突破长视频处理瓶颈：MA-LMM 长时记忆模型机制与效能研读报告

陈奕璇 22330009 张昕煜 22330139 朱丹仪 22330156

1 引言

随着计算机视觉和自然语言处理技术的融合，能够同时处理视觉和文本数据的大型多模态模型取得了显著突破。本报告聚焦于长视频理解领域的一项创新成果——通过引入长期记忆银行[1] 设计来增强大型多模态模型处理能力的最新研究。该方法通过解决长视频处理中的时序冗余和 GPU 内存限制等核心挑战，在视频问答、内容描述和长视频理解等任务中实现了最先进的性能表现。我们发现，这项研究不仅有效解决了传统方法在处理长视频时面临的一系列问题，还通过双重记忆库的设计实现了对视频信息的高效压缩和检索，为长视频理解领域的发展提供了新的方向和思路。

2 研究背景

多模态大模型通过整合图像、视频、文本等数据类型，实现视觉问答、图像描述和视频理解等任务。在前些年，受到 BERT、GPT 等大语言模型成功的启发，部分图像-语言模型如 CLIP、BLIP 和 Flamingo 也应运而生。但是将这类模型放到视频理解领域时，该如何处理时序信息成为了一大特殊挑战。与静态图像不同，视频需要模型捕捉帧与帧间的长期依赖关系。对于长视频而言，计算复杂度和内存需求问题尤为突出。。通过研读该领域近些年来的相关研究，我们发现已有部分学者尝试突破这些挑战，例如 VideoMAE[2] 采用掩码自编码器进行视频自监督预训练，AdaFrame 和 ScSampler 提出自适应帧选择算法来降低计算开销。虽然已经取得进展，但是受限于 GPU 内存和语言模型上下文长度，长视频处理仍具有挑战性。

3 论文方法与创新

本报告研究的 MA-LMM 模型采用的记忆银行创新方案为此提供了新的解决思路。其核心创新思想是加入一个” 记忆库” 来储存视频的过去信息，从而帮助模型理解视频的长远内容，突破长视频理解的瓶颈。而这个” 记忆库” 通过以下的三个核心模块来实现：视觉特征提取模块、长期记忆银行模块和文本解码模块。

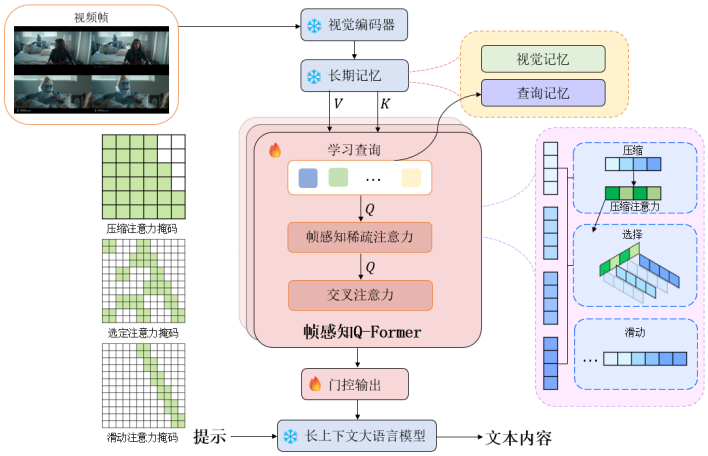

图 1: MALMM 总流程

3.1 视觉特征提取

受人类处理长期视觉信息的认知过程启发，MA-LMM 采用顺序处理视频帧的方法，将新的帧输入与长期记忆库中存储的历史数据动态关联，只保留有区分性的信息。这种选择性的信息保留方式，使得视频理解过程更加高效且可持续，并支持自动化的在线视频推理任务。

