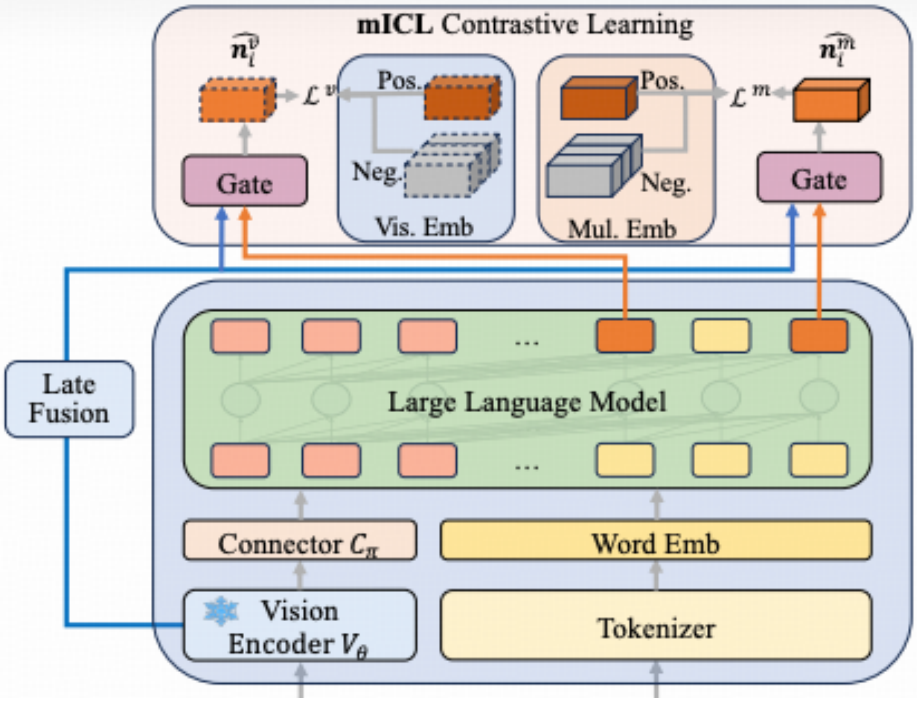


2024小红书：NoteLLM-2-融合多模态大模型，赋能推荐系统

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

随着互联网的发展，大多数在线平台都充斥着丰富的多模态信息，以促进用户参与度。多模态推荐作为这些平台的重要服务，根据用户对物品多模态信息的兴趣进行推荐。多模态[+]推荐的关键在于有效的多模态表示，它通过将多模态内容转化为嵌入并计算相似性，服务于检索、聚类和物品建模等任务。 已有的多模态表示模型如CLIP、METER和Siglip等在多模态处理上有所成就，但仍存在参数扩展空间。然而，研究往往侧重视觉模型的扩展，忽视了文本部分的优化。此外，纯文本导向的预训练不足导致文本分支在处理文本信息时表现不佳。

因此，当前亟需强化多模态表示模型的文本理解能力，以提升其全面且精准的多模态处理性能。大型语言模型[+]（LLMs）因其在文本处理的强大能力而受到关注。尽管已有研究探讨LLMs在生成文本嵌入上的优势，但对于如何增强多模态表示的理解，相关工作尚不多见。

