面对图片中的目标进行识别,多标签分类任务被提出.Sun 等人[67]认为现有的多分类任务利用不完整的标签来实现文本与图像空间的对齐会造成一定程度上的精度损失,为此其提出了 DualCoOp,利用预训练视觉语言模型的图文对齐能力来进行多分类.对每个类别分别设置了一组正向和反向分类提示,在提示微调或者推理阶段,每个类别都需要填入到这两种提示中进行二分类相似度计算.另外,为了避免通过池化聚合得到全图特征的操作带来的局部目标信息的丢失,作者针对每个图片区域都进行一次图文特征匹配计算,根据正负提示计算得到的数值进行 softmax 计算实现加权平均得到全图特征,最后使用非对称损失函数[159]优化提示参数来处理多分类任务中标签的不均衡问题.

Guo 等人[145]认为在实际场景中经常会遇见图片数据不充分的情况,而对比式的视觉语言预训练模型对于图片和文本提取的特征应该相似度很高.基于此,作者提出通过将易获得的文本作为图片的替代进行提示学习的方法 TaI.作者认为全局的提示适合进行整图的分类,但是不适用于多分类场景,容易造成对图片中除主类别之外其他类别的忽略.为此,TaI 设计了两个粒度的提示:全局提示用来学习全局信息,局部提示用来学习局部信息.其提出的名词过滤策略用来从每个给定的用于替代图片的文本描述中提取与目标相似度最高的分类的伪标签,并且滤除不相关的其他名词.在提示调优阶段将数据集的类别填入到全局和局部提示后经过文本编码器获取提示特征,将文本描述输入到文本编码器获取文本特征,基于伪标签进行参数优化.在测试阶段,文本描述以及对应的文本编码器替换成图片以及图片编码器,全局和局部的提示不变,即可实现对图片进行全局以及局部的多分类.

5.2.11　开放集识别(Open-set Recognition)

为了解决在有限类别数据上训练的模型能够在开放应用场景下准确识别已知类和未知类的问题,开放集识别任务[160-161]被提出.Esmaeilpour 等人[162]提出了 ZOC 方法,利用图片描述数据集在 CLIP 上额外训练了文本解码器用于从图片中生成对应的类别描述,然后在测试时通过将已知的类别和生成的类别串接到人工模版中进行零样本预测,实现已知类和未知类的识别.Liao 等人[68]认为 CoOp[40]形式的连续提示容易在已知类别上过拟合,不论是在提示调优阶段还是测试阶段,都只能从已知类别中选择一类作为预测.为此作者提出 CoHOZ 方法,引入除已知类别之外的词汇串接到连续提示中从而消除过拟合现象.此外,为了解决开放集识别中所有未知类都被统一识别成一个类别(也就是,未知类别)而忽略了其语义差别的问题,作者还设计了下游数据集对齐的语义层次树,通过从粗粒度到细粒度的逐层零样本预测实现了带有语义信息的开放集识别.基于 CoHOZ,R-Tuning[42]为了将方法扩展到大规模数据集上,提出了 CTT 策略,把在大规模数据集上的提示调优分解成在多个类别组上的独立提示调优,并且在预测阶段

通过综合多个提示的预测结果选择最优的子提示作为最终预测，实现了在小、中、大三个规模数据集上最好的开放集识别效果.

5.2.12 去偏差提示学习

随着大规模预训练模型的广泛应用以及取得的卓越性能，使用视觉语言预训练模型的安全性也逐渐引起研究人员的关注，其中很关键的一个部分就是消除预训练模型在下游应用中的预测偏差. Berg等人[70]首先提出了对偏差的评估方式，分别从Word Embedding Association Test（WEAT）引导的排序、信息检索引导的排序以及误分类三个角度进行了评估，并且提出了去偏差的方法，其中包括两个部分：一个是最小化偏差的目标函数，一个是优化目标函数的参数选择. 作者定义了一个对抗的去偏目标函数，在只给定图片与敏感文本相似度值的前提下预测属性. 提示向量则被串接在敏感句之前并且通过最大化对抗误差来进行优化. Menon 等人[163]认为虽然引入语言信息在很多下游任务上可以取得不错的效果，但是视觉表征本身存在任务层面的偏差，不同的图片对于不同的任务有不同的偏向性. 为此，对于一张图片以及其对应任务的标签集合，作者将这些标签输入到视觉语言模型的文本编码器中获取对应文本特征，另外对图片添加提示后经过图片编码器提取对应视觉特征，目标任务作为真值，通过交叉熵优化实现提示向量参数的优化.

Chuang 等人[69]认为不管是视觉语言模型还是生成模型如 StableDiffusion[164]，这些预训练模型可能由于预训练数据包含的固有内容会造成一定的偏差并且传递到下游任务应用当中. 现有的去偏差工作都需要额外采样的数据或者修改的目标来训练或者微调模型，这对于大模型来说不太现实. 为此作者提出了不需要额外数据的通过对输出的方向进行重映射实现去偏差. 作者首先在特征空间通过有偏差的提示来定义一些有偏差的方向，之后根据这些提示特征的正交方向来定义无偏差的映射矩阵. 然而仅仅靠有偏差方向形成的映射矩阵进行去偏差会造成不稳定以及噪声多的问题. 于是作者提出了校准损失函数来最小化一对提示的特征之间的差异，也就是去除两个提示之间有偏差的内容后最大化他们的相似度，这使得映射矩阵的定义不需要训练下游的数据或者标签. 作者发现只在文本特征空间进行基于校准映射矩阵的操作就足以实现去偏差并且改善零样本预测的性能.

Ma 等人[150]首先从梯度的角度探索了目前提示学习方法如 CoOp[40]产生过拟合现象的原因，发现在提示调优靠前的阶段特征泛化性强、不易过拟合，而提示调优靠后阶段特征虚假性高、易产生过拟合现象. 为了解决靠后阶段的过拟合现象，作者提出了 SubPT 方法，首先对 CoOp 进行了少数量周期的训练，并且利用靠前阶段保存的权重来计算主特征向量，这个主特征向量被认为是泛化性较强的方向；之后从相同的初始点重新优化 CoOp，并且在整个提示调优过程中将梯度投影到预先计算的主特征向量的空间来优化，梯度中导致过拟合部分的影响通过正交投影被消除，从而实现了去偏差.

