提示学习在视觉语言多模态领域中的应用

甘晴void

湖南大学信息科学与工程学院
学号：202108010XXX

摘 要 近年来，提示学习在自然语言处理与视觉语言多模态领域中应用广泛。多种提示学习方法百花齐放，为视觉语言多模态领域的多任务融合、多模型训练与其下游推理任务作出了很大贡献。本篇简要梳理了该领域的提示学习近年来的学界成果，并在篇章中对不同方法作出了对比。篇章包括了纯文本、视觉信息引导、外部知识引导、文本和视觉联合、面向特定成分组合、基于分布、多任务共享、梯度引导、无监督、建立颜色与标签关系、视觉映射到语言空间等提示学习方法，并就每个方法举了几个较为具体的工作进行展示。最后是学习后的心得。

The Application of Prompt Learning in Visual language multimodality

MEI Bing-Yin

College of Computer Science and Electronic Engineering,
HuNan University, Changsha Hunan

Abstract In recent years, prompt learning has been widely applied in the fields of natural language processing and multimodal visual language. A variety of prompt learning methods have emerged, making significant contributions to multi task fusion, multi model training, and downstream inference tasks in the field of visual language multimodality. This article briefly summarizes the academic achievements of prompt learning in this field in recent years, and compares different methods in the chapter. The chapter includes pure text, visual information guidance, external knowledge guidance, text visual collaboration, targeting specific component combinations, distribution based, multi task sharing, gradient guidance, unsupervised, establishing color label relationships, visual mapping to language space, and other prompt learning methods. Several specific works are presented for each method. Finally, the learning experience.

1 引言

自 Transformer[1]被提出以来，基于此的大规模预训练语言模型被相继提出，其应用也非常广泛。以 GPT[2]，BERT[3]，T5[4]等为主的“预训练-微调”范式模型与以 LAMA[5]、GPT-3[6]为主的“预训练-提示-预测”范式模型均在视觉单模态以及视觉语言多模态领域大放异彩。

针对自然语言处理领域中的提示学习方法已经有相关综述[7]展开了全面的介绍，针对视觉领域中的提示学习方法也有相关综述[8]展开了系统性的介绍。但是对于视觉语言多模态领域的综述中，对各方法的阐述尚不是很全面。因此，本文对提示学习方法在自视觉语言多模态领域的方法进行一个较为全面的总结，并进行简单的对比，作为本人在该领域初步学习的一个成果。

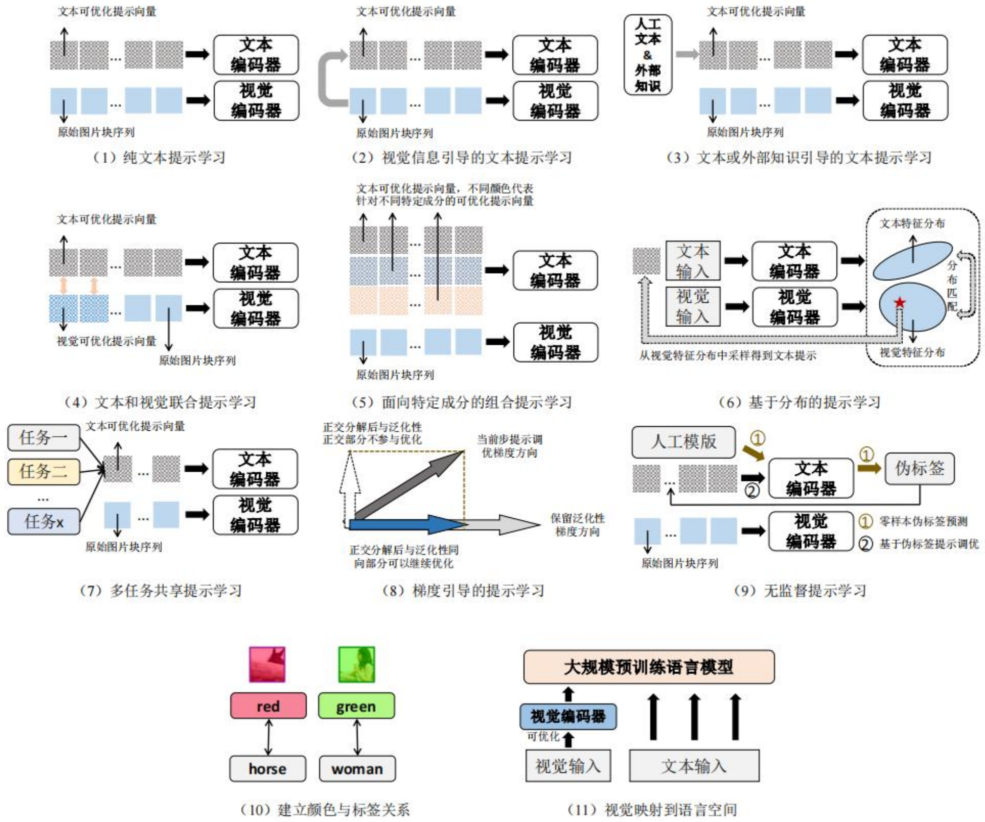
根据阅读综述，以及搜集资料，在视觉语言多模态领域，目前主流的提示学习方法包括以下几种：

- 纯文本提示学习
- 视觉信息引导的文本提示学习
- 文本或外部知识引导的文本提示学习
- 文本和视觉联合提示学习
- 面向特定成分的组合提示学习
- 基于分布的提示学习
- 多任务共享的提示学习
- 梯度引导的提示学习
- 无监督提示学习
- 建立颜色与标签关系
- 视觉映射到语言空间

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、基础到新类别泛化、领域泛化、领域适应、视觉问答、图片描述、图文检索、视觉蕴含、视觉推理、多标签分类、开放集识别、去偏差提示学习、组合零样本学习、图像分割等。

2 提示学习方法简析

下图是综述中关于这 11 种方法的形象阐释。接下来我将会逐个进行较为细致的分析。



（1）纯文本提示学习　（2）视觉信息引导的文本提示学习　（3）文本或外部知识引导的文本提示学习

（4）文本和视觉联合提示学习　（5）面向特定成分的组合提示学习　（6）基于分布的提示学习

（7）多任务共享提示学习　（8）梯度引导的提示学习　（9）无监督提示学习

（10）建立颜色与标签关系　（11）视觉映射到语言空间

图 1

2.1 纯文本提示学习

纯文本提示学习技术最先在如CLIP[9]这种双塔结构的多模态预训练模型上得以应用。CLIP 也就是本组完成课程设计时打比赛所用到的开源项目原型，该模型有文本侧和图像侧两个骨架，并提供分别训练或联合训练的接口。

