基于 Adapter 和多模板提示的 MaPLe 改进研究

王思杰

2024 年 12 月 5 日

摘要

MaPLe 综合考虑了文本和图像提示，取得了较好的改进效果。本文在此研究的基础上，提出了一种改进的 MaPLe 模型，主要是结合了 Adapter 和多种文本提示模板。首先，引入 Adapter 机制，通过插入小型、可训练的模块，以在不修改大量参数的情况下，使模型适应特定任务。其次，通过使用多种提示模板来增加文本输入的多样性，从而增强模型的鲁棒性和泛化能力。实验结果表明，改进后的 MaPLe 模型在多个下游任务中实现了显著的性能提升。

关键词：多模态；多模板提示；Adapter

1 引言

在当今人工智能技术的浪潮中，多模态学习以其强大的信息融合能力成为研究的前沿阵地。MaPLe 模型 [2] 作为多模态学习的杰出代表，通过整合文本和图像提示，在多个任务中展现了卓越的性能。然而，面对实际应用中模型参数微调复杂度高、对特定任务适应性不强以及输入多样性不足等挑战，如何进一步提升 MaPLe 模型的泛化能力和鲁棒性，成为当前亟待解决的问题。

在此基础上，本研究引入了 Adapter 机制 [1, 4]，这是一种高效的模型微调方法，通过插入小型、可训练的模块到预训练模型中，以在不修改大量参数的情况下，使模型适应特定任务。同时，为了增强模型的鲁棒性和泛化能力，本研究还设计了多种文本提示模板，以增加文本输入的多样性。这一创新性的改进思路，不仅为 MaPLe 模型的优化提供了新的方向，也为多模态学习领域的研究注入了新的活力。

从理论和实践两个层面来看，本研究具有重要意义。在理论层面，通过引入 Adapter 机制和多样性提示，本研究丰富了多模态学习的研究方法，为模型的高效微调和多样性提示提供了新的理论支撑。在实践层面，改进后的 MaPLe 模型在多个下游任务中实现了显著的性能提升，为图像描述生成、跨模态检索等实际应用提供了更为高效和鲁棒的解决方案。此外，本研究还为其他相关模型的改进提供了有益的参考和借鉴，推动了多模态学习领域的研究进展，拓展了人工智能技术的应用边界。

2 相关工作

自 CLIP 模型 [3] 在 2021 年被提出以后，视觉语言模型的相关研究迅速发展，并出现一系列相关的改进与应用策略。本节主要对视觉-语言模型（Vision-Language Model，VL 模型），提示学习（Prompt Learning）及其在 VL 模型中的应用进行介绍。

2.1 视觉-语言模型

视觉语言模型结合了自然语言监督与自然图像，这一结合方式在计算机视觉社区中引起了极大的兴趣。与传统的仅依赖图像监督学习的模型相比，视觉语言模型能够编码丰富的多模态表示，即同时处理和理解图像与文本信息。这种多模态表示能力使得视觉-语言模型在处理复杂视觉任务时具有更高的灵活性和准确性。近年来，一系列视觉-语言模型如 CLIP、ALIGN、LiT、FILIP 和 Florence 等 [5]，在包括小样本学习和零样本学习在内的广泛视觉识别任务上展示了卓越的性能。这些模型利用互联网上大量可用的图像-文本对数据，以自监督的方式学习图像与语言的联合表示。例如，CLIP 和 ALIGN 分别使用了约 4 亿和约 10 亿的图像-文本对来训练其多模态网络，从而获得了强大的泛化能力。

尽管这些预训练的视觉-语言模型已经学习了泛化的表示，但如何高效地将它们适应到下游任务中仍然是一个具有挑战性的问题。为了解决这个问题，许多研究提出了针对视觉-语言模型的定制方法，以在少样本图像识别、目标检测和分割等下游任务上实现更好的性能。这些方法通过调整视觉-语言模型的参数或结构，使其能够更好地适应特定任务的需求。MaPLe 模型提出了一种新颖的多模态提示学习技术，以有效地将 CLIP 模型适应到少样本和零样本视觉识别任务中。这种技术通过设计特定的文本提示，引导 CLIP 模型在有限的训练样本下，仍然能够准确地理解和识别图像内容。这种方法的优势在于，它不需要对 CLIP 模型进行大量的参数调整，而是通过巧妙地利用文本提示来激发模型的潜在能力，从而实现高效的模型适应。