5.2.13 组合零样本学习

Nayak 等人[71]认为现有的组合零样本学习方法依赖于独立训练的图文特征对齐来实现，缺少灵活性，并且这种方式限制了属性的组成数量，不能灵活控制组成的模式，此外 CLIP 模型在没有微调的前提下效果也不如现有的方法，并且在预训练阶段没有形成对物体属性的监督. 为此，作者提出 CSP 方法将属性以及目标类别都定义成可学习的词向量，并且在多个提示组合上对其进行微调. 通过这种方式就可以在训练集上学习到关于属性和目标类别对应的提示. 属性和目标类别的向量都是根据 CLIP 模型的预训练权重初始化的. 在训练阶段同样采用 CLIP 的对比学习预训练方式来微调属性以及目标类别对应的向量参数；在推理阶段，使用提示以及微调好的属性和类别向量组合作为文本，与测试图片进行相似度计算. Xu 等人[72]认为组合零样本学习的主要挑战是训练数据和测试数据之间的分布转移带来的模型过拟合现象. 因此作者提出 Prompt-CompVL 方法，主要包括两个部分：第一部分是可在连续空间连续优化的提示向量，第二部分是可连续优化的词向量，即把各个对应的属性和目标类别名字对应的词嵌入向量，由冻结的参数变成可学习的参数. 如此可以保证属性和目标类别对应的名词向量既保留预训练知识，又可以在下游任务上根据见到的数据进一步更新.

Allingham 等人[165]认为设计适合不同数据集零样本预测的最优人工提示组合通常会比较困难并且会花费很多人力成本，而一般的提示设计方法需要借助有标签的验证集，这在实际应用中是不现实的. 因此作者考虑能否将零样本预测的手工提示设计过程自动化，也就是给定一组可能用到的提示文本，如何在不接触有标签验证集的前提下选择最优的提示文本子集，最大化模型的零样本预测能力. 作

者发现提示学习展现出偏差的原因有两个:(1)预训练数据中存在的词频偏差;(2)测试数据中存在的伪概念频率偏差.基于此,作者提出了 ZPE 方法,将提示预测出来的概率分数进行相减正则化来消除词频偏差.具体地,作者利用 LAION400M 数据[166]模拟预训练数据,将所有数据在提示文本上的预测做一个概率期望值计算,这个期望值作为原始预测概率分数的相减项.对于测试数据中的伪概念频率偏差,则用提示对所有测试数据进行预测并计算概率期望值,作为消除概念频率偏差的减项.将原始预测概率减去这两个期望值即可得到去偏差的预测分布.另外,作者还发现提示产生的预测也会出现长尾现象,小部分的提示产生的分数较高,大部分提示没有比较高的预测分数.为此,作者提出对所有的提示分数进行 softmax 归一化处理,然后通过阈值比较选择合适的一部分提示进行最终的零样本预测任务.

5.2.14　图像分割

针对图像分割任务,Lüddecke 等人[73]提出了 CLIPSeg.此方法亮点在于分割目标可以任意由文本或者图像来指定.对于文本指定目标,该方法可以泛化到新的未见类别上;对于图像指定的目标,该方法探索了各种形式的视觉提示方案.其设计了轻量级的解码器,并且在解码器和 CLIP 的编码器的某几层之间设计了残差连接来实现信息的传递.通过在训练集上优化解码器的参数生成二值分割图.Kim 等人[74]认为现有分割方法通常需要添加额外的编码器生成区域 proposal 或者外部知识,为了避免这种通过添加模块实现分割的高代价方式,作者提出了 ZegOT,允许每个提示文本关注不同的语义特征,产生最佳对齐的像素－文本得分图.通过最优运输对齐的分数图可以允许不同类别分布之间的领域偏移,从而在未见类别上取得鲁棒的性能.

5.2.15　目标检测

Du 等人[135]针对目标检测任务提出 DetPro 方法,将 RPN 网络提取的 proposal 作为 CLIP 模型视觉编码器的输入从而学习文本提示.对于属于背景类别目标的检测,作者认为 proposal 上提取的视觉特征应该与任何一个前景类别目标的文本特征都不相似.由于前景类别数量多,作者将背景类的 proposal 与前景类文本特征的相似度优化目标定义为一个较小的数值,也就是前景类别数量分之一,与所有类别保持等距离且较远的关系.由于不同的 proposl 与前景的交叠比例不同,如果都使用一组连续提示向量来做优化可能导致相互冲突的效果,因此作者对 IoU 进行等间隔均匀划分,在每个划分区间内的 proposal 使用一组公用的连续提示.

5.2.16　多模态分类

现有的大部分视觉语言模型都假设在应用场景中视觉和文本两个模态的数据都是完整的.而在实际场景中经常会面临模态缺失的情况,也就是缺失文本或者图像.为了解决在模态缺失场景下的多模态分类问题,Lee 等人[50]基于 ViLT[109]预训练模型提出了 missing-awareprompts.对于输入缺失的模态,作者首先利用空字符或者空图片进行补齐使得模型的输入具有模态完整性.根据模态的缺失情况设计不同的提示,并且加入到模型中.提示包括两个部分,第一个是输入层的提示,可以串接到输入序列上;第二个是 Transformer 注意力层中串接在 key 和 value 上的提示,这样可以由于 query 的长度不变而保证输出序列的长度不变.在提示调优过程中只优化提示向量的参数以及预训练模型在文本综合分类表征上的池化层以及全连接层的参数.此方法在多模态分类数据集 MM-IDMb[167],UPMC Food-101[168]以及 Hateful Memes[169]上都取得了良好的性能表现.

5.2.17　增强预训练

提示学习不仅可以应用在下游任务上,还可以作为预训练中的增强手段.Wang 等人[170]认为现有的视觉语言模型缺少视觉定位的能力,为此其提出一种通过提示来增强预训练模型视觉定位能力的方法 PTP,并应用在了 CLIP[101]、BLIP[113]以及 ViLT[109]上.此方法核心在于通过在图像和文本中添加基于位置的共同参考标记,视觉理解可以重新表述为一个填空问题,最大限度地简化目标信息的学习.PTP 包括两个部分,第一部分是块标签生成,将图片分割成多个块区域,在每个块中通过目标检测器或者 CLIP 模型预测块内目标的类别,判断其中包含哪些物体;第二部分是文本提示生成,根据第一部分生成块标签的情况形成类似“The block [P] has a [O].”的文本提示模版.基于这个生成的提示,有两种方式可以将其融合到预训练模型中进行增强,一种方式是将提示与预训练的图文对中的文本串接在一起整合到预训练阶段进行预训练;另一种方式是把位置预测看作新的语言建模预训练任务,在生成的提示上进行自回归预测的预训练.