具体来说，给定 T 帧视频序列，将每一帧视频输入到预训练的视觉编码器中，提取得到视觉特征 $V = [v_1, v_2, ..., v_T]$，并通过位置嵌入层（PE）注入时间顺序信息：

$$f_t = v_t + PE(t)$$

3.2 长期记忆银行

长期记忆银行是 MA-LMM 的核心创新组件，这一长期时间建模机制旨在通过有效利用视觉和查询记忆库来增强视频理解，负责动态维护和更新历史视频信息。该建模整体使用的是与 BLIP-2 中相同的查询变压器（Q-Former）架构，用于将视觉嵌入对齐到文本嵌入空间，并通过处理学习到的查询来捕获视频帧中的时间信息。但与原始 Q-Former 只关注当前帧嵌入不同，本研究还设计了一个由视觉记忆库和查询记忆库组成的长期记忆库，它积累了历史视频信息并且增加了交叉和自我注意力层的输入，能够更有效地进行长期视频理解：

3.2.1 视觉记忆库

视觉记忆库存储从冻结的视觉编码器提取的原始帧特征。随着每一个新帧的输入，过去的视觉特征被串联成一个列表，从而使得模型可以对长时间的视觉上下文进行注意。在交叉注意力机制中，视觉记忆库充当静态的键-值对，为当前的查询提供全局视觉上下文：

$$K = F_t W_K$$

$$V = F_t W_V$$

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V$$

且因为 Q-Former 中的所有交叉注意层都关注相同的视觉特征，所以视觉记忆库共享于所有 Q-Former 块的跨注意力层。

3.2.2 查询记忆库

查询记忆库存储每个时间步的输入查询，通过存储这些查询，维持模型对每个帧的理解和处理的动态记忆。在自注意力机制中，查询记忆库充当动态键-值对：

$$K = Z_t W_K$$

$$V = Z_t W_V$$

不同于静态的视觉记忆库，输入查询会在 Q-Former 块之间演变，以增加的抽象级别捕获不同的视频概念和模式。因此每个自注意力层都有一个唯一的查询记忆库，其中包含的输入查询随着训练迭代更新。

3.2.3 记忆压缩算法

为了应对随着每一帧新数据的到来，记忆库不断增长，从而导致 GPU 内存和计算成本增加的挑战，模型通过一种新的记忆库压缩（Memory Bank Compression）技术来处理大量视频数据。具体流程是，首先计算相邻帧之间的余弦相似度来识别冗余的特征：

$$s_t^i = \cos(f_t^i, f_{t+1}^i), t \in [1, M], i \in [1, P]$$

接着，模型会选择最高相似度的 tokens 进行平均：

$$k = \text{argmax}_t(s_t^i)$$

$$\hat{f}_k^i = \frac{f_k^i + f_{k+1}^i}{2}$$

这种方法在减少记忆库大小的同时，还保留了具有辨识性的特征和时间上下文。当记忆库的大小超出设定阈值时，模型会在每个时间步应用这种压缩技术。

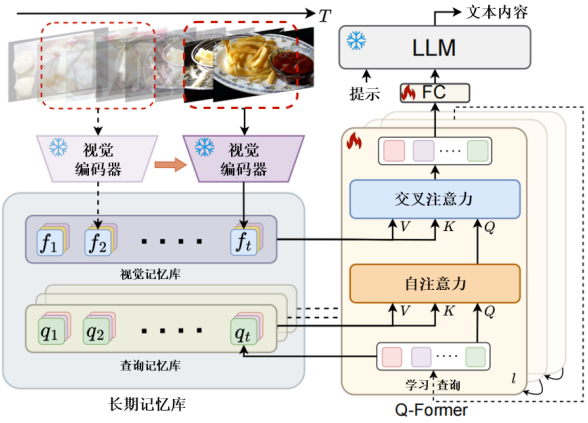

图 2: 长期记忆银行工作机制

3.3 文本解码

由于模型以自回归方式处理视频帧，Q-Former 在最后一个时间步的输出包含所有的历史信息，将其输入到 LLM 中。训练时，使用标准交叉熵损失进行监督：

$$\mathcal{L} = -\frac{1}{S}\sum_{i=1}^{S}\log P(w_i|w_{<i}, V)$$

4 论文实验复现结果

4.1 复现简述

原论文的下游任务一共有四个，分别是长视频理解、视频问答、视频字幕生成和在线动作预测。在本研究中，考虑到我们设备的算力，我们只选择复现其中的一项任务：长视频理解。

4.2 实验设置

- 环境配置: python3.10.17
- 硬件配置: 对于视觉编码器，我们采用 EVA-CLIP 中的预训练图像编码器 ViT-G/14，它还可以被替换为其他基于 Clip 的视频编码器。我们使用 InstructBLIP 中的预训练 Q-Former 权重，并采用 Vicuna-7B 作为 LLM。所有实验均在 4 个 4090 GPU 上进行。
- 数据集: LVU