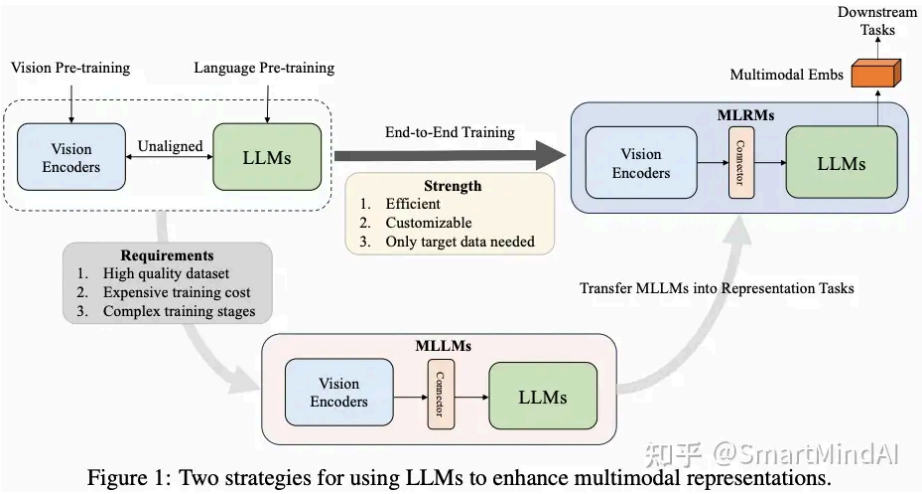


Figure 1: Two strategies for using LLMs to enhance multimodal representations.

当前，预训练MLLM对大规模、高质量多模态数据的依赖限制了个性化的定制训练。为解决这一问题，我们提出了多模态大模型（MLRMs）。MLRMs是一种专为增强多模态表示理解而设计的LLMs解决方案。我们研究了多模态上下文学习（mICL）和后期融合机制，这两种策略用于提升大型语言模型（LLMs）在多模态表示中的理解能力。mICL通过将多模态内容分离并分别压缩为模态

知乎

Preliminary Investigation

Preliminaries

问题陈述。在我们的场景中，每个项目代表一个笔记，由用户生成并表达他们的生活经验。我们在附录⁺H中展示了一些笔记。我们假设N = {n_1, n_2, ..., n_m}为笔记池，其中m是笔记的数量。每个笔记包含文本信息，如标题、主题和内容，以及图像。我们将第i个笔记表示为n_i = (t_i, tp_i, ct_i, v_i)，其中t_i, tp_i, ct_i, v_i分别表示标题、主题、内容和图像。给定查询笔记n_i，I2I推荐系统⁺根据多模态内容相似性对笔记池N\\{n_i}进行排名。该系统的目的是优先考虑与给定笔记相关的目标笔记的排名。

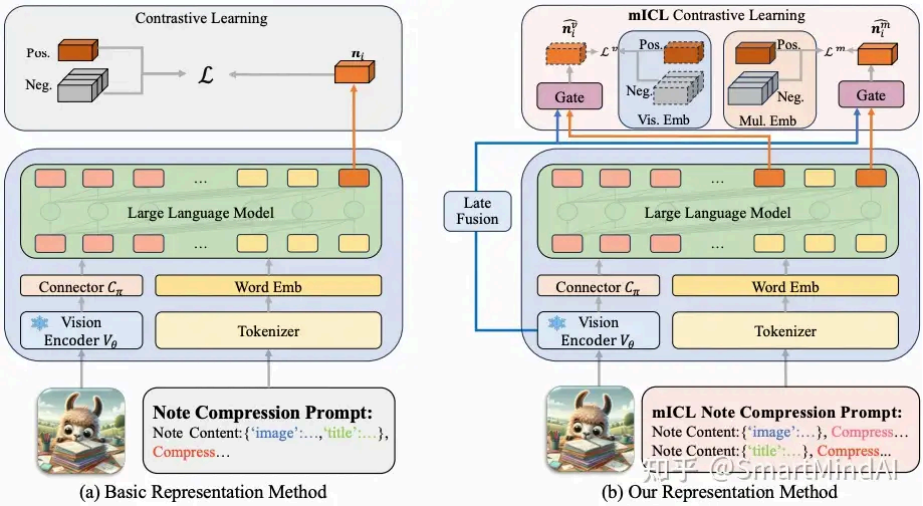
数据集构建。我们利用共现机制基于用户行为建立相关笔记对，而无需人工注释。这个机制假设经常一起阅读的笔记很可能是相关的。然后，我们计算用户查看笔记n_A并随后点击笔记n_B的共现分数如下：