该种方法最开始的设计是对类似如“a photo of a ___.”的人工提示模版在空白处添加类别词。接着分别将提示句和图片输入到文本编码器和视觉编码器中提取特征，最后再进行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。

这种人工模版需要很大的工作量，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[10]等连续提示学习方法启发，纯文本连续提示学习方法[11]被提出。如图 1（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。

如下图，在做小样本训练时，主模型(text encoder image encoder)不参与训练。我们使用梯度去更新它的 learnable context，去更新 learnable context，这样用较少的样本就可以训练一个更好的 learnable context，在测试的时候使用 learnable context 参与测试，CoOp[11]由此被提出，该方法在小样本图片分类上取得了非常显著的效果。

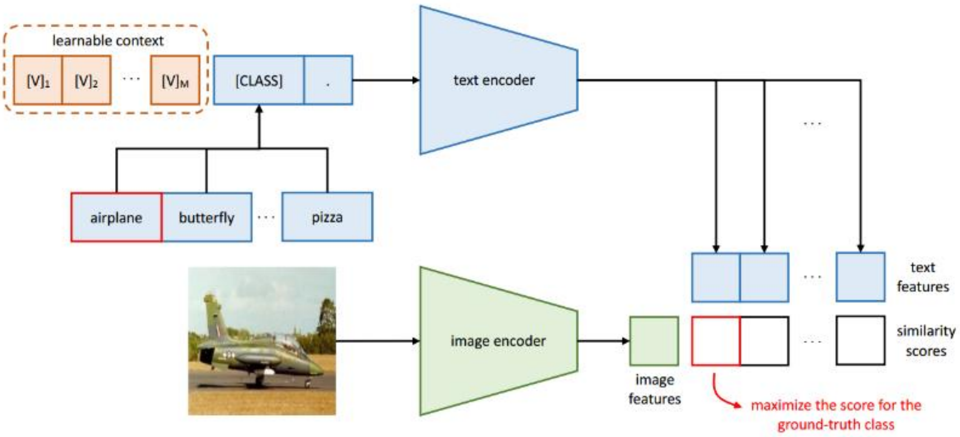


图 2

2.2 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。为了解决上述问题，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 1（2）所示。

围绕这个想法，有多种实现方式。

CoCoOp[12]、DPL、StyLIP 等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。在这之中，CoCoOp 是对于上面提到的 CoOp 的改进，这源于研究中发现的 CoOp 的一个问题：泛化性差，即学习的上下文向量不能推广到同一数据集中的未知类，这表明 CoOp 在训练时过拟合到了 base classes。CoOp 的 context 仅针对一组特定的类别上优化，而且在学习之后就固定了，一种简单的解决方法就是将 CoOp 中 unified context，变成 instance-adaptive context。这样对于每个样本都有一个特定的 prompt，从而会让 context 的聚焦从一组特定类别上，转移到每个样本的特征或者属性上，进而就减少了过拟合，对于 class shift 更加鲁棒。

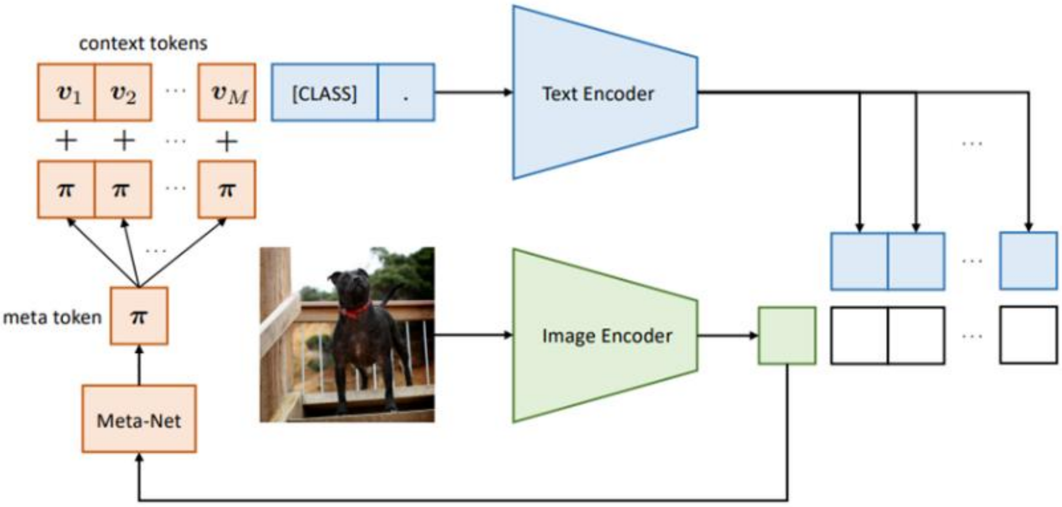


图 3

MAPL、PL-UIC[13]、LVP-M3 等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。下图以 PL-UIC[13]为例，其最左下角就是这个映射过程。

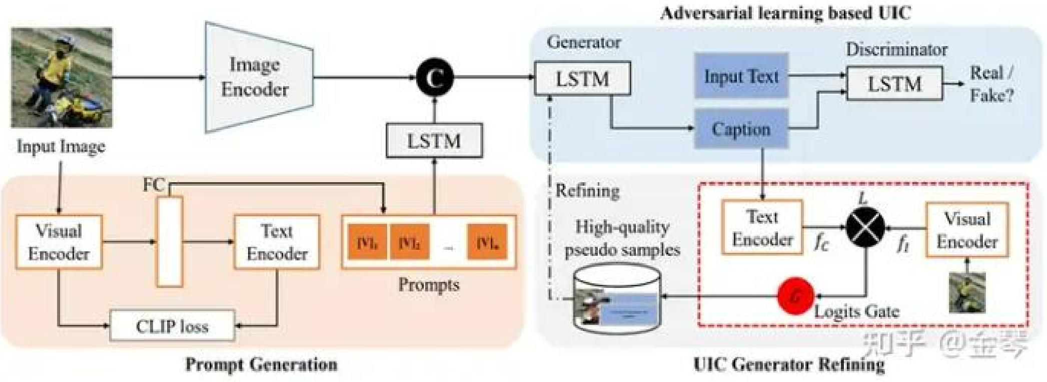


图 4

Img2Prompt[14]利用现有的图片描述模型将针对图片样本生成的描述（包括自提问与自回答）输入到语言模型中加强模态之间的理解（即下图中最右边的做法）。与之相对比的是，如下图所示，最左边的是训练一个额外的模块对齐视觉和语言向量，但是这种方法计算资源大并且有灾难性的遗忘性。如下图中间所示，直接采用自然语言作为图像的中间表示，不再需要昂贵的预训练，不需要将图片向量化表示。但是当没有样本时，其性能会显著下降。