类似地,Yao 等人[171]认为现有的视觉语言预训练模型虽然逐渐摆脱了对目标检测器的依赖,但是导致了在一些与位置预测相关的任务上的性能下

降,为此作者提出通过显式的目标位置建模来解决这个问题.基于 CLIP,BLIP 以及 ViLT,作者在预训练阶段将图片中目标的位置预测转变成一种掩码语言建模的形式.通过掩码预测的形式不仅建模了目标的位置信息,还建模了文本之间的信息,实现了将位置信息增强到预训练模型的目的.在下游任务中通过将所有任务给重构成广义掩码语言建模任务的形式实现在位置预测任务上的提示调优.

5.2.18 视觉定位

Yao 等人[56]认为在预训练阶段,大部分预训练模型均基于掩码语言建模任务,尝试从跨模态内容上恢复出掩码,在下游微调阶段,通常需要引入特定任务的参数在大规模有标签数据上将掩码预测映射到下游标签上.为了实现与预训练任务形式一致的视觉定位任务,作者提出了 CPT,将视觉定位任务重构成掩码语言建模形式的完形填空任务,这其中的关键在于建立图片区域与文本表示之间的联系.因此 CPT 包含两个设计:(1)视觉提示单独地将图片区域用颜色块进行标记;(2)文本提示将问题文本添加到基于颜色的提问模版中,如:"[CLS] a photo in [MASK] color [SEP]".通过这种方式,预训练模型就可以根据问题以及给定的带颜色的图片区域在掩码处预测出目标图片区域对应的颜色实现视觉定位.本方法中对图片区域与颜色的匹配尤为敏感,为此作者提出使用的颜色集合应该是预训练模型最为敏感的颜色.直接选择在预训练文本中出现频率较多的颜色会忽略视觉的表征,考虑到此,作者提出交叉模态提示搜索,联合考虑了视觉信息和文本语义表征的一致性,最终确定视觉区域与颜色的匹配关系.

5.2.19 多模态多语言机器翻译

Guo 等人[140]考虑到现有的多模态机器翻译任务仅限制应用在一对语言中,不同对语言之间的翻译需要训练不同的多个模型[172],计算成本高.为此,首次提出了多模态多语言机器翻译任务,并且提出了方法 LVP-M^3,实现了通过单个模型进行多种语言的翻译.其首先将图片通过预训练视觉模型提取特征得到视觉词,文本通过预训练语言模型得到文本词;之后设计了控制网络根据目标翻译语言动态生成映射网络的参数,映射网络则可以在语言翻译阶段根据视觉信息利用联合 Transformer 生成视觉引导的提示文本作为语言模型的输入;最后在语言模型的解码器处预测出翻译的结果.

5.2.20 序数回归(Ordinal Regression)

Li 等人[137]认为现有的数值回归方法将不同的排序看作独立的类别而忽略了他们之间的联系属性,并且容易在训练集上发生过拟合现象,为此提出 OrdinalCLIP 方法,它是从预训练模型 CLIP[101]丰富的语义空间中学习排序.该方法将每个排序标签转换为提示文本输入,将序数回归重构为一个图像-文本匹配问题.使用文本编码器来提取所有排序的文本原型,利用图像编码器从图像中提取视觉表征,然后选用一组最优的文本原型作为图像要匹配的文本特征.

5.2.21 图像编辑

相比于直接基于文本的图片区域生成,基于文本的图片编辑要求原来图像绝大部分区变化不大.而目前已有的纯文本引导的编辑方法只能修改图片的纹理等外观,而不能修改复杂的实体结构,比如把自行车换成一辆车.Hertz 等人[173]发现交叉注意力机制对于图片的布局控制很重要.为此,作者把交叉注意力图嵌入到扩散过程中,利用像素和文本之间的关系来控制生成,可以通过改变某个简单的词从而保证大部分场景不变,而改变小部分区域的置换.此外,该方法还可以全局地改变整个图或者加入一些新的信息.最重要的是,此方法不需要额外数据也不需要训练优化,只需要修改输入的提示语句即可实现图像编辑.

5.2.22 生成模型分布控制(Distributional Control of Generative Models)

生成模型基本以无监督的方式学习潜在的数据分布,但是许多应用都需要从生成模型的输出空间的特定区域或在一定特征范围内进行采样.为了实现在这些情况下的有效采样,Wu 等人[174]提出了生成视觉提示,通过合并任意现成模型的知识对预训练生成模型进行分布的控制.其提出了 PromptGen 方法,将分布控制定义为基于能量的模型 EBM[175].具体的,作者演示了提示方法如何控制如 StyleGAN2[176]等多种生成模型:(1)基于 CLIP[101]的模型控制,使用文本引导采样图像;(2)使用图像分类器控制,从一系列属性以及属性组合上消除生成模型的偏差;(3)基于反图模型控制采样与样本类似的图像;(4)CLIP 模型作为控制时会产生偏差的分布,而 PromtGen 可以通过迭代的方式消除这种偏差.

5.2.23 3D 识别

CLIP[101]针对各式各样的二维图像任务已经表现出了强泛化性,而它迁移到三维点云任务上的性能表现还有待提高.Zhu 等人[177]提出了针对开放世界的 3D 学习器,称为 PointCLIPV2.作者首先设计

了形状投影模块,以便 CLIP 的视觉编码器可以生成更真实的深度图,缩小投影点云与自然图像之间的领域差距;其次,利用大规模的语言模型,为 CLIP 的文本编码器自动设计了一个更具描述性的三维语义提示.在没有引入任何 3D 领域数据训练的情况下,PointCLIPV2 在零样本 3D 识别任务上超越了此前的方法,并且还可以扩展到小样本分类、零样本分割和零样本三维目标检测任务上. Hegde 等人[178]提出了一个 CG3D 框架,在 CLIP 已有知识的基础上,进一步采用“点云－图片－文本”三元组数据,用自然语言监督的方式训练一个 3D 识别网络.具体地,CG3D 通过提示调优将预训练的视觉编码器的输入空间从 CAD 对象渲染图像迁移到自然图像,从而更有效地用 CLIP 处理 3D 形状.此方法在零样本识别、语言场景理解和 3D 检索任务上都展现出了优异的性能表现.此外,CG3D 训练学习到的模型参数还可以作为多种 3D 识别任务模型的初始权重.

