ChatGPT

融合原型记忆与区域提示的图像字幕生成网络设计方案

模型整体架构设计

总体流程： 图像输入后，首先经过视觉特征提取，然后引入原型记忆模块生成视觉记忆原型，结合CLIP驱动的区域级文本提示信息，最后通过文本解码器生成图像字幕。整个架构在Transformer编码-解码框架下运作，以增强视觉语义对齐和描述一致性。

- **图像编码与特征提取：** 利用预训练视觉编码器（如CLIP的ViT模型或ResNet）提取图像的局部和全局特征表示。CLIP图像编码器将图像划分为若干patch并映射为向量表示，同时保留一个全局[CLS]特征[1]。这些特征既作为后续记忆模块的输入，也用于区域提示的语义映射。在CLIP预训练中，patch特征缺乏细粒度语义对齐监督，可能导致同一实体的patch语义不一致[2]。为此，我们引入FSGR方法对CLIP特征进行优化，增强patch级的语义一致性。

- **FSGR区域级文本提示融合：** 在编码阶段插入**CLIP跨模态区域提示模块**。具体而言，预先利用CLIP文本编码器构建一个**对象概念词典**（object-semantic codebook），收集多样的潜在对象和属性词的嵌入向量[3]。然后，对每个图像patch的视觉特征执行**语义量化**：将增强后的视觉patch特征映射到最近的词典概念，从而给每个视觉区域分配一个相应的文本语义标签[4]。由此生成**语言桥接的视觉图谱**，即用显式的文本嵌入来表示图像的细粒度语义内容，同时保留各区域的空间位置信息[5]。这一图谱可视作图像的区域级文本提示集合。接着，通过**对齐的跨模态交互**模块，将视觉patch与其对应的文本概念通过交叉注意力融合，聚合图像局部视觉信息与文本语义信息到统一的CLIP表示空间[6]。融合了区域文本提示的图像表征将作为解码器的条件输入，从而在解码时引导字幕生成更准确地涵盖图像中的关键语义。

- **原型记忆模块（PMA-Net记忆机制）：** 解码器在生成字幕时，除了通常对编码器图像特征执行交叉注意力外，还**接入原型记忆注意力**机制[7][8]。PMA-Net在训练过程中从过去样本的激活中提取**原型记忆向量**，存储于外部记忆库中[9]。这些原型记忆通过对过去训练样本的keys/values进行聚类和插值获得，能够浓缩训练集中常见的视觉-语言模式[10][11]。在解码每一步时，解码器的查询向量（当前时间步的隐藏状态）不仅与当前图像的编码特征交互，还通过一个额外的注意力头查询记忆库中的原型键/值对（**Memory Key/Value**）[7]。这样，模型可“回忆”训练集中类似场景下常见的描述模式，提高生成的准确性和多样性。例如，当描述滑雪场景时，记忆模块可能提供其他样本中出现过的短语如“雪山”“滑雪板”等，有助于生成更恰当的描述。原型记忆机制相当于让模型在生成过程中参考“自己的过去经验”，弥补单幅图像提供信息的不足[9]。值得注意的是，这种记忆查询仅在训练时生成原型，推理时并不会显著增加计算开销[12]。

- **解码与字幕生成：** 文本解码器采用Transformer解码架构，结合上述两类信息生成描述。一方面，解码器的跨模态注意力同时接收来自图像编码器的视觉特征和来自**区域文本提示图谱**的语义token作为Key/Value，从而将视觉和文本提示综合用于词生成。另一方面，解码器在每层还额外增加**原型记忆注意力子层**，以查询外部记忆键值对。解码时每一步的计算流程为：先自注意力聚合已有部分字幕上下文，再并行地对接收到的图像特征和文本提示执行跨注意力，然后对原型记忆执行注意力，最后经前馈网络输出新词概率。这样的双重信息源让生成的字幕既**贴近当前图像内容**，又参考了**大量训练样本的语义模式**。最终，解码器输出自然语言句子作为图像的字幕描述。