2.2 提示学习

提示学习是一种技术，其中，以句子形式给出的指令，被称为文本提示（text prompt），通常被给予 V-L 模型的语言分支。这些文本提示旨在帮助模型更好地理解任务，从而提高其性能。在 V-L 模型中，语言分支负责处理和理解文本信息，而视觉分支则负责处理和理解图像信息。通过向语言分支提供明确的文本提示，模型可以更容易地将图像与相关的文本信息关联起来，从而更准确地完成任务。这些文本提示可以是为特定下游任务手工制作的，也可以在微调阶段自动学习得到。当提示是在微调过程中自动学习时，这个过程被称为“提示学习”。提示学习最初在自然语言处理（NLP）领域得到应用，随后被扩展到 V-L 模型和仅视觉模型中。在 NLP 领域，提示学习已经被证明是一种有效的方法，可以提高模型在各种任务上的性能。随着 V-L 模型的兴起，提示学习也被引入到这一领域。通过向 V-L 模型提供明确的文本提示，研究人员可以引导模型更好地理解图像和文本之间的关系，从而提高模型在视觉识别等任务上的性能。

类似地，在计算机视觉领域，也有研究开始探索仅视觉模型的提示学习。这些研究通常通过向模型提供额外的视觉信息或上下文来帮助模型更好地理解图像内容。MaPLe 模型提出了一种新的设计，它使用了深层的“视觉”提示。这种设计不仅利用了文本提示来引导模型，还

通过深层的视觉处理来增强模型对图像的理解。然而，与之前的仅视觉提示设计不同，MaPLe 模型的设计是首个多模态提示设计。这意味着 MaPLe 模型的设计不仅考虑了文本提示，还结合了视觉提示，从而允许模型同时处理和理解图像与文本信息。

2.3 提示学习在 VL 模型的应用

在视觉语言模型（V-L 模型）中，特别是在 CLIP 这样的模型上，全面微调（Full Finetuning）和线性探测（Linear Probing）是两种常见的下游任务适应方法。全面微调涉及调整模型的所有参数，虽然这可以提高模型在特定任务上的性能，但往往会破坏模型之前学习到的视觉和语言联合表示。另一方面，线性探测仅调整模型输出层的参数，旨在保持大部分预训练知识的同时进行任务适应，但这种方法的局限性在于它限制了 CLIP 模型的零样本能力。为了克服这些限制，受自然语言处理（NLP）中提示学习（Prompt Learning）的启发，研究者们开始探索在 V-L 模型中应用提示学习的方法。这些方法通过在模型的语言或视觉分支上学习提示令牌（Prompt Tokens），以端到端的方式调整模型，从而实现对新任务的适应。例如，CoOp 方法 [7] 通过优化语言分支上的一组连续提示向量，实现了对 CLIP 的少样本迁移学习。而 Co-CoOp 方法 [6] 则解决了 CoOp 在新类别上表现不佳的问题，通过显式地将提示与图像实例关联起来，提高了模型的泛化能力。

然而，现有的提示学习方法大多遵循独立的单模态解决方案，要么在语言分支上学习提示，要么在视觉分支上学习提示，因此只是部分地适应了 CLIP 模型。这引发了一个重要的问题：即鉴于 CLIP 的多模态性质，是否全面提示（即在语言和视觉分支上都进行提示）更适合适应 CLIP。为了回答这个问题，MaPLe 模型首次研究了多模态提示学习的有效性。通过探索在语言和视觉分支上同时学习提示的方法，MaPLe 模型旨在改善视觉和语言表示之间的对齐，从而提高模型在下游任务上的性能。这种方法不仅保留了模型的大部分预训练知识，还通过全面提示的方式实现了对新任务的更好适应。

3 本文方法

3.1 本文方法概述

本文将要复现的论文为 MaPLe 模型。其关键的模型结构图如图1所示，图1展示了多模态提示学习（MaPLe）与标准提示学习方法的对比。

如图1(a) 所示，现有方法主要采用单模态提示技术来微调 CLIP 模型的表示。这些技术的一个显著特点是，它们只在 CLIP 模型的一个分支（语言分支或视觉分支）中学习提示。这种方法虽然在一定程度上能够提高模型在特定任务上的性能，但由于忽略了 CLIP 模型的多模态特性，即文本和图像表示之间的相互作用，因此其泛化能力可能受到限制。