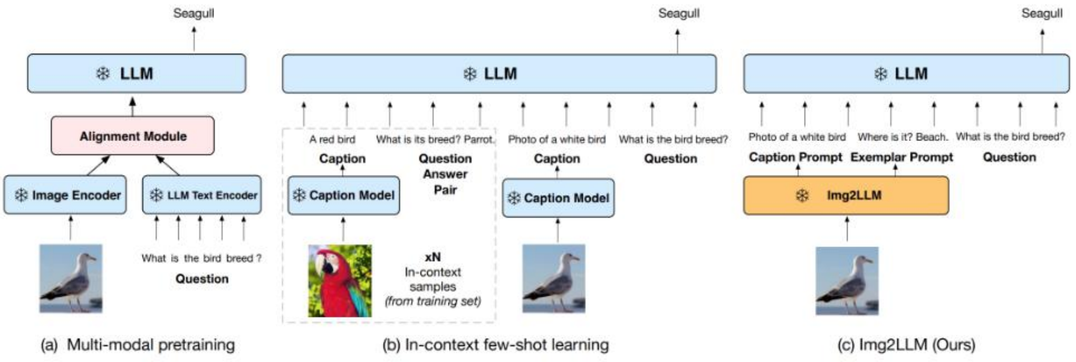


Figure 1. The illustrative comparison of three tyepes of methods that enable LLM to perform VQA tasks, where blue block denotes that the the inner parameters are frozen while pink block indicates the inner parameters are trainable.

图 5

2.3 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 1（3）所示，KgCoOp[15]在除了图文匹配的优化损失之外，还设计了对应的文本与外部知识之间的相似度损失，使得提示向量与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。以 KgCoOp 为例，如图所示，将手工提示和类别输入 CLIP，将可学习上下文和类别输入 CoOp 文本编码器，将图像输入 CoOp 图像编码器 CoOp

文本特征与图像特征计算相似概率，得到 p，p 与真实值 y 计算 Lce 损失。zero-shot CLIP 文本特征与 CoOp 文本特征计算 L2 范式，得到 Lkg 损失。将 Lce 和 Lkg 反传回去，最终的损失表达式是 Loss = Lce + Lkg，更新可学习参数。

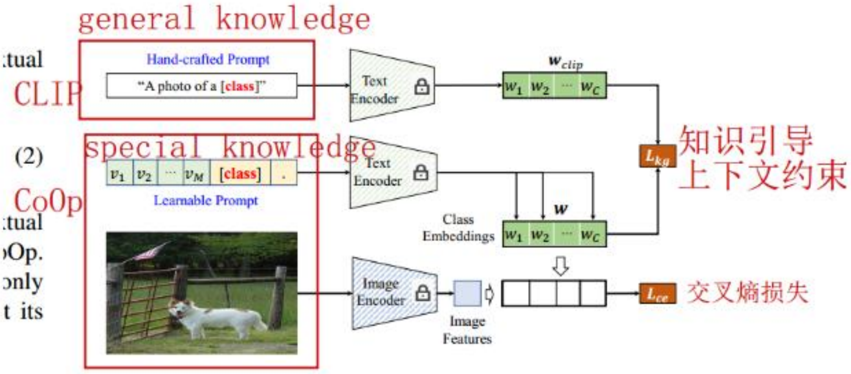


图 6

2. 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 1（4）所示 。UPT 设计了视觉和文本统一的提示向量。

MaPLe[16]、CAVPT、MetaPrompt、P3OVD 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，解决了原有的“独立的 V-L 提示”缺乏协同作用的问题。此外，Yang 等人[17]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。

以 MaPLe 为例，下图中，浅紫色部分为文本侧模型，浅绿色部分为图像侧模型，中间暗粉色部分为提示向量，通过耦合函数在语言提示上明确地设置视觉提示，以建立它们之间的交互。在学习过程中，只有暗粉色的提示向量部分参数会被调整，另外两侧的参数都是冻结的。用这种方法可以实现文本测和视觉侧信息的共享，准确来说是从视觉侧流向文本侧。

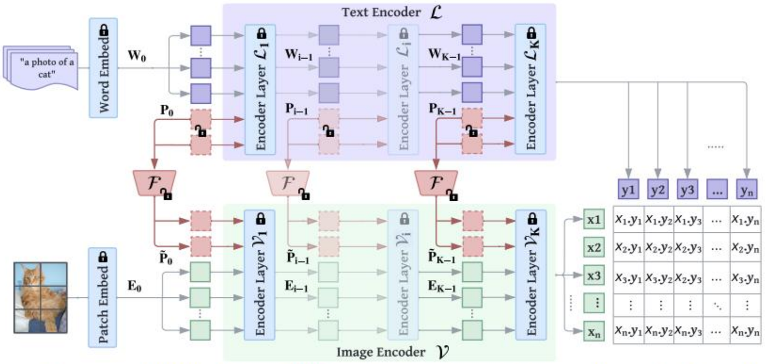

Text Encoder $\mathcal{L}$
"a photo of a cat"
Word Embed
W₀
P₀
Encoder Layer $\mathcal{L}_1$
W_{i−1}
P_{i−1}
Encoder Layer $\mathcal{L}_i$
W_{K−1}
P_{K−1}
Encoder Layer $\mathcal{L}_K$
$\mathcal{F}$
y1 y2 y3 ... yn
x1 x2 x3 ⋮ xn
Patch Embed
E₀
$\tilde{P}_0$
Encoder Layer $\mathcal{V}_1$
$\tilde{P}_{i-1}$
E_{i−1}
Encoder Layer $\mathcal{V}_i$
$\tilde{P}_{K-1}$
E_{K−1}
Encoder Layer $\mathcal{V}_K$
Image Encoder $\mathcal{V}$

Figure 2. Overview of our proposed MaPLe (**M**ulti-mod**a**l **P**rompt **Le**arning) framework for prompt learning in V-L models. MaPLe tunes both vision and language branches where only the context prompts are learned, while the rest of the model is frozen. MaPLe conditions the vision prompts on language prompts via a V-L coupling function $\mathcal{F}$ to induce mutual synergy between the two modalities. Our framework uses deep contextual prompting where separate context prompts are learned across multiple transformer blocks.