- **协同训练与参数策略：** 为成功融合PMA-Net和FSGR模块，我们采用分阶段和联合相结合的训练策略。首先，对CLIP图像编码器进行**局部语义一致性优化**：利用FSGR提出的patch级对比学习目标细调CLIP视觉分支参数（冻结大部分CLIP权重，仅添加少量可学习参数，实现高效调整[13][14]），使同一语义实

体的patch特征在CLIP空间更聚集，从而使后续语义分配更准确 [15] [16]。接着，构建对象概念词典，并在初始阶段将CLIP文本编码器权重冻结（或采用低学习率微调）以保持其对开放词汇的良好嵌入能力。随后，与图像字幕生成任务的联合训练开始：使用标准交叉熵损失训练Transformer编码器-解码器，同时包含原型记忆注意力和区域提示交互。**参数共享/冻结方面**，原型记忆模块主要引入少量新增参数（记忆键值的存储和读取操作），这些参数在训练中从随机初始化开始学习；CLIP编码器的大部分参数可冻结以降低训练难度，只针对特定层进行Adapter式调整 [17]。整个模型通过**多任务损失**共同训练：以字幕生成的语言模型目标为主，辅以FSGR的对比损失及可能的CLIP空间对齐损失作为正则。训练过程中，每个mini-batch动态更新外部记忆库（缓存在一定窗口内的键值并进行聚类形成原型） [18] [10]。在这种协同训练下，PMA的记忆机制和FSGR的区域提示模块能够彼此配合：记忆模块提供全局语义先验，区域提示模块确保局部语义精准，使模型更好地对齐视觉内容和生成文本。

结构设计创新点

这套融合模型在结构上具有以下显著创新，与现有图像字幕模型形成区别：

- **原型记忆与当前样本结合的双重注意力：** 与传统仅关注当前图像特征的Caption模型不同，我们引入了**跨样本的原型记忆注意力机制**。模型能够对训练集中其它样本的隐含语义模式进行检索利用 [19]。这种记忆增强的注意力让模型在生成过程中参考“见过的类似场景怎么描述”，据此丰富和纠正当前描述。这在Barraco等人提出的PMA-Net中已被验证可显著提升COCO数据集指标（CIDEr提高3.7分） [19]。因此，本模型通过保留并改进PMA-Net的记忆模块，实现**语义先验迁移**，这是对现有Transformer字幕模型的一大创新。

- **区域级文本提示的视觉语义对齐：** 相较于以往仅利用视觉特征或检测到的标签进行描述的方法，我们融合了**CLIP驱动的细粒度文本提示模块**。FSGR方法首次将CLIP的跨模态知识引入图像字幕生成的编码过程，通过给每个图像区域分配显式的文本标签来桥接视觉和语言 [4]。这一设计使模型具有**开放词汇的感知能力**：即使遇到训练集中未见过的新概念，只要CLIP词典中存在类似词语，模型也能通过提示获取相关语义。这种区域文本提示在细粒度语义对齐上优于以往方法。例如，以前一些工作尝试将图像检测到的属性或物体词嵌入Caption模型，但由于类别有限且视觉-文本嵌入空间不统一，容易出现语义错位或空间信息丢失的问题 [20]。我们的设计通过在CLIP公共空间中对齐视觉patch和文本概念，解决了以往视觉特征和文本概念对接不准的问题 [20]。因此，相比传统注意力模型，我们在结构上**显式嵌入了语言提示通道**，这是另一项重大创新。

- **多模态知识的协同融合：** 本模型将**训练数据的全局语义模式记忆**（通过原型记忆获得）与**图像内部的局部精细语义**（通过CLIP区域提示获得）统一在一个框架下。这种融合实现了**“全局语义先验 + 局部精细对齐”**的协同作用：记忆模块提供宏观的语义上下文（如场景常识、惯用描述），而区域提示模块确保具体视觉内容（如对象和属性）被正确描述并与语言空间对齐。两者互补，提升字幕的丰富度和准确性。这种双通道融合的结构在图像字幕领域尚属新颖，突破了单一模态信息来源的限制。