$$s_{n_A \to n_B} = \sum_{u \in U_{n_A \to n_B}} \frac{1}{N_u}$$

对于每一条笔记n_i，我们首先计算它与其他所有笔记的互动评分$s_{n_i \to n_j}$，其中$s_{n_A \to n_B}$代表笔记n_A与n_B的互动分数。设$s_{n_A \to n_B}$表示用户在观看n_A后选择n_B的观察次数$U_{n_A \to n_B}$是这些观察的用户集合N_u是用户u实际点击过的笔记总数。通过遍历所有笔记对，我们构建了笔记n_i的共现分数集$\mathcal{S}_{n_i}$，包含与之相关的所有非自我评分，即$(1 \le j \le m, i \ne j)$。

为了排除异常值，我们设置了上限up和下限low，移除$\mathcal{S}_{n_i}$中超出这些界限的高分或低分项。最后，从筛选后的共现分数集中，我们选择与笔记n_i最相关性的t个笔记，形成其关联笔记集。

MLRMs的多模态嵌入表示



我们延续之前的方法，对多模态内容进行整合，采用明确的词汇限制来生成嵌入。以JSON格式组织笔记内容，我们的处理流程如下：

1. 将笔记内容转化为结构化的JSON格式，包含文本信息和潜在的视觉元素。

2. 对文本部分，我们利用预训练LLM提取每个单词的嵌入。

3. 结合视觉信息，如果适用，利用视觉编码器⁺获取图像特征嵌入。

4. 将所有模态的嵌入合并，通常通过某种融合策略如注意力机制或者简单加权平均。

5. 进行词汇限制，确保嵌入的维度保持在可操作范围内，避免过拟合。

6. 最后，通过训练或微调，将压缩后的多模态嵌入融入到我们的多模态代表模型（MLRMs）中。

我们采用MLRMs的基础表示方法，处理多模态信息。首先，通过视觉编码器V_θ（如图(a)所示），使用\<IMG>替换为来自原始图像的视觉嵌入v_i，将其转换为维数为$L \times h_v$的视觉特征Z_v。接着，连接器C_π将这些视觉特征投影到LLMs的词嵌入+空间，生成维数为$L_c \times h_t$的视觉嵌入E_v。

文本部分，我们按照预训练LLM提取每个单词的嵌入，形成文本词嵌入E_t，其维度为$T \times h_t$，其中T是文本的令牌数量。当遇到\<IMG>时，我们插入相应的视觉嵌入，形成多模态嵌入E_m，其维度扩展至$(L_c + T - 1) \times h_t$。

最后，利用LLMsLLM_μ处理多模态嵌入E_m，得到最终的隐藏状态H，维度也是$(L_c + T - 1) \times h_t$。这个隐藏状态n_i代表了笔记n_i的压缩多模态表示，确保了LLMs能够以预测下一个令牌的形式整合和压缩多模态上下文信息。然而，由于LLMs主要通过语言模型任务训练，与表示学习的目标不同，我们引入了对比学习来弥合两者间差距。对比学习，作为一种常见于嵌入学习中的策略，被用来优化。 具体操作是，每次迭代处理B对相关笔记对(n_i, n_i^+)，这构成一个小批量，总计$2B$笔记。我们依照先前的步骤（引用文献或前文对应的部分），通过梯度下降法+来最小化对比损失函数，表达如下：

$$L_{contrastive} = -\sum_{(n_i, n_i^+)} log \frac{\exp(sim(E_{m_i}, E_{m_i^+})/\tau)}{\sum_{j=1}^{2B} \exp(sim(E_{m_i}, E_{m_j})/\tau)}$$

这里$sim(\cdot)$表示嵌入间的相似度度量τ是temperature参数，确保分布不对称E_{m_i}和$E_{m_i^+}$分别代表笔记n_i和相关笔记n_i^+的多模态嵌入。通过这种方式，LLMs在对比学习中学习如何区分相关和不相关笔记，从而更有效地适应表示学习任务。

$$\mathcal{L}(\pi, \mu) = -\frac{1}{2B} \sum_{i=1}^{2B} log \frac{e^{sim(n_i, n_i^+) \cdot e^\tau}}{\sum_{j \in [2B] \setminus \{i\}} e^{sim(n_i, n_j) \cdot e^\tau}}$$