图 7

2.5 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特征属性丰富的数据表征能力不够的问题。为了解决这个问题，如图 1（5）所示，PTP[18]由图像和图像原型之间的相似性决定这个预测在多大程度上依赖于相应的提示原型，即给图像相似的图片分配相似的提示原型。具体过程如下图所示。

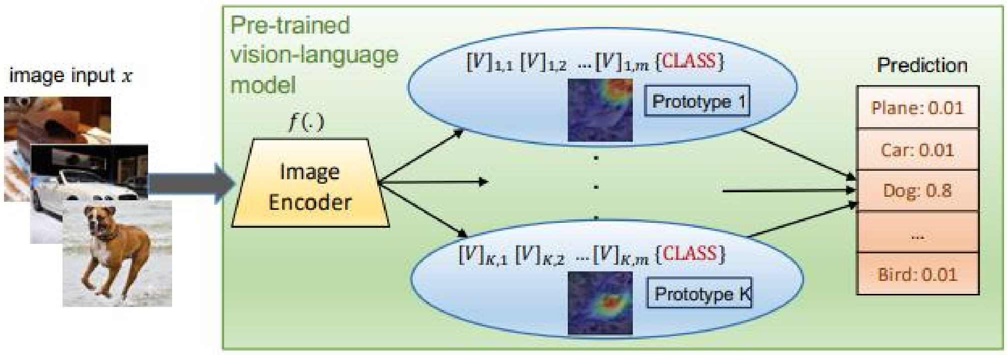


image input x
Pre-trained vision-language model
$f(.)$
Image Encoder
$[V]_{1,1}$ $[V]_{1,2}$ ...$[V]_{1,m}$ {CLASS}
Prototype 1
$[V]_{K,1}$ $[V]_{K,2}$ ...$[V]_{K,m}$ {CLASS}
Prototype K
Prediction
Plane: 0.01
Car: 0.01
Dog: 0.8
...
Bird: 0.01

Figure 1: The overall architecture of our model PTP. PTP mainly consists of: i) an image encoder, which is a part of PVLM; ii) K prototype components, where each component consists of an image prototype $\mathcal{P}_k$ and a prompt prototype $\mathcal{T}_k$; iii) a fixed PVLM. During training, we learn the lightweight parameters related to prompting, i.e., image prototypes and prompt prototypes, and keep the pre-trained vision-language model **frozen**.

图 8

Lee 等人[19]的工作与前人的区别主要在于其处理的场景更为贴近真实，允许模态的缺失。这种工作对模态齐全、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。如下图，这种提示可以放在两个地方：在输入层中，该提示向量可以串接到输入序列上，在 transformer 注意力层中该提示向量可以串接在 key 和 value 上，之后通过

控制query的长度不变而使得输出序列的长度不变。最终选择多模态transformers中与文本相关的任务标记作为最终输出特征，并将其输入池器层和全连接（FC）层进行类别预测。只有粉红色阴影区块需要训练，而其他区块则被冻结。

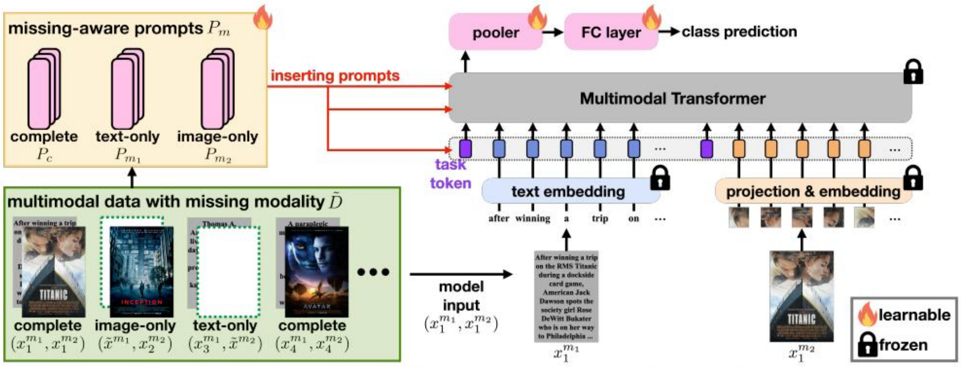

missing-aware prompts P_m
complete P_c
text-only P_{m_1}
image-only P_{m_2}
inserting prompts
pooler
FC layer
class prediction
Multimodal Transformer
task token
text embedding
projection & embedding
after winning a trip on ...
multimodal data with missing modality $\tilde{D}$
complete $(x_1^{m_1}, x_1^{m_2})$
image-only $(\tilde{x}^{m_1}, x_2^{m_2})$
text-only $(x_3^{m_1}, \tilde{x}^{m_2})$
complete $(x_4^{m_1}, x_4^{m_2})$
model input $(x_1^{m_1}, x_1^{m_2})$
$x_1^{m_1}$
$x_1^{m_2}$
learnable
frozen

Figure 2. The overview of our proposed prompt-based multimodal framework. We first select the missing-aware prompts P_m according to the missing case (e.g., complete, text-only, image-only in vision-language tasks) of the multimodal inputs $(x_i^{m_1}, x_i^{m_2})$, in which the dummy inputs $\{\tilde{x}^{m_1}, \tilde{x}^{m_2}\}$ respectively for text and image are adopted for the corresponding missing modality. Then we attach missing-aware prompts into multiple MSA layers via different prompting approaches (see Figure 3 and Section 3.3). We select the text-related task token of the multimodal transformer as our final output features, and feed them to the pooler layer and fully-connected (FC) layers for class predictions. Note that only the pink-shaded blocks require to be trained while the others are frozen.

图 9

TaI[20]设置了全局和局部的提示分别学习整图与区域的信息。如下图中G（Global）和L（Local）所示。

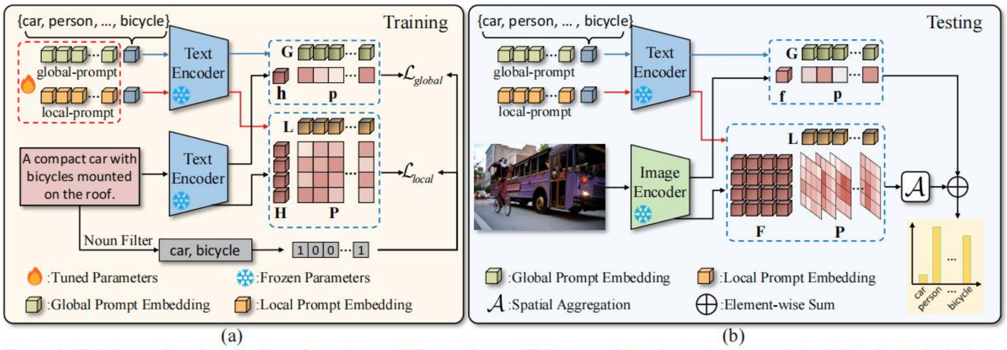


Figure 2. Training and testing pipeline of our proposed Text-as-Image (TaI) prompting, where we use text descriptions instead of labeled images to train the prompts. (a) During training, we use two identical text encoders from pre-trained CLIP to extract the global & local class embeddings (**G**&**L**) and overall & sequential text embeddings (**h**&**H**) respectively from the prompts and text description. The corresponding cosine similarity (**p**&**P**) between the embeddings are guided by the derived pseudo labels with ranking loss. (b) During testing, we replace the input from text descriptions to images. The global and local class embeddings can discriminate target classes from global & local image features (**f**&**F**). The final classification results are obtained by merging the scores of the two branches.