- 评估指标:Top-1 分类准确率（表示模型预测结果中排名第一的类别与真实类别相符的比例，能直观反映模型对视频内容分类的准确程度，比例越高，说明模型在分类任务上的表现越好）

4.3 关键结果展示

我们将自己跑出来的 MALMM 结果与论文中展示的其他模型的结果进行对比如下表所示：

表 1: 与 LVU 数据集上的最新方法比较，加粗表示排名第一的结果

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
S5	**67.1**	42.1	73.5	67.3	**65.4**	51.3	48.0	59.2
MALMM	58.2	**44.8**	80.3	**74.6**	61.0	**70.4**	**51.9**	**63.0**
复现	57.9	**44.8**	**80.4**	74.5	60.6	**70.4**	51.8	62.9

如表 1 所示，我们的复现结果与原始论文报告指标的相对误差均小于 1%（如 MALMM 在 Genre 任务上差异-0.6%），部分差异可能源于硬件差异导致的随机性。

5 实验思考与改进

5.1 模型优化方向

MALMM 模型提出了 Visual Memory Bank 和 Query Memory Bank，通过 MBC 减少冗余降低内存占用，而它沿用了 Qformer 的 Attention 机制，我们猜想通过压缩、滑动窗口和选定路径协同视觉和查询内存库，为长时间多模态推理提供高效、查询感知的稀疏注意力，从而改进 Qformer 的 Attention 机制，在尽可能不影响内存占用的情况下取得更高的精度。

5.2 优化尝试结果

我们将 Qformer 部分改写，融合了 Native Sparse Attention 的 Self Attention（改进后的代码会在 demo 中详细解释），并在 LVU 数据集上进行测试，得到如下结果：

表 2: 与 LVU 数据集上的最新方法比较，加粗表示排名第一的结果

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
S5	**67.1**	42.1	73.5	67.3	65.4	51.3	48.0	59.2
MALMM	58.2	44.8	80.3	**74.6**	61.0	70.4	51.9	63.0
Ours	60.6	**45.0**	**81.1**	73.9	**70.1**	**72.3**	**52.3**	**65.0**

结果显示，优化后的方法在 LVU 数据集上以 65.0% 的平均准确率取得最佳性能，略优于 MALMM（63.0%）。在 Metadata 任务（如'Genre'、'Writer'、'Year'）上表现突出，同时在整体推理能力上验证了优化注意力机制的有效性，兼顾了精度与内存效率。

6 总结

MA-LMM 引入的长期记忆银行，通过顺序处理视频帧和存储历史数据，解决了大语言模型的上下文长度和 GPU 内存限制问题, 提高了模型的高效性和灵活性。长期记忆库可即插即用，实验表明该方法具有显著优势，为长视频理解研究提供了有价值的见解。本报告系统梳理了该模型的技术方案，整合了模型的优点，并且对实验进行了复现和相关改进，在原论文的基础上达到了更理想的效果。

参考文献

[1] He B, Li H, Jang Y K, et al. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13504-13514.

[2] Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[J]. Advances in neural information processing systems, 2022, 35: 10078-10093.

[3] Tang Y, Bi J, Xu S, et al. Video understanding with large language models: A survey[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[4] Jin P, Takanobu R, Zhang W, et al. Chat-univi: Unified visual representation empowers large language models with image and video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13700-13710.

[5] Song E, Chai W, Wang G, et al. Moviechat: From dense token to sparse memory for long video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 18221-18232.