在这个过程中，我们引入τ作为可调节的温度参数，它控制着相似度的软化程度。$sim(a, b)$计算的是向量a和b的点积+归一化值。我们使用这样的点积比较来定义嵌入之间的相似度。

为了优化，我们使用对比学习，通过对比每对笔记对(n_i, n_i^+)的多模态嵌入E_{m_i}和$E_{m_i^+}$，计算对比损失$L_{contrastive}$。这个损失鼓励模型区分正样本（相关笔记）和负样本（随机选取的其他笔记），使得LLMs能更好地理解并适应表示学习任务，而不是仅仅依赖语言模型的训练。 特别地，我们只更新连接器C_π和LLMsLLM_μ，保持视觉编码器V_θ不变，这样做的目的是扩大批量规模以提升性能，同时保持对视觉信息的原有处理。

Datasets and Experimental Settings

在模型训练中，我们利用8台配备80GB Nvidia A100 GPU，每台分配16个批次，总计128个批次。每个批次包含256份笔记数据。初始设置中，我们设定温度参数τ为3。这些详细信息，包括但不限于硬件配置和训练策略，已在附录中给出，供进一步参考。为了衡量MLRMs的表现，我们在测试集中以查询笔记为基础，对所有笔记进行排名。我们通过计算目标笔记在排名列表中的精确位置来评估，Recall\@100、Recall\@1000和Recall\@10000三个指标。这涵盖了测试集、短查询与短目标的全面评估。

Performance of Multimodal Representation of Fine-tuned MLRMs

我们在附录中探讨了多模态语言模型的零样本适应性问题，发现尽管零样本方法无法有效提升模型在表示任务上的性能，它仍落后于基线BM25。为此，我们设计了三种端到端的多模态学习模型：

MTomato-Base，基于持续预训练的LLM Tomato，CLIP ViT-B作为视觉模块，以及随机初始化的Q-Former；

MQwen-Base是MTomato-Base的变体，更换为Qwen-Chat；

MQwen-bigG则用ViT-bigG替换原模型的视觉部分。为了提高效率，视觉嵌入尺寸设为16。

对比实验中，我们选取了**BLIP-2**+和**Qwen-VL-Chat**作为预训练模型，所有视觉编码器保持冻结状态。同时，我们还包含了其他多模态表示模型作为基线，详细信息在附录中列出，以供详尽比

多模态能力，但对视觉编码器大小敏感，CLIP ViT-B的改进不如预期，MTomato-Base和MQwen-Base相对于原模型改进有限，分别只有1.31%和1.36%的提升。尽管MQwen-bigG在效率上领先，但与Qwen-VL-Chat的性能差距仍然存在，这表明优化策略和预训练模型的选择对提升多模态语言模型的表示能力至关重要。

Table 2: Representation performance of fine-tuned MLRMs under different modal inputs (%).

