【MoLE：通过分层专家混合解码缓解大型视觉语言模型中的幻觉问题】

——【MoLE:Decoding by Mixture of Layer Experts Alleviates Hallucination in Large Vision-Language Models】

1 相关资源

pdf: https://ojs.aaai.org/index.php/AAAI/article/view/34056
ppt:
短视频:
数据集:
源码: https://github.com/Rainlt/MoLE
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: AAAI 2025
【给出具体会议名称和年限，不要仅仅写 ACM，IEEE】
论文类别: Large Language Model
【论文的类别，比如移动计算、轨迹处理、深度学习等】
论文关键字: LLMs, Hallucination Mitigation
推荐程度：3 （其他说明可标注）
（5 非常棒，建议认真研读、小组讨论和复现；4 好，建议细读，考虑复现；3 可以，部分内容值得注意；2 一般，简单浏览即可；1 没有意义，不建议阅读）

3 工作团队

作者: Tian Liang, Yuetian Du, Jing Huang, Ming Kong, Luyuan Chen, Yadong Li, Siye Chen, Qiang Zhu
单位:

1. Zhejiang University
2. Beijing Information Science and Technology University
3. Ant Group

团队情况描述:

4 论文介绍

(1) 研究目的

【研究背景是什么？本文工作有什么用?】

大型视觉-语言模型（LVLM）已能在图像描述、视觉问答等任务中生成流畅且上下文相关的文本，但在医疗、自动驾驶、机器人等高可靠性场景中，模型经常“脑补”出图像中并不存在的物体（即对象幻觉），导致错误决策。

本文的工作旨在通过一种新的混合解码的方法来减少 MLLMs 中的幻觉现象，从而提高模型在多模态任务中的准确性和可靠性。通过这种方法，本文希望为多模态大语言模型的发展提供一种新的思路，使其能够更好地理解和生成与视觉输入一致的文本信息。

(2) 研究现状

【当前的最好研究做到什么程度了？存在的问题是什么？这里采信论文的说法，可以给出自己的评点】

当前，针对 LVLMs 中的幻觉问题，研究者们已经探索了多种方法。其中，对比解码技术是一种主流的解决方案，它通过对比“业余模型”和“专家模型”的输出来过滤幻觉内容。这些方法虽然在一定程度上减少了幻觉，但存在一些局限性。例如，依赖于较弱的业余模型可能无法提供最准确的输出，而且这些方法通常会增加推理时的计算开销。此外，现有方法在处理复杂多模态任务时，仍然难以完全避免幻觉现象。

本文指出，现有方法的一个关键问题是，它们主要依赖于单一的专家模型，而忽略了模型内部不同层可能具有的不同专长和信息。这限制了模型在生成过程中对不同信息的综合利用，从而影响了输出的准确性和可靠性。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文提出了一种新的解码方法 Mixture of Layer Experts，它通过利用 LVLMs 内部不同层的专家层进行协同解码，从而减少幻觉。

(4) 创新与优势

【本文的创新之处是什么？新场景？新发现？新视角？新方法？请明确指出】

【本文工作的贡献或优点是什么?】

1. 将专家混合 MoE 概念引入 LVLM 解码过程，从传统的基于业余模型的对比解码转向协作式分层方法，显著提升输出忠实度。
2. 开发新型分层专家解码方法 MoLE，利用来自 LVLM 不同层的最终专家、第二意见专家和提示保留专家，结合定制门控机制提升解码准确性。
3. 通过在两个多模态幻觉基准上对三个领先 LVLM 进行广泛实验，验证 MoLE 能显著减少幻觉，同时保持计算效率且不依赖额外工具。

(5) 解决思路

【本文是怎样解决问题的？包括方法、技术、模型等，以自己理解的方式表述清楚】

作者提出无需训练的层专家混合解码方法 MoLE。该方法采用启发式门控机制动态选择 LVLMs 的多个层级作为专家层：最终决策专家、第二意见专家和提示保留专家。通过专家协作，MoLE 增强了生成过程的鲁棒性和忠实度。

该方法识别三个关键专家：来自最后一层的最终专家（负责优化最终输出）；选自最后几层的第二意见（SO）专家（提供参考性替代见解）；以及来自最佳保留原始提示层的提示保留（PR）专家（确保输出忠实于输入视觉和指令）。通过单次前向传播协调这些专家，MoLE 以最小计算开销有效减少幻觉。