如图1(b) 所示，与此不同，MaPLe 方法引入了分支感知的层次提示。这种方法不仅考虑到了 CLIP 模型的语言分支，还同时考虑了视觉分支，通过同时适应这两个分支来改进模型的泛化能力。具体来说，MaPLe 学习了一组层次化的提示，这些提示能够捕捉到文本和图像表示之间的复杂关系，并在微调过程中保持这种关系。这种方法使得 CLIP 模型能够更好地适应新的任务和数据集，从而在多种场景下表现出更好的性能。

如图1(c) 所示，该部分展示了 MaPLe 在 11 个不同的图像识别数据集上对于新颖类别泛

化任务的表现。实验结果表明，MaPLe 方法显著超越了当前最先进的方法。这意味着，当面临未见过的类别时，MaPLe 能够更好地利用 CLIP 模型的多模态特性来进行预测，从而实现了更高的准确率。这一结果进一步证明了 MaPLe 方法的有效性和实用性，为未来的视觉语言模型研究提供了新的思路和方法。

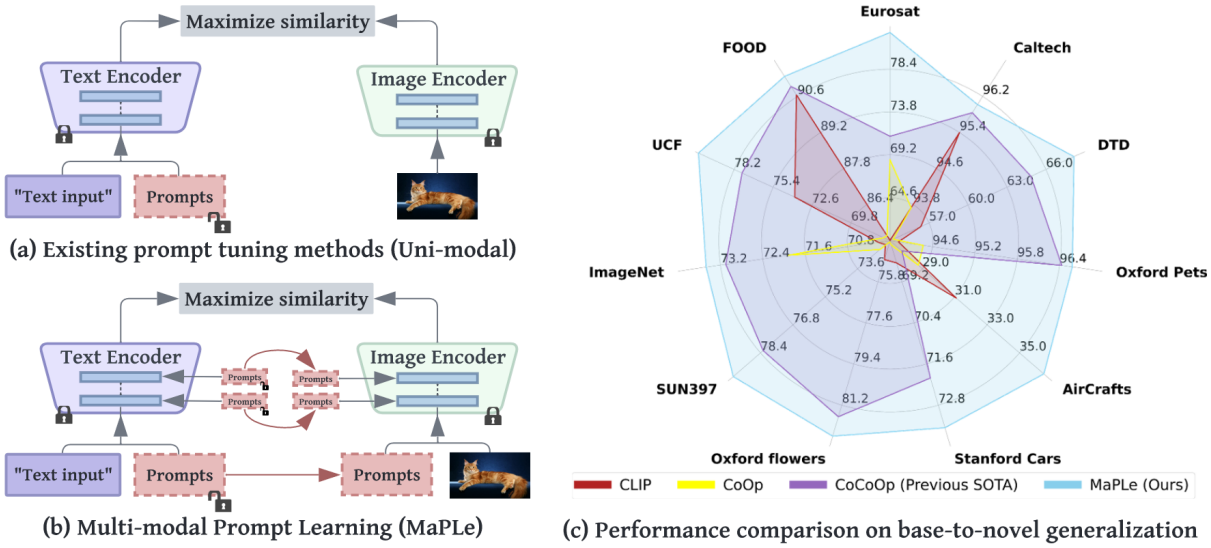

图 1. MaPLe 模型关键结构

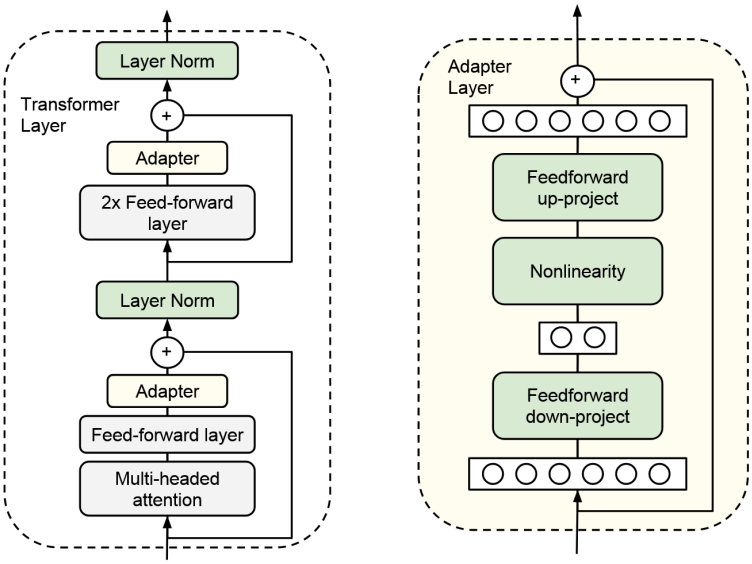

图 2. Adapter 的具体结构

3.2 Adapter

Adapter（Low-Rank Adaptation）是一种针对大型语言模型（LLMs）的参数高效微调方法，其具体结构如图2所示。这种方法的核心思想是通过在预训练模型的中间层插入小型、可训练的模块（Adapter），在不大规模调整原始模型参数的情况下，使模型能够高效地适应特定任务。Adapter 模块仅微调其自身的参数，而固定预训练模型的其他部分，从而显著减少需要更新的参数量，同时保持预训练模型的强大表示能力。这种方法能够提升迁移学习的效率，尤其在多个任务或计算资源有限的情况下，具有明显的优势。