图 10

M-Tuning[21]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。该论文还明确指出，提示向量最好还要引入开放词。如下图(a)中的当前提示调整方法仅涉及已知类，导致(b)中的标签偏差，这表明开集和闭集数据之间的闭集最大概率分布显著重叠。M-Tuning在(c)中引入了开放词，扩展构成文本的词的范围。在(d)中减轻标签偏差后的结果表明，闭集和开集数据是明显分离的。显然引入了开放词的效果更好。

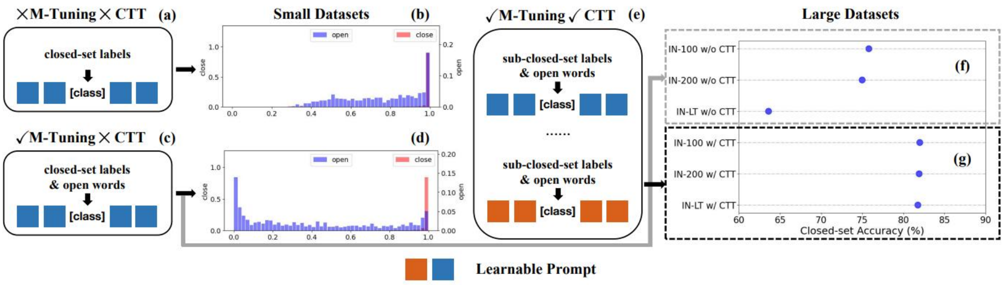
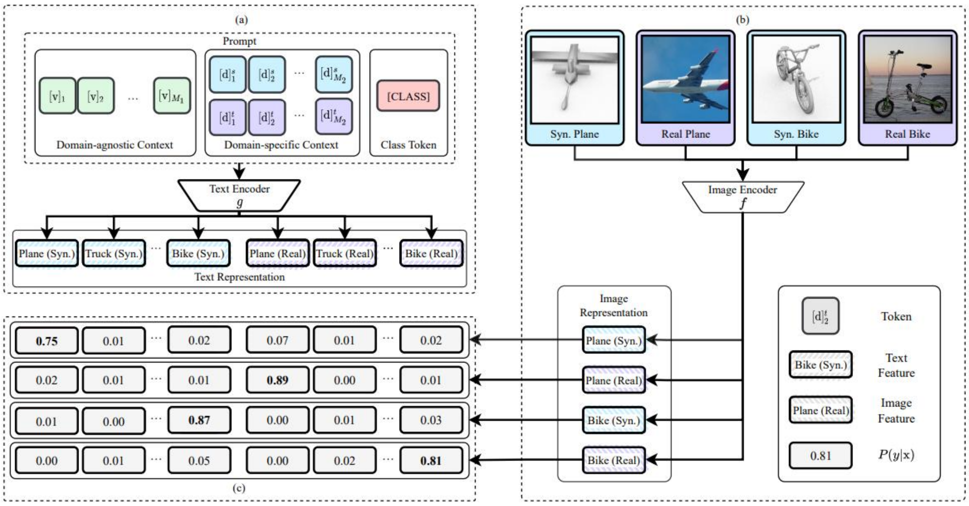


Fig. 1 Small datasets with fewer classes. Current prompt tuning methods in (a) only involve known classes, causing the label bias in (b), which shows the distributions of the closed-set maximum probability overlaps significantly between open-set and closed-set data. M-Tuning in (c) introduces open words to extend the range of words forming texts. The results after mitigating the label bias in (d) show the closed-set and open-set data are clearly separated. **Large datasets with more classes.** In (f) and (g), ‘IN’ is the abbreviation of ImageNet. Prompts represented by different colors are mutual independent. When applying M-Tuning on large datasets directly as in (c), the closed-set accuracy in (f) is low. Combining the CTT strategy in (e), M-Tuning is performed on the decomposed groups with fewer classes, contributing to higher closed-set accuracy in (g).

图 11

针对领域适应任务，DAPL[22]分别设置了领域通用、领域特定以及类别对应的提示。该工作引入了一种新的 UDA 提示学习范式，即通过提示学习进行领域适应(DAPL)。它使用了预训练的视觉语言模型，并且只优化了很少的参数。主要思想是将领域信息嵌入到提示中，这是一种由自然语言生成的表示形式，然后用于执行分类。该域信息仅由来自同一域的图像共享，从而根据每个域动态调整分类器。



0.75	0.01	···	0.02	0.07	0.01	···	0.02
0.02	0.01	···	0.01	**0.89**	0.00	···	0.01
0.01	0.00	···	**0.87**	0.00	0.01	···	0.03
0.00	0.01	···	0.05	0.00	0.02	···	**0.81**

图 12

2.6 基于分布的提示学习

除了面向特定成分的提示学习，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 1（6）所示。这种方法被称作基于分布的提示学习。

ProDA[23]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。

如下图所示。

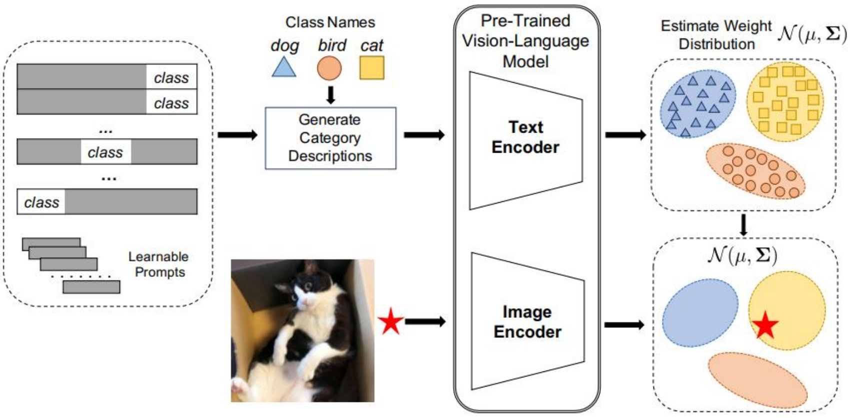

class
class
...
class
...
class
Learnable Prompts
Class Names
dog
bird
cat
Generate Category Descriptions
Pre-Trained Vision-Language Model
Text Encoder
Image Encoder
Estimate Weight Distribution
$\mathcal{N}(\mu, \Sigma)$
$\mathcal{N}(\mu, \Sigma)$

Figure 3. **Overview of the architecture of ProDA.** The class names and various learnable prompts are integrated to generate diverse category descriptions on the downstream recognition task. The output embeddings of these descriptions as the weights of linear classifiers are used to estimate the weight distribution. Given the weight distribution, we can minimize the empirical classification error and predict the classes of test samples.