Method	Input	All Pair			Short Query Pair			Short Target Pair		
		R@100	R@1k	R@10k	R@100	R@1k	R@10k	R@100	R@1k	R@10k
Multimodal Training										
BLIP-2	Image	24.72	45.21	69.68	26.07	48.78	75.59	27.59	49.78	75.97
	Text	56.85	75.56	88.26	32.85	47.28	64.66	34.88	51.68	72.98
	Multimodal	68.38	88.22	97.44	52.97	77.83	94.40	54.00	79.74	95.29
MTomato-Base	Image	0.60	1.49	6.98	0.54	1.51	9.15	0.84	2.13	9.70
	Text	71.49	87.59	95.31	43.80	61.74	79.26	44.72	63.89	83.26
	Multimodal	71.94	88.22	96.13	44.77	63.77	83.31	45.83	66.08	87.02
MQwen-Base	Image	1.92	6.59	23.43	2.86	8.54	28.83	3.05	9.38	29.72
	Text	73.14	88.35	95.34	44.50	61.59	78.78	46.12	64.41	83.97
	Multimodal	74.02	89.65	97.22	48.15	68.35	88.22	49.82	70.91	91.43
MQwen-bigG	Image	18.32	38.50	64.56	19.76	41.56	71.32	21.30	42.72	71.54
	Text	70.63	86.09	93.81	39.94	55.77	72.17	43.03	60.02	79.77
	Multimodal	77.64	92.89	98.91	57.45	80.92	95.73	57.90	83.49	96.99
Qwen-VL-Chat	Image	24.67	45.51	70.08	25.94	48.32	75.32	26.94	49.59	75.67
	Text	71.44	86.78	94.43	42.55	58.59	75.35	43.97	61.11	80.17
	Multimodal	78.54	93.77	99.03	60.43	83.26	96.60	61.63	85.48	97.89

我们对多种模态输入进行了评估，显示在表格中。对于仅图像输入，MTomato-Base和MQwen-Base的性能明显下降，这说明它们在理解和利用图像信息方面存在不足。相比之下，MQwen-bigG凭借其强大的视觉编码器，能更好地表达视觉内容，但即便如此，它在多模态预训练模型如BLIP-2和Qwen-VL-Chat面前，表现仍有所落后，这强调了预训练方法和模态融合的重要性。

Exploring the Bias of Fine-tuned MLRMs

我们通过探究MLRMs在不同LLM层次的注意力模式，深入研究微调后的模型中的偏见。首先，我们从10,000个训练样本中随机选择笔记，按批次计算每种模态（文本和图像）与压缩词的注意力权重。具体步骤包括：

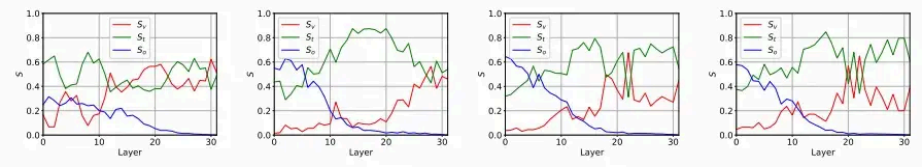
1. 分别获取各LLMs特定层的注意力矩阵。

2. 对每个模态令牌，计算与压缩词的注意力得分。

3. 求和所有模态在该层的注意力分数，得出累积注意力分数。

4. 计算每种模态的平均注意力分数，作为注意力权重。

5. 通过比较权重，分析不同LLMs和层次中模态信息的重要性及潜在偏见。



(a) BLIP-2 (b) MTomato-Base (c) MQwen-Base (d) MQwen-bigG

Figure 3: Relative sizes of S_v, S_t and S_o in different layers of different MLRMs.

这种方法有助于揭示MLRMs在处理和融合不同模态信息时的策略，从而评估潜在的偏见。

$$I_l = \sum_h \left| A_{h,l}^\top \frac{\partial \mathcal{L}}{\partial A_{h,l}} \right|$$

在数学表达中，设$A_{h,l}$代表第l层第h个注意力头的注意力矩阵元素$\mathcal{L}$是对比损失，我们通过累加所有注意力头得到的是第l层的总注意力矩阵I_l，其中$I_l(i,j)$表示从第j到第i位置的信息流动量重要性。为研究压缩词对不同模态信息的贡献，我们将I_l分解为三部分，可以这样表示：

$$I_l = I_l^{text} + I_l^{image} + I_l^{compr}$$

知乎

Methodology

我们的研究表明，未经预训练的MLRMs在处理时忽视了视觉信息，因此我们专注于改进对视觉信号的处理。我们设计了一个名为**NoteLLM-2**的新训练框架，它包含两种策略： 一种是mICL（model-based Instructional Contrast Learning），通过提示（prompt）引导注意力聚焦于视觉信息；另一种是架构层面的后期融合和视觉提示强化，通过延迟融合来增强视觉信息对最终表示的影响。 在实践中，对于笔记n_i，mICL不是直接压缩多模态信息，而是将它转化为两个独立的单模态笔记。接着，我们应用类似于ICL（Instructional Contrast Learning）的方法来整合这两种模态的信息。压缩笔记的提示被重新表述为： "' (通过引入视觉提示，对视觉内容进行改写的提示) "' 这样的方法旨在优化MLRMs对不同模态信息的理解和整合，以提升其表示能力。

mICL Note Compression Prompt.