- **增强的视觉-语言一致性和泛化能力：** 由于CLIP提示模块将视觉表示转换为显式的语言形式，我们的解码器实际上是在同时读取视觉特征和“用词描述的视觉概念”。这使生成的句子在措辞用语上更贴近人类描述习惯，显著提高了视觉-语言的一致性。另一方面，CLIP提供的开放词汇和跨域知识，以及记忆模块提供的训练集多样经验，都有利于模型对**新颖场景和开放域图像**的泛化。相比仅在COCO上训练的模型，我们的方法在跨域应用中预期表现更好——这也与FSGR报告的跨域能力提升一致 [21]。综合而言，本融合架构在设计上实现了**更深的跨模态交互**（视觉patch与文本概念对齐）以及**更广的知识调用**（历史经验记忆），是对现有图像描述模型的明确改进。

潜在应用场景

融合原型记忆和区域文本提示的高级图像描述能力，使该模型具备广泛的应用前景：

- **图像搜索与检索：** 更加精准详细的图像字幕有助于图像检索系统理解图像内容。当每张图片都附有一致且语义丰富的描述时，用户可以通过自然语言检索找到匹配的图片。本模型产生的字幕在视觉-语义对齐上表现更佳，提升了以图搜图、以文搜图的效果。例如，在电商平台中，通过查询“红色双肩包户外场景”，系统可检索到描述中提及类似关键词的图片。

- **多模态内容推荐：** 在社交媒体或内容分发平台上，模型生成的高质量字幕可用作跨模态推荐的桥梁。一方面，字幕提炼了图像的关键信息，便于与文本内容计算相似度；另一方面，原型记忆确保描述风格符合常见语义，使推荐系统更准确地匹配用户兴趣。例如，根据一张风景照生成的字幕“**蓝天白云下的群山**”，系统可以推荐相关文章或诗歌等相关内容。

- **辅助弱视/盲人用户：** 对于视力障碍用户，我们的方法可生成**高度详尽且准确的图像描述**，帮助其“听见”图像内容。由于融合了记忆常识和局部细节，描述将既涵盖图像中的主要对象和动作，也包括场景氛围等信息，让弱视用户对图像有完整的理解。例如，一张复杂街景照片，我们的字幕不仅会提到“有人行走在街道上”，还可能补充“街道两旁是复古建筑，路面湿滑反射灯光”等细节。

- **图文内容审核与生成：** 模型生成的描述在视觉-语言对齐上的可靠性使其可用于内容审核，检测图像与其说明文字是否匹配，以及辅助生成图文内容。比如，在新闻媒体中，给一张新闻照片自动生成描述以辅助记者撰写说明文字；或者在用户上传图片时自动生成标题和标签，便于审核人员快速了解图片内容是否违规。

- **专业领域图像解读：** 在医疗、生物等专业领域，我们的方法同样适用。例如医疗影像报告场景，模型可以结合记忆中学到的医学术语和影像特征，生成描述性报告辅助医生诊断（如“**肺部X光显示右下叶有局灶性阴影**”）。又如博物馆文物图像描述，模型可以识别出文物的材质和用途并生成说明。在遥感影像分析中，模型可以描述卫星图像中的地物分布。这些应用得益于模型开放词汇的获取和对细节的重视，使其能够适应各种专业图像描述任务。

实现可行性与优化建议

实现可行性： 从工程角度看，将PMA-Net和FSGR模块融合是较为可行的。PMA-Net的开源实现（GitHub: aimagelab/PMA-Net）基于HuggingFace Transformer架构，已经验证了**在Transformer中集成外部记忆**的效果 [22]。FSGR的方法（Gao等人, 2025）在论文中提供了清晰的算法流程和模块划分 [23]；尽管可能没有完整开源代码，但其核心组件（CLIP模型、语义分配、对比学习）均有现成实现可参考。例如，CLIP模型可直接使用OpenAI提供的预训练权重，S-CLIP（用于生成patch语义标签）等也有开源项目。本融合方案主要涉及将**CLIP编码器与Transformer字幕模型对接**，以及**增设记忆查询操作**，这些都在当前框架下可编程实现。因此，在PyTorch或TensorFlow中复现该架构应当比较直观：可重用PMA-Net代码框架，将CLIP编码器嵌入其中，增添FSGR的语义处理步骤。