本研究的引入的 Adapter 模块是文本编码器和图像编码器之间的共享 Adapter，如图3所示，其中包括专门针对不同模态编码器进行调优的两个独立投影层（“Down” 和 “Up”），以及一个旨在加强视觉和语言分支之间紧密联系的共享投影层（“Shared”）。这一设计不仅允许系统对各个模态的信息进行精细化的处理，还促进了跨模态信息的有效整合与交互，从而显著提升了多模态信息处理的效率和准确性。

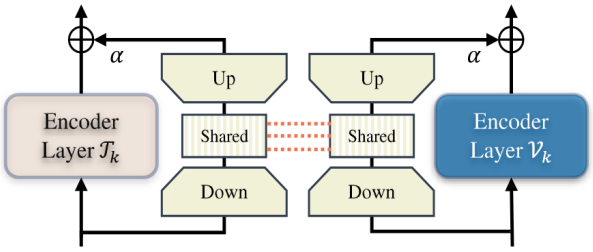

图 3. 跨模态 Adapter 原理示意图

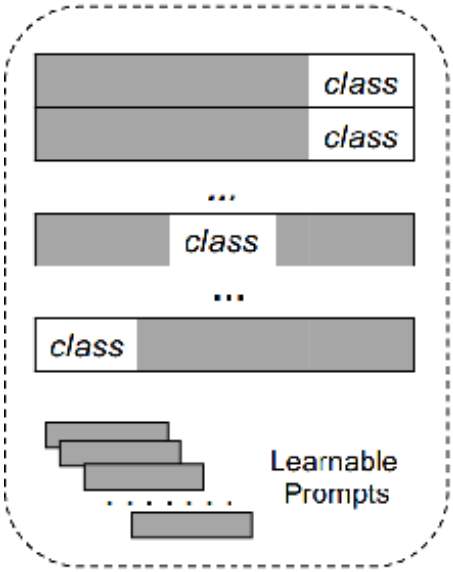

图 4. 多种提示模板示意图

3.3 多种模板提示

提示学习是一种先进的微调方法，其核心在于利用模板提示来引导模型理解并完成任务，从而提升性能。多种模板提示在此过程中的必要性和作用不容忽视。首先，从任务适应性的角度来看，不同的下游任务往往要求不同的输入格式和输出形式。通过设计多样化的模板提示，可以将这些任务需求转化为预训练模型能够理解并处理的形式，从而增强模型的适应能力。这种灵活性使得提示学习能够广泛应用于各种自然语言处理任务中。其次，模板提示在知识引导方面发挥着重要作用。它们包含了与任务相关的关键信息和指令，这些信息能够引导模型在推理过程中有效利用预训练阶段学到的知识。通过这种方式，模型能够更准确地理解任务要求，并作出正确的决策。此外，多种模板提示还有助于减少多任务学习中的负迁移现象。在多任务场景中，不同任务之间可能存在差异和冲突，这可能导致模型在训练过程中产生混淆。通过为每个任务设计独立的模板提示，可以降低任务间的相互干扰，从而提高模型的稳定性和准确性。更重要的是，多种模板提示的引入还增强了模型的可解释性。模板提