图 13

PLOT[24]和 ZegOT 将视觉特征和提示对应的文本特征看成两个离散的分布，通过最优运输来实现跨模态的匹配。

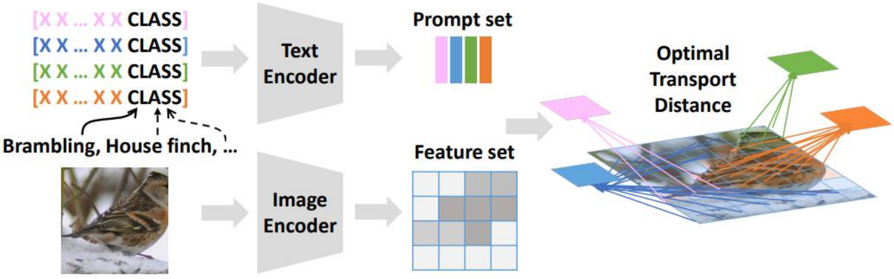


Figure 2: **The framework:** **PLOT** first describes each category with multiple prompts and obtains a set of prompt features by text encoder. The image is also encoded as a set of local features. Then the optimal transport is used as the metric between prompts and visual features.

图 14

2.7 多任务共享的提示学习

以上方法针对不同任务分别设计不同的提示，因此任务之间的相关性被忽略了，这样信息没有充分得到共享。为了增强不同任务间信息的关联性，如图 1（7）所示，Shen 等人[25]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。这样就实现了从多个任务中抽取信息并应用到单个目标任务的提示上。

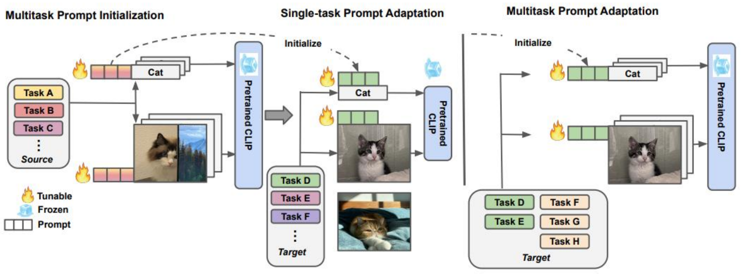


Figure 2. An illustration of our *multitask prompt initialization* (left) and *multitask prompt adaptation* (right) approaches in MVLPT. **Left**: We learn single generic source prompt vector on various *source tasks*, which is then used to initialize the prompt for each single *target task*. **Right**: After use the source prompt vector for initialization. We group relevant *target tasks* together and perform multitask prompt tuning within each group. Noted that grouping one task means single-task adaptation. (see Section 3 for details).

图 15

SoftCPT[26]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。具体来说，有一个任务共享元网络，利用任务名称和可学习的任务上下文作为输入，为每个任务生成提示上下文。该元网络的参数和任务上下文在所有任务的联合训练集上进行调整。因此，所有任务的提示语境都将以软方式共享。下图比较清晰地解释了这个过程。

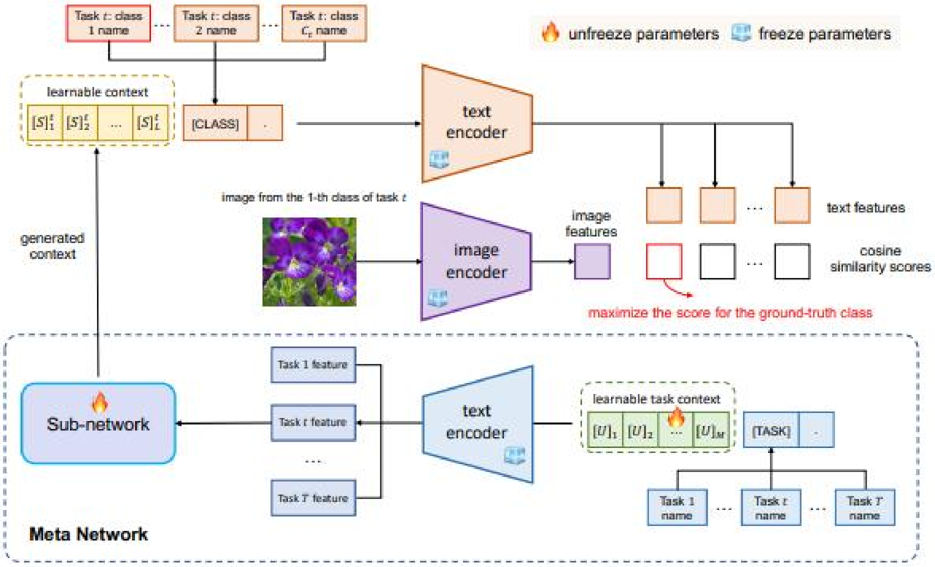


图 16

2.8 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，一种解决方案是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。如图 1（8）所示。

ProGrad[27]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。这个想法也是比较易于理解的，下图较好地显示了这个过程。

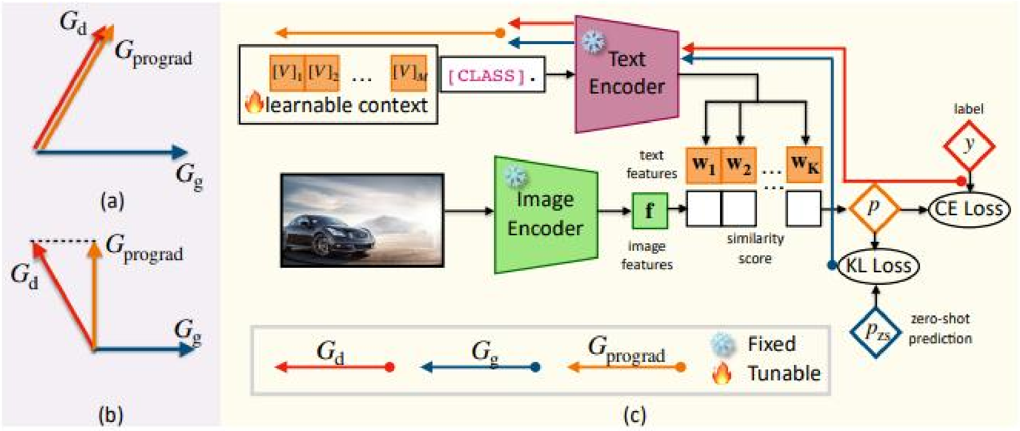


Figure 3: (a) If G_d is aligned with G_g, we set $G_{prograd}$ as G_d. (b) If G_d conflicts with G_g (*i.e.*, their angle is larger than 90°), we set $G_{prograd}$ as the projection of G_d on the orthogonal direction of G_g. (c) Training pipeline of our ProGrad. Only the context vectors are learnable.