显存占用与优化： 融合模型的复杂性会带来一定的显存开销，主要来自CLIP图像编码器（ViT大型模型）、存储原型记忆的额外张量，以及Transformer解码器多一个注意力子层。为控制显存与加速训练，提出以下策略： - 分阶段训练降低峰值内存： 可先训练或细调CLIP编码器的patch对比模块，获得优化的视觉特征，再冻结CLIP大部分参数，仅保留Adapter层参与后续Caption训练 [17]。这样在大部分训练时间里CLIP参数不反传梯度，节省显存。原型记忆聚类的计算也可离线进行一定程度的预处理。 - 记忆模块的轻量实现： 原型记忆键和值向量数目m是可控的 [24]。可根据GPU容量选择适当的m（比如几十到一两百），避免存储过多记忆向量。PMA-Net论文指出增加聚类数和记忆库大小会提升性能但也提高开销 [25]，需要折中选择。实现中可利用FAISS库加速近邻搜索构建原型 [22]，并采用半精度（FP16）存储记忆向量以减半显存占用。 - 动态更新与缓存策略： 记忆库不必每步都完整更新，可采用滑动窗口或间隔更新策略 [18] [26]。例如每隔若干iteration进行一次

聚类刷新，平时固定使用最近的原型。这减少了频繁更新开销。也可以对历史样本激活进行随机采样加入记忆，限制每次聚类的数据规模。 - 训练速度调优： 可以使用混合精度训练和梯度累积来缓解单卡显存压力。针对融合后的复杂损失函数（caption交叉熵 + 对比损失等），可调整各部分损失权重的平衡，先以较高权重训练主要任务（字幕生成），再逐步增加对比损失权重以细调语义对齐。这样确保模型先学会生成句子，再学细节对齐，不至于因训练难度过大而收敛变慢。

实用性验证： 值得一提的是，PMA-Net和FSGR各自的论文报告了可靠的性能提升 [19] 。两者融合后的方法因为兼顾宏观和微观语义，预期能在COCO等基准上进一步提升评价指标，同时保持对未知对象的描述能力。模型复杂度虽有所增加，但仍主要由Transformer模块主导，原型记忆查询在推理阶段等价于一层额外注意力，CLIP编码器也可提前计算好图像特征。因此推理速度上仍可接受，有潜力应用于实际系统。综合来看，该融合路径在当前软硬件条件下具备良好的可实现性和扩展性。

论文撰写大纲

- **标题（建议）：** 融合原型记忆和CLIP区域提示的图像字幕生成 （或 Prototype Memory-augmented Image Captioning with CLIP-based Region Prompts）。

- **摘要（范例）：** 本文提出了一种创新的图像字幕生成模型，将原型记忆网络（PMA-Net）与CLIP支持的细粒度区域文本提示融合，以提升视觉语义对齐和字幕描述一致性。模型采用Transformer架构：通过CLIP编码器提取图像patch特征并生成对应的语言标签提示，结合原型记忆模块从训练语料中提炼语义原型供解码器检索。在MS COCO基准上，我们的方法在保持跨域泛化能力的同时取得了比现有方法更高的BLEU、CIDEr等指标 [19] 。定性分析显示，融合模型生成的字幕在细节刻画和用词准确性方面显著优于对比模型。所提方法证明了跨样本记忆与区域级跨模态提示相结合的有效性，推动了图像描述任务向更高水平发展。

- **章节结构：**

- **引言：** 介绍图像字幕生成任务背景和挑战，指出视觉特征语义间隙和生成一致性问题；引出本文以原型记忆和区域文本提示融合来解决上述问题的思路，概述主要贡献和实验结果。