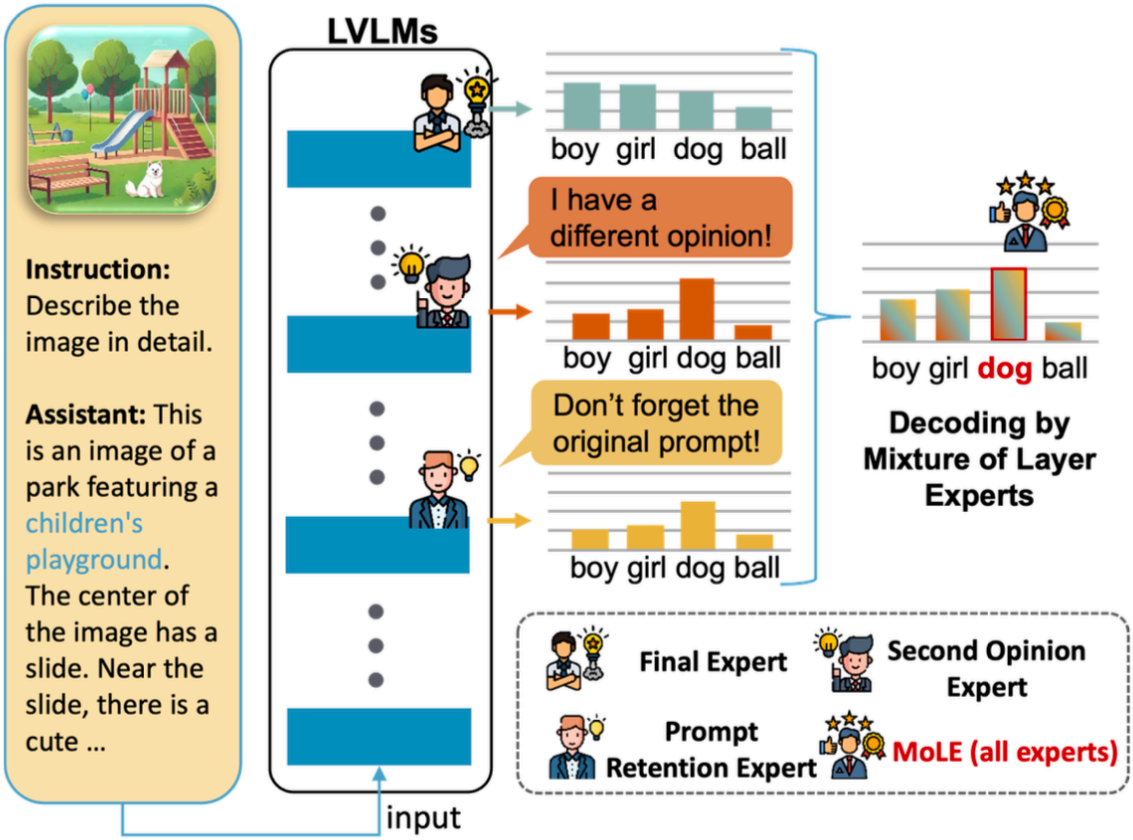

图 1: An illustration of Mixture of Layer Ex- perts(MoLE) Decoding.

第二意见专家

在许多决策过程中，尤其是医学和法律等领域，人们常会寻求第二意见以减少错误并提供更平衡的视角。类似地，作者在 LVLM 框架中引入了第二意见专家。该专家选自最后层中的某一层，这些层已知编码了预训练过程中积累的不同层次的世界知识。

第二意见专家的核心思想是引入一个能提供与最终专家不同视角的层，特别是在关键 token 上保持差异，同时在多数 token 上维持一致性。这是通过评估最终层与候选层的 logits 差异实现的。作者使用 Jensen-Shannon 散度 JSD 来度量这些 logits 分布之间的差异：

$$\begin{aligned} &d(q_N(\cdot \mid PT, x_{<t}), q_j(\cdot \mid PT, x_{<t})) \\ &= JSD(q_N(\cdot \mid PT, x_{<t}) \parallel q_j(\cdot \mid PT, x_{<t})) \end{aligned} \tag{1}$$

第二意见专家的门控机制：为确保 SO 专家提供有价值的意见，作者选择的层需在关键（Top-k）token 上呈现最大差异，同时在多数 token 上与最终专家保持一致。为简化公式，将第 j 层 $q_j(x_{top-k} \mid PT, x_{<t})$ 预测的 top-k token 对应 logits 记为 $q_j^{(top-k)}$，非 top-k 预测的 logits 记为 $q_j^{(majority)}$。因此公式为：

$$M_{top-k} = \arg\max_{j \in \mathcal{N}} d\left(q_N^{(top-k)}, q_j^{(top-k)}\right) \tag{2}$$

$$M_{majority} = \arg\min_{j \in \mathcal{N}} d\left(q_N^{(majority)}, q_j^{(majority)}\right) \tag{3}$$

$$M_{SO} = \begin{cases} M_{top-k}, & \text{if } M_{top-k} = M_{majority} \\ -1, & \text{otherwise} \end{cases} \tag{4}$$