图 17

Ma 等人[28]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，其提出子空间提示调谐(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。该工作还提出新颖特征学习器(NFL)来增强泛化能力，该部件无需任何图像训练数据。具体来说，NFL 鼓励在 CoOp 和零样本 CLIP 之间新类别的产生，以此增加泛化性。

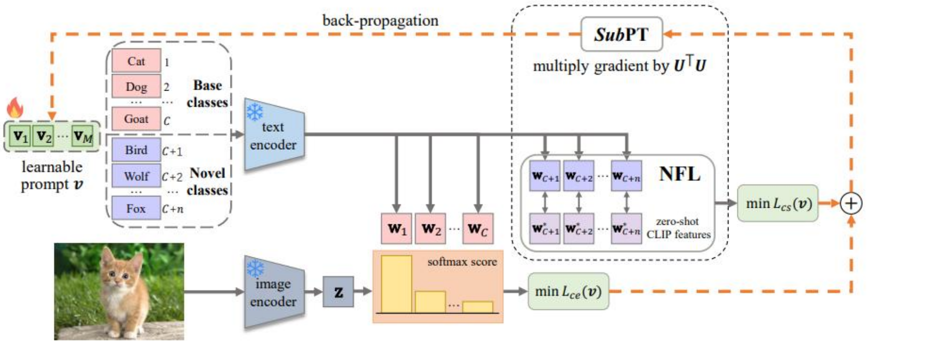


Fig. 3. **Overview of the proposed Subspace Prompt Tuning (*Sub*PT) and Novel Feature Learner (NFL)**, surrounded by the black dotted box. To eliminate spurious components and mitigate overfitting, we project the gradient onto the low-rank **subspace** spanned by the dominant eigenvectors U of early-stage gradient flow during back-propagation. **NFL** learns text features towards zero-shot CLIP features on novel categories to enhance the generalization ability of the learned prompt embedding beyond the training set.

图 18

2.9 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 1（9）所示，Huang 等人[29]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，使用多个视觉模型 CLip 来进行对比学习，按照 top-k 策略挑选伪标签，从而给无标签的

数据打上伪标签。之后参考 CoOp[30]基于伪标签在连续空间上进行提示调优。

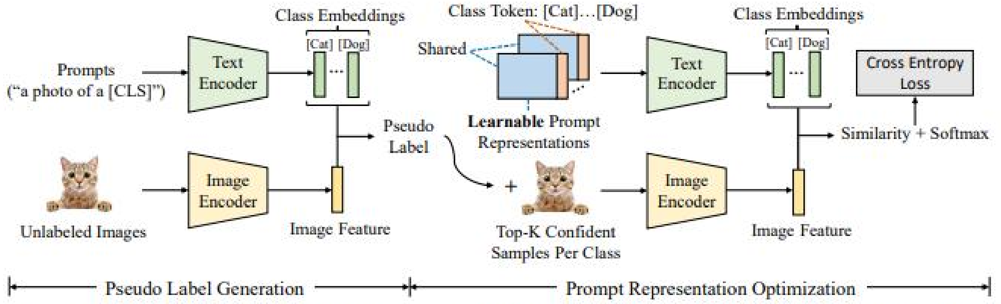


Figure 2: Overview of the proposed unsupervised prompt learning (UPL) framework. Our UPL mainly contains two parts, namely pseudo label generation and prompt representation optimization. We first use CLIP with a simple prompt (e.g., "a photo of a [CLS]") to generate pseudo labels for target datasets and select top-K confident samples per class for subsequent training. Then we define a learnable prompt representation which is optimized on selected pseudo-labeled samples. For inference, we simply swap out the hand-crafted prompts with the well-optimized prompt representations.

图 19

2. 10 建立颜色与标签关系

目前主流的 vision-language 任务，基本上服从 pre-train 和 fine-tuning 的框架。先在大型 vision-language 数据对上进行预训练学习，然后在下游任务上进行特征的微调，以取得更好的下游任务结果。这种范式极大地推动了 vision-language 领域的发展，很多模型都取得了更好的精度。但是这种范式的主要问题是，pre-train 和下游任务的学习显得有点分离。同时，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 1（10）所示，CPT[31]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。在这个工作中，受到最近 pre-trained model 的启发，CPT 这种新的范式进行 vision-language 预训练模型的微调。关键点就在于：添加 color-based co-referential markers in both image and text，visual grounding 可以重新定义为“完形填空”问题，最大程度上缩小 pre-training 和 fine-tuning 之间的差异。