- **相关工作：** 回顾图像字幕模型的发展，包括基于注意力的编码-解码模型、记忆增强的Caption方法（如Meshed-Memory等）、以及利用CLIP等预训练多模态模型改进字幕生成的最新工作（如ClipCap、FSGR等），突出现有方法的局限并为本文方法奠定基础。
- **方法：** 详细描述融合模型架构。首先介绍PMA-Net原型记忆模块的机制和公式 [27] ；然后介绍FSGR模块，说明如何通过对比学习优化CLIP视觉编码、构建对象概念词典、进行语义量化形成语言桥接图 [28] ；接着阐述两模块在Transformer中的集成方式（双重交叉注意力结构），给出模型信息流图和关键算法步骤。公式推导模型各部分的前向计算和损失函数构成。
- **实验：** 分别在MS COCO、NoCaps等数据集上进行实验。与几种代表性模型（Transformer基线、带记忆的Meshed Transformer、ClipCap、以及单用FSGR的模型等）进行定量比较，报告BLEU、METEOR、ROUGE、CIDEr等指标，突出我们方法的提升幅度。进行消融实验：分别去除原型记忆或区域提示模块，验证两者对性能的贡献；改变记忆容量m和区域提示规模等超参数，分析对结果的影响。还将结果与当前SOTA模型比较，突出本方法在有无SCST微调情况下的领先表现。
- **结果与分析：** 给出定性结果展示，包括模型生成的字幕案例与真实描述的对比。用可视化图展示模型的注意力权重，例如哪个原型记忆向量在起作用、解码器在生成某词时关注了哪几个图像区域及对应文本提示（可绘制热力图）。分析我们的模型如何更好描述了图像细节（举例说明有无区域提示时字幕差异）以及应对未见概念的能力（举例模型成功描述了罕见对象）。讨论可能的错误和局限，如过度依赖记忆导致的陈词滥调风险，以及CLIP提示错误匹配的情况，并提出改进思路。

- **结论：** 总结全文，重申本文融合原型记忆和CLIP区域提示的框架如何改善图像字幕生成效果，指出其在更大范围多模态应用中的潜力。展望未来工作，例如引入视频描述、结合更大型的多模态模型（如视觉大模型）等扩展。

- **主要图表清单：**

- **图1：模型架构示意图。** 展示完整流程：图像编码（标注CLIP提取patch特征和文本提示分配）、原型记忆模块（绘制若干记忆单元插入解码器注意力）、以及Transformer解码器如何同时利用图像特征、文本提示和记忆向量生成字幕。该图突出我们方法的新组件（用颜色或虚线标出FSGR模块和记忆模块）相对于标准Transformer的区别。

- **表1：性能对比表。** 列出本模型与现有模型在COCO等数据集的评价指标（BLEU-4, CIDEr等）得分。用粗体标出最佳成绩，注明我们的模型在各指标上相对提升的百分比。可能还包括一行展示模型参数量或推理速度以供参考。
- **图2：消融实验柱状图/折线图。** 例如，用柱状图表示去除记忆或去除区域提示对CIDEr得分的影响；或折线图表示不同记忆原型数m对性能和计算开销的权衡。此图体现各模块对性能的贡献。
- **图3：定性结果示例。** 显示若干输入图像，分别给出Baseline模型生成的描述和我们模型生成的描述，对比之下突出我们方法更准确丰富的地方。可以在图中高亮模型关注的图像区域及其对应提示词。例如在一张复杂场景图上，用颜色标示出模型通过提示识别的关键对象（如车辆、行人）及生成的相关描述片段。这张图有助读者直观理解本方法在语义捕捉上的优势。

以上方案详细阐述了一篇融合PMA-Net与FSGR模块的图像字幕生成论文的构思。从模型设计到创新点、应用前景、实现细节及论文写作框架均已给出，可作为后续撰写高质量学术论文的基础。通过该方案，读者将清楚地理解我们所提出方法的动机与价值。 19 4

---

1 2 3 4 5 6 13 14 15 16 17 20 21 23 28 √CLIP—交叉注意力—Transformer.pdf
file://file_00000000773462o6b62640bd13a0d3fe

7 8 9 10 11 12 18 22 24 25 26 27 With a Little Help from Your Own Past: Prototypical Memory Networks for Image Captioning
https://openaccess.thecvf.com/content/ICCV2023/papers/Barraco_With_a_Little_Help_from_Your_Own_Past_Prototypical_Memory_ICCV_2023_paper.pdf

19 Image captioning using transformer-based double attention network | Request PDF
https://www.researchgate.net/publication/374354044_Image_captioning_using_transformer-based_double_attention_network