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{sigmoid}(\boldsymbol{W}[\boldsymbol{v}, \boldsymbol{n}_i^v] + \boldsymbol{b}), \\ \hat{\boldsymbol{n}_i^v} &= \boldsymbol{z} \odot \boldsymbol{v} + (\boldsymbol{1} - \boldsymbol{z}) \odot \boldsymbol{n}_i^v, \end{aligned}$$

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{sigmoid}(\boldsymbol{W}[\boldsymbol{v}, \boldsymbol{n}_i^m] + \boldsymbol{b}), \\ \hat{\boldsymbol{n}_i^m} &= \boldsymbol{z} \odot \boldsymbol{v} + (\boldsymbol{1} - \boldsymbol{z}) \odot \boldsymbol{n}_i^m, \end{aligned}$$

接下来，我们使用融合后的笔记嵌入进行对比学习：

1. 融合的笔记嵌入：$\hat{\boldsymbol{n}_i^v}$和$\hat{\boldsymbol{n}_i^m}$分别代表经过融合处理的视觉和多模态嵌入。
2. 联合表示： 将两者通过串联 '$[\cdot, \cdot]$' 合并，形成一个融合后的联合向量 $\boldsymbol{z_i} \in \mathbb{R}^{2h_t}$ 其中$2h_t$是嵌入的总维度。
3. 学习参数： 使用可学习的参数$\boldsymbol{W}$和$\boldsymbol{b}$来进一步处理这个联合表示。
4. 对比损失：通过计算每个样本（包括目标正样本和负样本）与目标嵌入的相似度与负样本的差异，应用对抗性学习。使用 cosine similarity 或 triplet loss，定义为：

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_{target_i} / \tau)}{\sum_{j=1}^{N_{neg}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_{neg_j} / \tau) + \epsilon}$$

其中τ调整了相似度的尺度N_{neg}是负样本的数量ϵ防止除以零。 对比学习的目标是优化两个模态嵌入的相对关系，增强模型在处理多模态信息时的理解和区分能力，从而提升整体性能。

$$\mathcal{L}^v(\pi, \mu) = -\frac{1}{2B} \sum_{i=1}^{2B} log \frac{e^{sim(\hat{n_i^v}, \hat{n_i^{v,+}}) \cdot e^{\tau}}}{\sum_{j \in [2B] \setminus \{i\}} e^{sim(\hat{n_i^v}, \hat{n_j^v}) \cdot e^{\tau}}}$$

$$\mathcal{L}^m(\pi, \mu) = -\frac{1}{2B} \sum_{i=1}^{2B} log \frac{e^{sim(\hat{n_i^m}, \hat{n_i^{m,+}}) \cdot e^{\tau}}}{\sum_{j \in [2B] \setminus \{i\}} e^{sim(\hat{n_i^m}, \hat{n_j^m}) \cdot e^{\tau}}}$$

最终的损失函数定义为：

$$\mathcal{L}_{total} = \mathcal{L}^v(\pi, \mu) + \mathcal{L}^m(\pi, \mu)$$

其中$\mathcal{L}^v$关于视觉笔记$^+$嵌入（π和μ）的损失，而$\mathcal{L}^m$则针对多模态笔记嵌入。将这两个部分的损失相加，模型同时关注视觉和多模态信息，通过对比学习来协调和优化这两种模态在不同层次的注意力分配和融合效果。这样做旨在提升模型对多模态数据的整体理解和生成能力。

$$\mathcal{L}^f(\pi, \mu) = \frac{\mathcal{L}^v(\pi, \mu) + \alpha \mathcal{L}^m(\pi, \mu)}{1 + \alpha}$$

$$\mathcal{L}^f(\pi, \mu) = \alpha \times \mathcal{L}^v(\pi, \mu) + (1 - \alpha) \times \mathcal{L}^m(\pi, \mu)$$