5.2.24　视频相关任务

提示学习还可以应用于视频相关任务.受限于目标检测器的性能表现以及高昂的计算代价等因素,现有的视频语言多模态预训练模型仍然没有有效解决视频数据和语言数据的细粒度对齐. Li 等人[179]提出了 ALPRO 方法,在预训练模型常用的视频文本匹配、视频文本对比以及掩码语言建模任务的基础上,额外设计了 PEM 模块,用于学习视频细粒度区域和语言实体匹配对齐关系,通过类似 CLIP 的方式来预测任意给定的视频切片中的实体名字. Gao 等人[180]提出了 Re-Pro 方法,解决在开放集视频视觉关系检测任务中针对现有的方法容易朝着特定的主客体组合以及模式产生偏差的问题. Re-Pro 在设计提示词的过程中充分考虑主客体两种不同的语义角色,在提示调优的过程中充分考虑主客体组合的不同关系.此外,还分别设置了由主语和宾语指定的组合式提示表示,基于这种设计,此方法可以根据语义角色,即主题或对象,对提示上下文进行建模.

在视频理解任务中,Ju 等人[181]基于一个图片理解模型,在其视觉编码器后额外串接了一个可学习的 Transformer 模块,用来建模获取时序关系特征,最终应用于视频动作识别、视频－文本检索和动作定位等任务中.在文本视频检索任务中,VoP[182]在每个模态编码器的每一层都添加提示进行调优,其中包括三种类型的提示,分别是建模帧之间相关位置信息的位置特定的视频提示、将帧之间的上下文信息整合到单帧内部的上下文特定的视频提示以及学习帧内与帧间关系的功能特定的视频提示.对于零样本视频动态检索任务,Wang 等人[183]在时域候选框生成阶段,用连续的空白图像随机掩码序列帧,以同时获得局部和全局的视觉上下文特征;在伪 query 生成阶段,使用现成的目标检测器来识别视觉目标,同时屏蔽动作词,通过提示调优构建生成动词的提示,最终名词和生成的动词一起构成伪查询.在视频多模态追踪任务中,目前仍然存在严重的标签数据不足问题,导致模型中的多模态融合模块难以进行有效训练.在多模态跟踪任务中,需要视觉模态以及其他模态数据的辅助,而视觉模态数据的分辨度影响着跟踪性能,其他辅助模态数据在某些特定场景下才有效.考虑到辅助模态数据提供的信息量通常较少,Yang 等人[184]提出利用提示学习将多模态输入转化成单模态的输入形式,即把辅助模态的数据通过提示转变到可视模态中从而减少分布的差异,之后将转变后的多模态数据输入到预训练模型的 RGB 跟踪器中进行训练学习.

6　展　望

在自然语言处理领域,提示学习方法通过将各类下游任务重构成预训练任务的形式,实现多任务统一的下游应用,并且取得了出色的效果.这得益于自然语言处理领域的两个特性:(1)各类大规模语言模型的预训练任务相似度高,即:通过语言建模的方式,基于可见的文本序列预测生成不可见的目标文本.这些预训练任务有效实现了模型对语料的综合理解表征能力.(2)下游任务的解都可以作为语言输出表达,因此可以很自然地将不同的语言下游任务都重构成预训练阶段的语言建模任务,将下游的解通过预训练方式进行生成式预测.

我们基于视觉及多模态的提示学习方法与自然语言处理领域提示学习方法的对比,从以下几个方面进行分析:

首先,目前视觉及多模态提示学习方法均是面向特定任务进行特殊设计的方法,少有实现多任务统一的且与预训练任务保持一致的提示学习方法.我们认为存在以下两个主要原因:(1)视觉和多模态模型的预训练任务繁多复杂,例如有监督的分类预训练任务[91-100]、对比式自监督预训练任务[101]、图像恢复式的生成式预训练任务[105-106]等,这些预训练

任务形式各不相同;(2)下游任务的解的形式也各不相同,如图像分类任务需要输出分类标签,目标检测任务需要输出具体的目标坐标位置及对应类别,语义分割任务需要输出分割的区域图等.由于预训练任务之间的巨大差异、下游任务的解之间的巨大差异,以及预训练和下游任务形式之间的巨大差异,要基于现有的视觉或多模态预训练模型实现多任务统一且保留预训练和下游任务一致性的提示学习方法存在很大的困难.

其次,虽然在自然语言处理、视觉以及多模态领域中,提示学习可以实现参数高效的微调,但是在基于下游数据调优的整个过程中梯度依旧需要进行回传优化,并没有带来时间上的高效性.

此外,目前视觉和多模态领域的大部分模型参数规模都要远小于自然语言处理领域模型的参数规模,限制了模型可学习的知识量以及最终对数据的理解表征能力.

基于以上分析,我们从以下几个角度给出未来研究的可行方向:

(1)现有的大部分视觉提示学习方法还没有实现像文本提示学习方法中的将各类下游任务统一成预训练的形式,而是面向特定任务进行特定设计的提示学习方法.因此,探索将视觉里的各种任务重构成预训练任务形式从而保持一致性的提示学习方法是未来研究方向之一.

(2)现有的视觉和多模态提示学习方法绝大多数都是基于判别式预训练模型进行设计的,反观自然语言处理领域,提示学习方法基本都是基于自监督生成式预训练模型进行设计的.因此,探索在生成式预训练视觉和多模态模型上的提示学习方法是未来一个研究方向.

(3)探索多模态统一、多任务统一的预训练方式,并且通过提示学习实现与预训练任务形式一致的下游应用也是未来的研究方向之一.

(4)从模型参数规模角度,视觉模型参数量相比语言预训练模型的参数量要小得多.提示学习作为一种高效地利用预训练模型进行下游任务的方法,性能表现很大程度上取决于模型本身在预训练阶段学习到的知识.因此,进一步扩大视觉和多模态预训练模型的规模,从而提升模型的知识学习以及理解表征能力是未来方向之一.

(5)在预训练阶段考虑将提示学习整合进来,在预训练阶段就做到预训练形式与下游任务的形式对齐,从而无需可以在下游任务阶段花费大量时间进行提示调优.这也是未来的一个研究方向.

(6)将 GPT 系列模型[185-186]进行小型化,并且研究使其能解决视觉中更复杂困难的任务也成为未来的研究方向.

(7)对于以文本生成为主的下游任务,现在的提示学习方法还不具有较好的与人类指令对齐的能力.通过收集高质量的指令数据来指导模型根据指令提示输出与人类需求对齐的答案成为未来研究方向之一.