示通常包含人类可理解的关键词、短语或句子，这使得模型的决策过程更加透明和易于理解。这有助于研究人员更好地分析模型的性能，并进行针对性的优化和改进。最后，多样化的模板提示还支持了模型的泛化能力。通过引入不同形式的模板提示，模型能够学习到更通用的特征表示，从而增强其在未知任务或数据集上的适应能力。这种泛化能力对于应对不断变化的自然语言处理需求至关重要。

综上所述，多种模板提示在提示学习中发挥着不可或缺的作用。它们不仅提高了模型的性能和可解释性，还支持了多样化的任务需求，并促进了模型的泛化能力。因此，在设计提示学习方案时，应充分考虑模板提示的多样性和有效性，以充分发挥其潜力，这里介绍的多个提示如图所示。

在本研究中，考虑到图像分支的提示是由文本分支的提示通过线性变换得到的，因此可以仅考虑增加文本端提示的多样性，以使得模型性能进一步提升。具体来说，就是将文本端的输入取平均。

4 复现细节

4.1 与已有开源代码对比

本研究使用了论文 MaPLe 在 GitHub 上开源的代码，作为基准，并在此基础上引入了 Adapter 机制，在输入端引入了多种提示模板以增加提示的多样性。

4.2 实验条件

为了进行实验，需确保以下配置：

- Python 版本：Python 3.9 需被安装并配置在系统上。
- PyTorch 版本：需安装与 CUDA 11.x 兼容的 PyTorch 1.12 版本（确保版本与所安装的 CUDA Toolkit 相匹配）。
- GPU 要求：系统中需配备 NVIDIA V100 GPU，并正确安装 NVIDIA 驱动程序及 CUDA Toolkit。

完成上述配置后，可通过 nvidia-smi 命令检查 GPU 状态，并在 Python 环境中验证 PyTorch 是否成功检测到 GPU。

4.3 创新点

如上文提及的内容所示，本研究的创新点主要在于 Adapter 机制的引入以及多模板提示的使用，这两个创新点不仅丰富了模型优化与训练的理论与实践，更预期对 MaPLe 模型产生显著的改进作用。

首先，Adapter (Low-Rank Adaptation) 机制的引入，是本研究的一大创新亮点。Adapter 机制通过引入小型、可训练的模块，使得模型在微调过程中仅需调整少量的参数，即可实现对预训练模型的有效适配。这种方法的优势在于，它能够在不改变预训练模型原始参数的情况下，实现对模型性能的定制化提升。相较于传统的全参数微调方法，Adapter 机制极大地降

低了计算资源和时间的消耗，同时保持了模型性能的稳定性。这一创新点的引入，预期将使得 MaPLe 模型在面对多样化任务时，能够更加高效地进行模型适配，从而提升模型的灵活性和泛化能力。

其次，多模板提示的使用，是本研究另一个重要的创新点。在自然语言处理任务中，模板提示被广泛应用于指导模型生成符合特定格式或风格的输出。然而，传统方法往往采用单一的模板提示，这限制了模型的生成能力和适应性。本研究通过引入多模板提示，为模型提供了更多样化的输入指导，从而增强了模型的生成能力和灵活性。多模板提示的使用，预期将使 MaPLe 模型在生成文本时，能够更准确地捕捉和理解输入信息的多样性，从而生成更加准确、自然和符合预期的文本输出。

这两个创新点的结合，预期将对 MaPLe 模型产生显著的改进作用。一方面，Adapter 机制的引入将降低模型微调的门槛，使得 MaPLe 模型能够更加高效地进行定制化适配；另一方面，多模板提示的使用将提升模型的生成能力和灵活性，使得 MaPLe 模型在面对多样化任务时，能够生成更加准确、自然和符合预期的文本输出。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

本研究选用的数据集有 OxfordPets、Food101 和 StanfordCars 三个数据集，以下是模型在这几个数据集上的实验结果。

表1显示了各模型在 OxfordPets 数据集上的泛化能力，通过对实验结果的分析可以看出，Adapter 和 MTT 的引入都能在 MaPLe 基准模型的基础上带来性能提升，尤其是在新类别（Novel）和整体性能（HM）上。单独使用 Adapter 提升了模型在新类别识别的能力，而 MTT 的提升幅度较小。然而，Adapter 和 MTT 的结合则展现了最为显著的提升，尤其在新类别识别和整体性能上，表明两者的联合优化能够更有效地增强模型的泛化能力。总体来看，结合 Adapter 和 MTT 是提高 MaPLe 模型在 OxfordPets 数据集上泛化能力的最佳方案。