在评估时，我们用$\hat{\boldsymbol{n}_i^m}$来评估包含多模态信息的笔记嵌入。$\mathcal{L}^f$代表综合损失，它是由视觉部分($\mathcal{L}^v$)和多模态部分($\mathcal{L}^m$)通过权重α（一个调节参数）按比例组合而成。这个权重调整允许我们平衡两者在模型训练中的相对重要性，以达到最优的多模态理解和生成效果。

Experiments

知乎

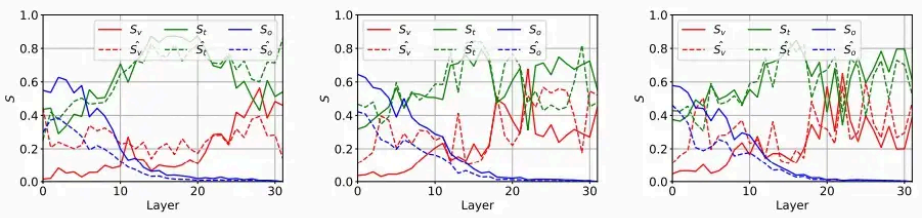
Table 3: Performance of NoteLLM-2 based on three different MLRMs (%).

Method	All Pair			Short Query Pair			Short Target Pair		
	R@100	R@1k	R@10k	R@100	R@1k	R@10k	R@100	R@1k	R@10k
MTomato-Base	71.94	88.22	96.13	44.77	63.77	83.31	45.83	66.08	87.02
+ mICL	74.16	90.95	98.26	51.56	75.86	93.84	52.71	78.12	**95.82**
+ late fusion	74.51	91.37	98.45	54.19	78.02	94.50	54.00	79.71	95.80
+ NoteLLM-2	**74.61**	**91.50**	**98.47**	**54.87**	**78.81**	**94.69**	**54.52**	79.75	95.46
only late fusion	74.22	91.25	98.43	53.21	77.75	94.65	53.65	**79.77**	95.59
MQwen-Base	74.02	89.65	97.22	48.15	68.35	88.22	49.82	70.91	91.43
+ mICL	75.82	91.57	98.55	53.32	76.40	94.38	54.52	79.31	96.05
+ late fusion	75.92	91.49	98.28	52.74	76.03	93.65	54.06	78.89	95.53
+ NoteLLM-2	**76.50**	**92.18**	**98.58**	**55.72**	**78.91**	**94.94**	**56.13**	**80.65**	**96.24**
only late fusion	75.84	91.49	98.35	53.36	76.23	93.57	54.27	78.39	95.30
MQwen-bigG	77.64	92.89	98.91	57.45	80.92	95.73	57.90	83.49	96.99
+ mICL	**77.97**	93.19	**98.92**	**58.50**	**82.08**	**96.66**	58.58	**84.01**	**97.20**
+ late fusion	77.43	92.92	98.78	57.47	81.77	95.89	58.50	83.99	97.16
+ NoteLLM-2	77.56	**93.38**	98.82	57.94	81.98	96.25	**58.79**	83.87	97.05
only late fusion	76.24	92.31	98.67	55.21	79.22	95.58	55.15	81.13	96.70

在实验中，我们对（多模态学习模型）上进行了测试，以验证其有效性。同时，我们也设立了一个基线，即'late fusion'方法，仅在输出阶段合并图像和文本信息，不直接将图像嵌入到LLMs（语言模型）中。实验结果显示

Saliency Scores of Enhanced MLRMs

我们通过比较原模型和增强后的注意力分数，进一步研究了。原模型的注意力分配为S_v S_t和S_o，增强后变为$\hat{S}_v$ $\hat{S}_t$和$\hat{S}_o$。我们把视觉注释视为视觉嵌入E_v的一部分进行分析。



(a) MTomato-Base (b) MQwen-Base (c) MQwen-bigG

Figure 4: Relative sizes of saliency scores in different layers of different MLRMs.