7 总 结

在本文中,我们首先回顾了自然语言处理领域的预训练模型以及基于它们的提示学习方法;其次,我们分别介绍了视觉以及多模态领域提示学习适配的预训练模型;之后,我们分别从方法设计角度和下游任务角度进行简单分类,详细介绍了视觉以及多模态领域的提示学习方法;最后,我们对比了自然语言处理领域以及视觉和多模态领域提示学习方法的现状,并且进行了分析,给出了视觉和多模态领域提示学习未来可能的研究方向.

参 考 文 献

[1] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. OpenAI, 2018:1-12

[2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[3] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010

[5] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 7871-7880

[6] Petroni F, Rocktäschel T, Riedel S, et al. Language Models as Knowledge Bases? //Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019:80-89

[7] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Online, 2020: 1877-1901

[8] Qin G, Eisner J. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021: 5203-5212

[9] Gao T, Fisch A, Chen D. Making Pre-trained Language Models Better Few-shot Learners//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 3816-3830

[10] Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online, 2021: 3045-3059

[11] Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online, 2021: 255-269

[12] Cui L, Wu Y, Liu J, et al. Template-based named entity recognition using BART//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 1835-1845

[13] Rajani N F, McCann B, Xiong C, et al. Explain yourself! Leveraging language models for commonsense reasoning//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4932-4942

[14] Ettinger A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 2020, 8(2): 34-48

[15] Khashabi D, Min S, Khot T, et al. UNIFIEDQA: Crossing format boundaries with a single QA system//Findings of the Association for Computational Linguistics: EMNLP 2020. Online, 2020: 1896-1907

[16] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, 2020: 9729-9738

[17] Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers//Proceedings of the CVF International Conference on Computer Vision (ICCV). Online, 9620-9629

[18] Jia M, Tang L, Chen B C, et al. Visual prompt tuning//Proceedings of the Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, Part XXXIII. Cham, Switzerland: Springer, 2022: 709-727

[19] Wang Z, Zhang Z, Lee C Y, et al. Learning to prompt for continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 139-149

[20] Li A, Zhuang L, Fan S, et al. Learning Common and Specific Visual Prompts for Domain Generalization//Proceedings of the Asian Conference on Computer Vision. Cham: Springer Nature Switzerland, Macau, China, 2022: 578-593

[21] Bahng H, Jahanian A, Sankaranarayanan S, et al. Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274, 2022, 3: 11-12

[22] Wu J, Li X, Wei C, et al. Unleashing the Power of Visual Prompting At the Pixel Level. arXiv preprint arXiv: 2212.10556, 2022

[23] Chen A, Lorenz P, Yao Y, et al. Visual prompting for adversarial robustness//Proceedings of the ICASSP 2023－2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece, 2023: 1-5

[24] Nie X, Ni B, Chang J, et al. Pro-tuning: Unified prompt tuning for vision tasks. IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2023.3327605

[25] Loedeman J, Stol M C, Han T, et al. Prompt generation networks for efficient adaptation of frozen vision transformers. arXiv preprint arXiv:2210.06466, 2022

[26] Wang H, Chang J, Luo X, et al. Lion: Implicit vision prompt tuning. arXiv preprint arXiv:2303.09992, 2023

[27] Huang Q, Dong X, Chen D, et al. Diversity-aware meta visual prompting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 10878-10887

[28] Zheng Z, Yue X, Wang K, et al. Prompt vision transformer for domain generalization. arXiv preprint arXiv:2208.08914, 2022

[29] Yang Z, Sha Z, Backes M, et al. From visual prompt learning to zero-shot transfer: Mapping is all you need. arXiv preprint arXiv:2303.05266, 2023

[30] Chen A, Yao Y, Chen P Y, et al. Understanding and improving visual prompting: A label-mapping perspective//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19133-19143

[31] Liao N, Shi B, Zhang X, et al. Rethinking visual prompt learning as masked visual token modeling. arXiv preprint arXiv:2303.04998, 2023

[32] Zhang Y, Zhou K, Liu Z. Neural prompt search. arXiv preprint arXiv:2206.04673, 2022

[33] Wang Z, Zhang Z, Ebrahimi S, et al. Dualprompt: Comple-

mentary prompting for rehearsal-free continual learning// European Conference on Computer Vision. Cham: Springer Nature Switzerland, Tel Aviv, Israel, 2022: 631-648

[34] Smith J S, Karlinsky L, Gutta V, et al. CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 11909-11919

[35] Gao Y, Shi X, Zhu Y, et al. Visual prompt tuning for test-time domain adaptation. arXiv preprint arXiv:2210.04831, 2022

[36] Wang S, Chang J, Wang Z, et al. Fine-grained retrieval prompt tuning//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA, 2023, 37(2): 2644-2652

[37] Liu L, Chang J, Yu B X B, et al. Prompt-matched semantic segmentation. arXiv preprint arXiv:2208.10159, 2022

[38] Dong B, Zhou P, Yan S, et al. Lpt: Long-tailed prompt tuning for image classification. arXiv preprint arXiv:2210.01033, 2022

[39] Li H, Feng C M, Zhou T, et al. Prompt-driven efficient open-set semi-supervised learning. arXiv preprint arXiv: 2209.14205, 2022

[40] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022, 130(9): 2337-2348

[41] Yu T, Lu Z, Jin X, et al. Task residual for tuning vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 10899-10909

[42] Liao N, Zhang X, Cao M, et al. R-Tuning: Regularized Prompt Tuning in Open-Set Scenarios. arXiv preprint arXiv: 2303.05122, 2023

[43] Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA, 2022: 16816-16825

[44] Guo J, Li J, Li D, et al. From images to textual prompts: Zero-shot vqa with frozen large language models. arXiv preprint arXiv:2212.10846, 2022

[45] Bulat A, Tzimiropoulos G. Language-aware soft prompting for vision & language foundation models. arXiv preprint arXiv: 2210.01115, 2022

[46] Yao H, Zhang R, Xu C. Visual-language prompt tuning with knowledge-guided context optimization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada,2023: 6757-6767

[47] Zang Y, Li W, Zhou K, et al. Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225, 2022

[48] Khattak M U, Rasheed H, Maaz M, et al. Maple: Multi-modal prompt learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19113-19122

[49] Zhang Y, Fei H, Li D, et al. Prompting through prototype: A prototype-based prompt learning on pretrained vision-language models. arXiv preprint arXiv:2210.10841, 2022