表2和表3分别分别显示了模型在 StanfordCars 和 Food101 数据集上的泛化能力，通过对实验结果的分析可以看出，在 StanfordCars 和 Food101 数据集上，Adapter 和 MTT 的引入均能提高 MaPLe 模型的性能，尤其是在 Novel 类别和整体性能（HM）上。在 StanfordCars 数据集上，Adapter 和 Adapter + MTT 组合分别提升了 0.36 和 0.42 的 Novel 性能，而 Food101 数据集上，二者的改进更为显著，尤其是 Adapter + MTT 组合，分别在 Base、Novel 和 HM 上提升了 0.96、0.73 和 0.84。总体来看，Adapter 主要提升了 Novel 类别的性能，而 MTT 则改善了整体性能，二者结合能够最大化提升模型的泛化能力。

6 总结与展望

本研究通过引入 Adapter 机制，实现了模型在微调过程中的高效性与稳定性。Adapter 通过小型、可训练的模块，大幅减少了需要调整的参数数量，从而在保持模型性能稳定的同时，降低了计算资源和时间的消耗。这一创新使得 MaPLe 模型在面对多样化任务时，能够更快速、更灵活地进行适配，进而提升了模型的泛化能力。在自然语言处理任务中，模板提示的应用至关重要。本研究通过引入多模板提示，打破了传统单一模板的限制，为模型提供了更多样

	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
Co-CoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
+ Adapter	95.67	98.13	96.89
+ MTT	95.59	97.98	96.78
+ Adapter + MTT	95.71	98.32	96.99
	+0.28	+0.56	+0.41

表 1. 各模型在 OxfordPets 数据集上的泛化能力

	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
Co-CoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
+ Adapter	73.11	74.36	73.73
+ MTT	73.06	74.12	73.59
+ Adapter + MTT	73.16	74.42	73.79
	+0.22	+0.42	+0.32

表 2. 各模型在 StanfordCars 数据集上的泛化能力

	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
Co-CoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
+ Adapter	91.58	92.63	92.10
+ MTT	91.32	92.39	91.85
+ Adapter + MTT	91.67	92.78	92.22
	+0.96	+0.73	+0.84

表 3. 各模型在 Food101 数据集上的泛化能力

化的输入指导。这一创新不仅增强了模型的生成能力，还提高了其灵活性。预期将使 MaPLe 模型在生成文本时，能够更准确地捕捉和理解输入信息的多样性，从而生成更加准确、自然和符合预期的文本输出。

尽管本研究在 Adapter 机制和多模板提示方面取得了一定的创新成果，但在实现过程中仍存在一些不足：

- Adapter 机制的实验验证：虽然 Adapter 机制在理论上具有显著优势，但本研究在实验验证方面可能还不够充分。未来需要更多实验来验证 Adapter 机制在不同任务、不同数据集上的表现，以确保其稳定性和可靠性。
- 多模板提示的多样性：虽然多模板提示的使用提高了模型的生成能力和灵活性，但如何更好地设计和利用多样化的模板提示仍需进一步研究。未来需要探索更多样化的模板设计策略，以进一步提升模型的性能。
- 模型性能评估：本研究在评估模型性能时可能还存在一些局限性。未来需要采用更多元化的评估指标和方法，以更全面、更准确地评估 MaPLe 模型在不同任务上的表现。

基于当前研究的成果和不足，未来可进一步开展以下研究方向：

- Adapter 机制的深度优化：针对 Adapter 机制的实验验证不足问题，未来可以开展更深入的优化研究。通过调整低秩分解的参数、优化适配层的结构等方式，进一步提升 Adapter 机制的性能和稳定性。
- 多模板提示的创新设计：针对多模板提示的多样性问题，未来可以探索更多创新性的模板设计策略。例如，结合自然语言生成技术自动生成多样化的模板提示，或者利用深度学习技术学习并提取多样化的模板特征等。
- 模型性能评估体系的完善：针对模型性能评估的局限性问题，未来可以构建更完善、更全面的评估体系。通过引入更多元化的评估指标和方法，更准确地评估 MaPLe 模型在不同任务上的表现，并为其后续优化提供有力支持。

参考文献

[1] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[2] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024.

[5] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.