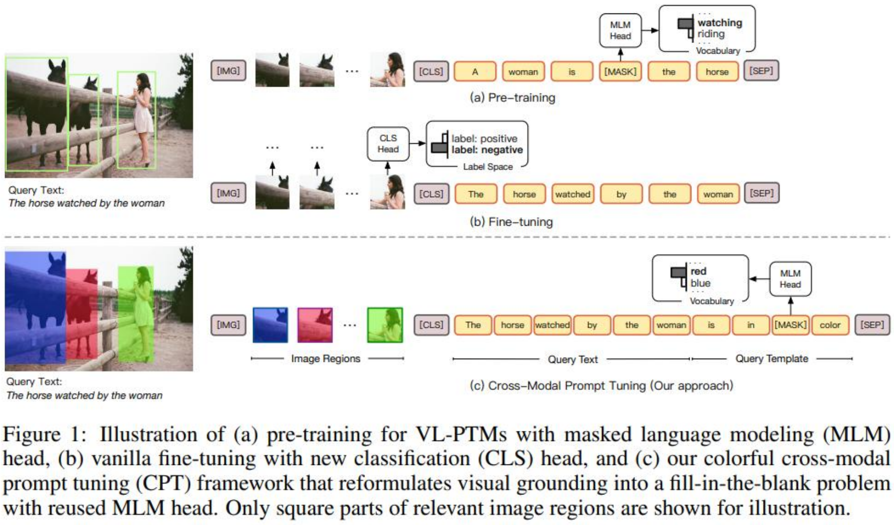
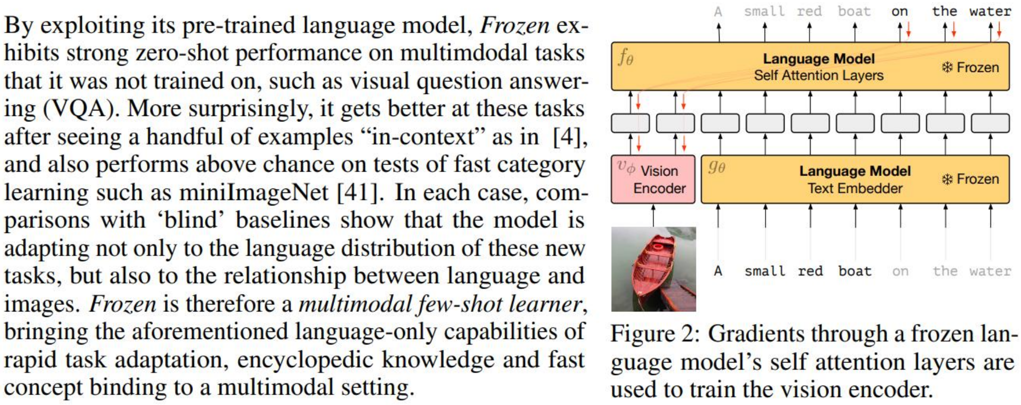


Figure 1: Illustration of (a) pre-training for VL-PTMs with masked language modeling (MLM) head, (b) vanilla fine-tuning with new classification (CLS) head, and (c) our colorful cross-modal prompt tuning (CPT) framework that reformulates visual grounding into a fill-in-the-blank problem with reused MLM head. Only square parts of relevant image regions are shown for illustration.

图 20

2.11 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 1（11）所示，Tsimpoukelli 等人[32]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是可训练的。



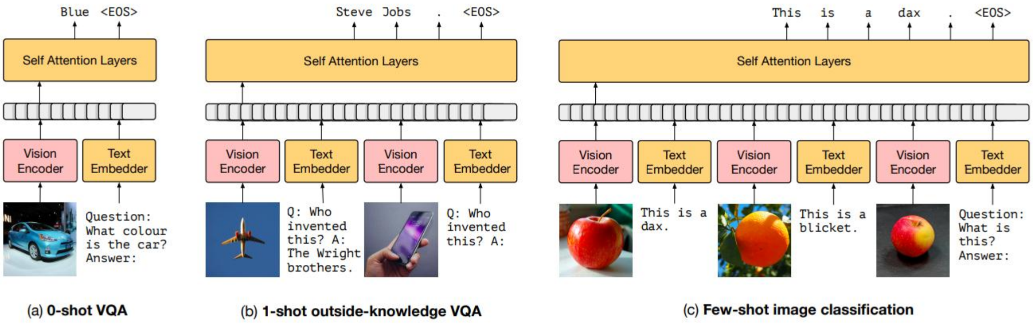
A small red boat on the water
f_θ
Language Model
Self Attention Layers
❄ Frozen
v_ϕ Vision Encoder
g_θ
Language Model
Text Embedder
❄ Frozen
A small red boat on the water

By exploiting its pre-trained language model, *Frozen* exhibits strong zero-shot performance on multimdodal tasks that it was not trained on, such as visual question answering (VQA). More surprisingly, it gets better at these tasks after seeing a handful of examples "in-context" as in [4], and also performs above chance on tests of fast category learning such as miniImageNet [41]. In each case, comparisons with 'blind' baselines show that the model is adapting not only to the language distribution of these new tasks, but also to the relationship between language and images. *Frozen* is therefore a *multimodal few-shot learner*, bringing the aforementioned language-only capabilities of rapid task adaptation, encyclopedic knowledge and fast concept binding to a multimodal setting.

Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

图 21

该工作的作者表示，他们开发 frozen 并不是为了最大限度地提高任何特定任务的性能，而是更泛化地处理之前没有见过的数据并给出令人信服的结果。期望达到的效果如下。



(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

图 22

3 学习心得与思考

经过对于文献综述的阅读，并以此为索引对于视觉语言多模态领域中的提示学习进行了较为广泛与深入的调研。我收获很多，大概有以下几点：知识、方法与乐趣。我对于该领域的提示学习方法有了一个较为全面的了解，并对其中一些较感兴趣的方法进行了深入思考。我不仅学习到了知识本身，更从一个方法的提出背景等，看到了这个方法为什么会被提出，被提出解决了什么问题，以及为什么这个方法能成功，它的好处在哪里。

这些创新性的提示学习方法，目的主要是增强视觉语言模型的性能和泛化能力。通过阅读论文，我发现这些方法体现了研究者对提示学习的深入理解和创新思路。梯度引导、子空间调优、无监督学习等技术手段,显示了作者对优化过程的细致分析和对泛化性提升的独到见解。我不禁感叹，这些才是推动该领域进步的关键所在。

创造性的方法针对性地解决了视觉语言模型的一些痛点,如过度拟合、特征学习不足、标注数据缺乏等。通过独特的设计,有望在一定程度上缩小预训练和微调之间的性能差距,提高模型在新场景下的适应性。这对于实际应用来说无疑是一大进步。

完整阅读完感兴趣的工作后，我感叹它们体现了研究者对模型内部机制和跨模态对应关系的深入思考。从梯度分解到特征映射再到颜色标记,都显示了他们对模型内部工作原理的精准把握。这种深入理解为未来的模型优化和拓展提供了坚实的基础。

总的来说，这些论文中提出的创新性提示学习方法,在理论和应用价值上都令人印象深刻。它不仅推动了视觉语言模型的性能提升,也为该领域的发展提供了新的思路和方向。作为一个初学者，我很高兴能阅读到该领域的这些前人工作，很高兴能学习到这些知识。我有自信，在不远的将来，我能够在更深入理解的基础上，投身于继续创造能够提高性能、提高泛化性、在单个任务或多个任务上有更好表现的方法。

参考文献

[1] 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.

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

[3] 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.

[4] 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.

[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] 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.

[8]Liao N, Cao M, Yan J C. Visual Prompt Learning: A Survey, 2024(4). 廖宁，曹敏,严骏驰. 视觉提示学习综述.[J];计算机学报;2024 年 04 期

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Yang H, Lin J, Yang A, et al. Prompt Tuning for Generative Multimodal Pretrained Models. arXiv preprint arXiv:2208.02532, 2022.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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

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

[30] 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.

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

[32] 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.