[50] Lee Y L, Tsai Y H, Chiu W C, et al. Multimodal prompting with missing modalities for visual recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14943-14952

[51] Lu Y, Liu J, Zhang Y, et al. Prompt distribution learning// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA, 2022: 5206-5215

[52] Derakhshani M M, Sanchez E, Bulat A, et al. Variational prompt tuning improves generalization of vision-language foundation models//ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models. Kigali, Rwanda, 2023

[53] Shen S, Yang S, Zhang T, et al. Multitask vision-language prompt tuning//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 5656-5667

[54] Zhu B, Niu Y, Han Y, et al. Prompt-aligned gradient for prompt tuning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15659-15669

[55] Huang T, Chu J, Wei F. Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649, 2022

[56] Yao Y, Zhang A, Zhang Z, et al. Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797, 2021

[57] Tsimpoukelli M, Menick J L, Cabi S, et al. Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, Online, 2021, 34: 200-212

[58] Bose S, Jha A, Fini E, et al. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 5542-5552

[59] Ge C, Huang R, Xie M, et al. Domain adaptation via prompt learning. IEEE Transactions on Neural Networks and Learning Systems,doi: 10.1109/TNNLS.2023.3327962

[60] Fahes M, Vu T H, Bursuc A, et al. PODA: Prompt-driven zero-shot domain adaptation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 18623-18633

[61] Liu T Y, Wu Z X, Chen J J, et al. Multimodal pre-training method for vision-language understanding and generation. Journal of Software, 2022, 34(5): 2024-2034(in Chinese)
刘天义,吴祖煊,陈静静,等.面向视觉语言理解与生成的多模态预训练方法.软件学报,2022,34(5):2024-2034

[62] Jin W, Cheng Y, Shen Y, et al. A good prompt is worth

millions of parameters: Low-resource prompt-based Learning for vision-language models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 2763-2775

[63] Wang N, Xie J, Wu J, et al. Controllable image captioning via prompting//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(2): 2617-2625

[64] Zhu P, Wang X, Zhu L, et al. Prompt-based learning for unpaired image captioning. IEEE Transactions on Multimedia, doi: 10.1109/TMM.2023.3265842

[65] He X, Yang D, Feng W, et al. CPL: Counterfactual prompt learning for vision and language models//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 3407-3418

[66] Chen Z, Zhou Q, Shen Y, et al. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. arXiv preprint arXiv: 2301.05226, 2023

[67] Sun X, Hu P, Saenko K. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems, New Orleans, USA, 2022, 35: 30569-30582

[68] Liao N, Liu Y, Xiaobo L, et al. CoHOZ: Contrastive multimodal prompt tuning for hierarchical open-set zero-shot recognition//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 3262-3271

[69] Chuang C Y, Jampani V, Li Y, et al. Debiasing vision-language models via biased prompts. arXiv preprint arXiv: 2302.00070, 2023

[70] Berg H, Hall S, Bhalgat Y, et al. A prompt array keeps the Bias away: Debiasing vision-language models with adversarial learning//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Online, 2022: 806-822

[71] Nayak N V, Yu P, Bach S H. Learning to compose soft prompts for compositional zero-shot learning. arXiv preprint arXiv:2204.03574, 2022

[72] Xu G, Kordjamshidi P, Chai J. Prompting large pre-trained vision-language models for compositional concept learning. arXiv preprint arXiv:2211.05077, 2022

[73] Lüddecke T, Ecker A. Image segmentation using text and image prompts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA, 2022: 7086-7096

[74] Kim K, Oh Y, Ye J C. Zegot: Zero-shot segmentation through optimal transport of text prompts. arXiv preprint arXiv:2301.12171, 2023

[75] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9

[76] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 1-35

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

[78] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens//Proceedings of the International conference on machine learning. Hawaii, USA, 2022: 2206-2240

[79] Black S, Biderman S, Hallahan E, et al. Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745, 2022

[80] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy,2019: 1441-1451

[81] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv: 1904.09223, 2019

[82] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 13063-13075

[83] Bao H, Dong L, Wei F, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training//Proceedings of theInternational conference on machine learning. Online, 2020: 642-652

[84] Ouyang X, Wang S, Pang C, et al. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic,2021: 27-38

[85] Song K, Tan X, Qin T, et al. MASS: Masked sequence to sequence pre-training for language generation//International Conference on Machine Learning. Long Beach, USA, 2019: 5926-5936

[86] Hu S, Ding N, Wang H, et al. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 2225-2240

[87] Li X L, Liang P. Prefix-Tuning: Optimizing continuous prompts for generation//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 4582-4597

[88] Yuan W, Neubig G, Liu P. Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, Online, 2021, 34: 27263-27277

[89] Wallace E, Feng S, Kandpal N, et al. Universal adversarial triggers for attacking and analyzing NLP//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019

[90] Shin T, Razeghi Y, Logan IV R L, et al. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 4222-4235

[91] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the2009 IEEE conference on computer vision and pattern recognition. Miami, USA, 2009: 248-255

[92] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA, 2016: 770-778

[93] Zagoruyko S, Komodakis N. Wide residual networks//British Machine Vision Conference 2016. British Machine Vision Association, York, UK, 2016

[94] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA, 2017: 1492-1500

[95] Mahajan D, Girshick R, Ramanathan V, et al. Exploring the limits of weakly supervised pretraining//Proceedings of the European conference on computer vision (ECCV). Munich, Germany, 2018: 181-196

[96] Kolesnikov A, Beyer L, Zhai X, et al. Big transfer (bit): General visual representation learning//Proceedings of the-Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, Part V 16. Springer International Publishing, 2020: 491-507

[97] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of theInternational Conference on Learning Representations. Online, 2020

[98] Radosavovic I, Kosaraju R P, Girshick R, et al. Designing network design spaces//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, 2020: 10428-10436

[99] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention//International conference on machine learning. Online, 2021: 10347-10357

[100] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF international conference on computer vision. Online, 2021: 10012-10022

[101] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//International conference on machine learning. PMLR, Online, 2021: 8748-8763

[102] Peng Z, Dong L, Bao H, et al. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366, 2022

[103] Bar A, Gandelsman Y, Darrell T, et al. Visual prompting via image inpainting. Advances in Neural Information Processing Systems, New Orleans, Louisiana, USA, 2022, 35: 25005-25017