结果显示，增强模型在表示层对图像的关注度显著增强，同时在浅层，对目标输出S_o的关注度减少，而对文本信息S_t基本保持不变。这种变化是由mICL通过压缩提示对两种模态进行相同处理，使得模型能识别并强化图像特征的共同模式实现的。这表明，同时保持了对文本信息的理解，从而优化了多模态任务的表现，没有忽视其他模态的贡献。

Hyper-Parameter Analysis

视觉令牌长度和**文本与视觉损失比例**的分析。

Table 4: Impact of the number of visual tokens L (%).

#Visual Tokens	All Pair			Short Query Pair			Short Target Pair		
	R@100	R@1k	R@10k	R@100	R@1k	R@10k	R@100	R@1k	R@10k
8	74.17	90.85	98.16	52.80	76.13	93.76	53.19	77.87	94.48
16	**74.61**	91.50	98.47	**54.87**	78.81	94.69	**54.52**	79.75	95.46
32	74.55	91.41	98.38	54.56	**78.91**	94.94	54.21	79.83	95.63
48	74.30	**91.57**	**98.48**	54.31	78.83	**95.13**	53.73	**80.06**	**95.92**

Table 5: Impact of textual loss and visual loss ratio α (%).

α	All Pair			Short Query Pair			Short Target Pair		
	R@100	R@1k	R@10k	R@100	R@1k	R@10k	R@100	R@1k	R@10k
1	73.94	91.11	98.38	54.13	78.16	94.67	54.06	79.00	95.38
3	74.38	91.41	98.42	54.15	77.71	94.59	54.06	79.31	**95.53**
9	**74.61**	**91.50**	98.47	**54.87**	**78.81**	94.69	**54.52**	79.75	95.46
19	74.53	91.33	**98.54**	54.56	78.68	**95.00**	53.92	**80.04**	**95.53**

对于视觉令牌长度，我们观察到将长度从16缩短到8时，在处理短序列的任务上表现下降，这暗示了需要在性能和计算效率之间找到一个平衡点。 至于文本与视觉损失比例（通过参数α调节），我

知乎

原文《NoteLLM-2: Multimodal Large Representation Models for Recommendation》

发布于 2024-06-14 10:42 · IP 属地北京

推荐系统 小红书 多模态大模型

理性发言，友善互动

1 条评论 默认 最新

xlhlhy

跟 NoteLLM 第一篇文章相比，最大的区别是啥呢?

09-09 · 上海 回复 喜欢

推荐阅读

AI PaperReading第75篇 2023 CVPR PMR方法 多模...

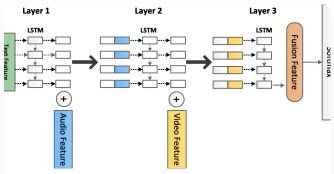
0.基础信息paper: http://arxiv.org/abs/2211.07089 code: none keywords: #多模态平衡 importance: #star4 TLDR: 问题：解决多模态训练中的模态不平衡问题。 名词： prototype 就是...

李晓敏 发表于Paper...

CVPR 2022 oral | OGM-GE: 解决多模态不平衡问题

Title：Balanced Multimodal Learning via On-the-fly Gradient Modulation Link：https://arxiv.org/abs/2203.15332I address (open source)：https://github.com/GeWu-...

Resel... 发表于Resel...



多模态特征融合三部曲

养生的控制... 发表于建模控制与...

视频级多模态生成模型-理 向综述2024

多模态现在已经成了LLMs的 尽管去年已经有很多多模态的 作，但大多停留在探索层面， 用性很远，更不要说对标一些 模型。 这里汇总 多模态尤其 级方向的一些论文和工作。主

金天