[104] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, 2021: 12873-12883

[105] Bao H, Dong L, Piao S, et al. BEiT: BERT pre-training of image transformers//International Conference on Learning Representations. Online, 2021

[106] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16000-16009

[107] Sohn K, Chang H, Lezama J, et al. Visual prompt tuning for generative transfer learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19840-19851

[108] Chang H, Zhang H, Jiang L, et al. Maskgit: Masked generative image transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11315-11325

[109] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision//International Conference on Machine Learning. Online, 2021: 5583-5594

[110] Wang P, Yang A, Men R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//International Conference on Machine Learning. Baltimore, USA, 2022: 23318-23340

[111] Wang B, Komatsuzaki A. GPT-J-6B: A 6 billion parameter autoregressive language model. 2021

[112] Brock A, De S, Smith S L, et al. High-performance large-scale image recognition without normalization//Proceedings of theInternational Conference on Machine Learning. Online, 2021: 1059-1071

[113] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of theInternational Conference on Machine Learning. Baltimore, USA, 2022: 12888-12900

[114] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:

2205.01068, 2022

[115] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP//Proceedings of theInternational Conference on Machine Learning. Long Beach, USA, 2019: 2790-2799

[116] Hu E J, Wallis P, Allen-Zhu Z, et al. LoRA: Low-rank adaptation of large language models//Proceedings of theInternational Conference on Learning Representations. Online, 2021

[117] Garcia X, Firat O. Using natural language prompts for machine translation. arXiv preprint arXiv:2202.11822, 2022

[118] Min S, Lewis M, Zettlemoyer L, et al. MetaICL: Learning to learn in context//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, Washington, USA, 2022: 2791-2809

[119] Frankle J, Carbin M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv: 1803.03635, 2018

[120] McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: The sequential learning problem//Psychology of learning and motivation. Academic Press, 1989, 24: 109-165

[121] Farquhar S, Gal Y. Towards robust evaluations of continual learning. arXiv preprint arXiv:1805.09733, 2018

[122] Hadsell R, Rao D, Rusu A A, et al. Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences, 2020, 24(12): 1028-1040

[123] Wang J, Lan C, Liu C, et al. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 8, pp. 8052-8072, 1 Aug. 2023, doi: 10.1109/TKDE.2022.3178128

[124] Teh E W, DeVries T, Taylor G W. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis//Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, Part XXIV 16. Springer, 2020: 448-464

[125] Moskvyak O, Maire F, Dayoub F, et al. Keypoint-aligned embeddings for image retrieval and re-identification//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Online, 2021: 676-685

[126] Zhang Y, Zhou K, Liu Z. What makes good examples for visual in-context learning? arXiv preprint arXiv: 2301.13670, 2023

[127] Kang B, Xie S, Rohrbach M, et al. Decoupling representation and classifier for long-tailed recognition//Proceedings of the8th International Conference on Learning Representations. Online, 2020

[128] Li M, Cheung Y, Lu Y. Long-tailed visual recognition via gaussian clouded logit adjustment//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans,USA, 2022: 6929-6938

[129] Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA, 2019: 9268-9277

[130] Menon A K, Jayasumana S, Rawat A S, et al. Long-tail learning via logit adjustment//International Conference on Learning Representations. Online, 2020

[131] Zhang Y, Kang B, Hooi B, et al. Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9):10795-10816

[132] Li T, Wang L, Wu G. Self supervision to distillation for long-tailed visual recognition//Proceedings of the IEEE/CVF international conference on computer vision. Online, 2021: 630-639

[133] Zhou B, Cui Q, Wei X S, et al. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, 2020: 9719-9728

[134] Shu M, Nie W, Huang D A, et al. Test-time prompt tuning for zero-shot generalization in vision-language models//Proceedings of theAdvances in Neural Information Processing Systems, New Orleans, USA, 2022

[135] Du Y, Wei F, Zhang Z, et al. Learning to prompt for open-vocabulary object detection with vision-language model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA, 2022: 14084-14093

[136] Feng C, Zhong Y, Jie Z, et al. Promptdet: Expand your detector vocabulary with uncurated images. arXiv preprint arXiv:2203.16513, 2022

[137] Li W, Huang X, Zhu Z, et al. Ordinalclip: Learning rank prompts for language-guided ordinal regression. Advances in Neural Information Processing Systems, New Orleans, USA, 2022, 35: 35313-35325

[138] Zhang X, Gu S S, Matsuo Y, et al. Domain prompt learning for efficiently adapting clip to unseen domains. Transactions of the Japanese Society for Artificial Intelligence, 2023, 38(6): B-MC2_1-10

[139] Mañas O, Lopez P R, Ahmadi S, et al. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Ceoatia, 2023: 2515-2540

[140] Guo H, Liu J, Huang H, et al. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 2862-2872

[141] Xing Y, Wu Q, Cheng D, et al. Class-aware visual prompt tuning for vision-language pre-trained model. arXiv preprint

arXiv:2208.08340, 2022

[142] Zhao C, Wang Y, Jiang X, et al. Learning domain invariant prompt for vision-language models. arXiv preprint arXiv: 2212.04196, 2022

[143] Long Y, Han J, Huang R, et al. P3OVD: Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection. arXiv preprint arXiv:2211.00849, 2022

[144] Yang H, Lin J, Yang A, et al. Prompt tuning for generative multimodal pretrained models. arXiv preprint arXiv: 2208.02532, 2022

[145] Guo Z, Dong B, Ji Z, et al. Texts as images in prompt tuning for multi-label image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 2808-2817

[146] Chen G, Yao W, Song X, et al. PLOT: Prompt learning with optimal transport for vision-language models//The Eleventh International Conference on Learning Representations. Online, 2022

[147] Villani C. Optimal transport: old and new. Berlin,Germany: springer, 2009

[148] Liu X, Wang D, Li M, et al. Patch-token aligned bayesian prompt learning for vision-language models. arXiv preprint arXiv:2303.09100, 2023

[149] Ding K, Wang Y, Liu P, et al. Prompt tuning with soft context sharing for vision-language models. arXiv preprint arXiv:2208.13474, 2022

[150] Ma C, Liu Y, Deng J, et al. Understanding and mitigating overfitting in prompt tuning for vision-language models. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9):4616-4629

[151] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems. Harrahs and Harveys, USA, 2013, 26

[152] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017: 1501-1510

[153] Wang Y, Huang Z, Hong X. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 5682-5695

[154] Krishna R, Zhu Y, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 2017, 123: 32-73

[155] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks//Advances in Neural Information Processing Systems, Montréal, Canada,2015, 28

[156] Chen T H, Liao Y H, Chuang C Y, et al. Show, adapt and tell: Adversarial training of cross-domain image captioner//Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017: 521-530

[157] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning//International Conference on Learning Representations. Caribe Hilton, Puerto Rico, 2016

[158] Han X, Yang J, Hu H, et al. Image scene graph generation (sgg) benchmark. arXiv preprint arXiv:2107.12604, 2021

[159] Ridnik T, Ben-Baruch E, Zamir N, et al. Asymmetric loss for multi-label classification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Online, 2021: 82-91

[160] Bendale A, Boult T. Towards open world recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, Massachusetts, USA, 2015: 1893-1902

[161] Sun X, Yang Z, Zhang C, et al. Conditional gaussian distribution learning for open set recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2020: 13480-13489

[162] Esmaeilpour S, Liu B, Robertson E, et al. Zero-shot out-of-distribution detection based on the pre-trained model clip//Proceedings of the AAAI conference on artificial intelligence. Virginia, USA,2022, 36(6): 6568-6576

[163] Menon S, Chandratreya I P, Vondrick C. Task Bias in Vision-Language Models. arXiv preprint arXiv: 2212.04412, 2022

[164] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 10684-10695

[165] Allingham J U, Ren J, Dusenberry M W, et al. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models//International Conference on Machine Learning. Hawaii, USA, 2023: 547-568

[166] Schuhmann C, Kaczmarczyk R, Komatsuzaki A, et al. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs//NeurIPS Workshop Datacentric AI. Jülich Supercomputing Center, Online, 2021 (FZJ-2022-00923)

[167] Arevalo J, Solorio T, Montes-y-Gómez M, et al. Gated multimodal units for information fusion. arXiv preprint arXiv: 1702.01992, 2017

[168] Wang X, Kumar D, Thome N, et al. Recipe recognition with large multimodal food dataset//2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Turin, Italy, 2015: 1-6

[169] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems, Online, 2020, 33: 2611-2624

[170] Wang J, Zhou P, Shou M Z, et al. Position-guided text

prompt for vision-language pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 23242-23251

[171] Yao Y, Chen Q, Zhang A, et al. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 11104-11117

[172] Lin H, Meng F, Su J, et al. Dynamic context-guided capsule network for multimodal machine translation//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 1320-1329

[173] Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with cross-attention control//Proceedings of theEleventh International Conference on Learning Representations. Online, 2022

[174] Wu C H, Motamed S, Srivastava S, et al. Generative visual prompt: Unifying distributional control of pre-trained generative models. Advances in Neural Information Processing Systems, New Orleans, USA, 2022, 35: 22422-22437

[175] LeCun Y, Chopra S, Hadsell R, et al. A tutorial on energy-based learning. Predicting structured data, 2006, 1(0), pp. 191-241

[176] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, 2020: 8110-8119

[177] Zhu X, Zhang R, He B, et al. PointCLIP V2: Adapting CLIP for powerful 3D open-world learning. arXiv preprint arXiv:2211.11682, 2022

[178] Hegde D, Valanarasu J M J, Patel V. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 2028-2038

[179] Li D, Li J, Li H, et al. Align and prompt: Video-and-language pre-training with entity prompts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4953-4963

[180] Gao K, Chen L, Zhang H, et al. Compositional prompt tuning with motion cues for open-vocabulary video relation detection//The Eleventh International Conference on Learning Representations. Online, 2022

[181] Ju C, Han T, Zheng K, et al. Prompting visual-language models for efficient video understanding//Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV. Cham, Switzerland: Springer, 2022: 105-124

[182] Huang S, Gong B, Pan Y, et al. VoP: Text-video co-operative prompt tuning for cross-modal retrieval//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6565-6574

[183] Wang G, Wu X, Liu Z, et al. Prompt-based zero-shot video moment retrieval//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 413-421

[184] Yang J, Li Z, Zheng F, et al. Prompting for multi-modal tracking//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 3492-3500

[185] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, New Orleans, USA, 2022, 35: 27730-27744

[186] OpenAI. GPT-4 technical report. arXiv preprint arXiv: 2303.08774, 2023

LIAO Ning, Ph. D. candidate. His main research interests include large multimodal model, prompt/instruction learning, open-setrecognition.

CAO Min, Ph. D., associate professor. Her main research interests include vision-language learning, anomaly detection.

YAN Jun-Chi, Ph. D., professor. His main research interests include machine learning, as well as the intersection with combinatorial optimization and quantum computing.

Background

With the rapid development of large language models, exploring efficient ways to leverage them in downstream applications becomes a significant problem. The prior paradigm, namely "pre-train, fine-tune", exhibits strong performance yet requires optimizing the entire pre-trained models with a great amount of parameters with different specific objectives in the face of different tasks. Such a paradigm is prohibitive. Recently, a new paradigm known as "pre-train, prompt, predict" has been proposed in the natural language processing (NLP) field. It aims at reformulating various downstream tasks in the same form as the pre-training one to narrow down the task gap, by which downstream tasks could be resolved as that in pre-training phase. This paradigm is parameter- and data-efficient and has succeeded in NLP.

Inspired by the above, prompt learning has been widely studied in unimodal vision and multimodal vision-language,

which pursues the same goal as learning in NLP, i. e. , leveraging large-scale pre-trained models to perform downstream applications efficiently. However, there exists much difference between NLP and vision-language fields, both in the pre-training and downstream tasks. To this end, we aim to conduct a comprehensive survey on the prompt learning methods in unimodal vision and multimodal vision-language fields. Specifically, we first introduce the pre-trained models and prompt learning methods in NLP as the preliminary. Then, we give a brief introduction on the pre-trained models in the unimodal vision and multimodal vision-language fields. After that, we introduce the current prompt learning methods in unimodal vision and multimodal vision-language areas. We briefly summarize them from method design, and give the details from the perspective of downstream tasks.

Till now, prompt learning in vision and multimodal areas still faces the difficulty in unifying the downstream tasks as the pre-training ones, due to the inherent difference between tasks. It is because of the inherent difference between pre-training and applications. In these two fields, models are pre-trained by supervised classification, contrastively self-supervised learning, or image modeling, while downstream tasks are much denser than the pre-training ones. Based on this, we deliver our opinions in future research. We hope this survey could attract more researchers' attention and contribute in this direction.