其中 $\mathcal{N} \in \{0, 1, \ldots, 31\}$ 表示 LVLLM 的索引集。如图所示，选定的第二意见专家层 M_{SO} 需满足关键 token 差异最大化而多数 token 差异最小化的标准。该专家生成的 logits 可表示为：

$$q_{SO} = \alpha 1_{\{M_{top-k}=M_{majority}} q_t^{M_{SO}} \tag{5}$$

此处 α 是控制专家层 M_{SO} 强度的比例因子，$1_{\{...}$ 为指示函数（括号内条件为真时取值 1，否则为 0），$q_t^{M_{SO}}$ 表示时间步 t 时 M_{SO} 层的 logits。

提示保持专家

随着序列生成的推进，模型对初始提示的关注度逐渐降低，这会导致幻觉现象，尤其在长序列中更为明显。这是因为随着生成 token 增多，模型逐渐丢失了提示提供的原始上下文。为此作者引入提示保持专家，该专家层专为在序列生成全程维持对提示的高关注度而选定。

提示保持专家的门控机制：作者计算各层对提示 token 的注意力分数总和，选择总和最高的层作为提示保持专家：

$$S_t^{(j)} = \sum_{k=1}^{m} \text{softmax}\left(\frac{(h_t^{(j)} \cdot (hp_k^{(j)})^\top}{\sqrt{d}}\right) \tag{6}$$

$$M_{PR} = \arg\max_{j \in \mathcal{N}} S_t^{(j)} \tag{7}$$

此处 $S_t^{(j)}$ 表示第 j 层对提示 token 的注意力分数总和，M_{PR} 为选定的提示保持专家层。为确保该专家影响力随序列长度增加而增强，作者使用时变权重计算提示保持专家的 logits：

$$q_{PR} = \left(1 - e^{-\frac{t}{\lambda}}\right) \cdot q_t^{PR} \tag{8}$$

其中 q_t^{PR} 表示对应 M_{PR} 的逻辑值，λ 是控制 q_{PR} 随时间增长速率的温度系数。这个随时间变化的权重 $\left(1 - e^{-\frac{t}{\lambda}}\right)$ 会随着序列推进增加提示保留专家的影响力，从而缓解模型偏离原始提示的倾向。

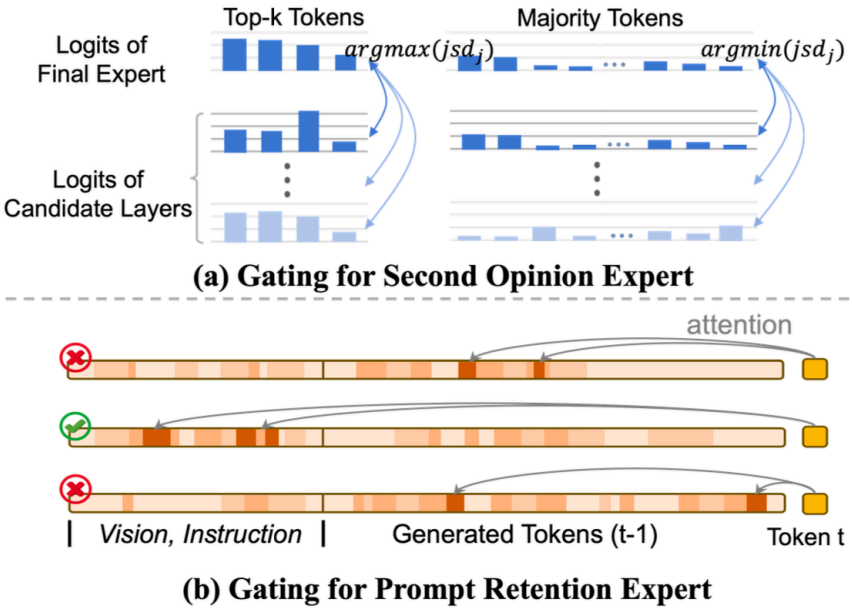

图 2: (a) An illustration of gating mechanism of the Sec- ond Opinion Expert. jsdj represents the JSD between the logits of the Final Expert and the j-th layer. (b) An illustra- tion of gating mechanism of the Prompt Retention Expert. The layer with max attention on vision and Instruction will be selected.

分层专家混合解码

在确定最终专家、第二意见专家和提示保留专家后，我们通过整合它们的逻辑值生成最终预测。与传统对比解码方法（从业余模型逻辑值中做减法）不同，我们的方法对这些专家层逻辑值做加法运算，充分发挥其互补优势：

$$p_{\text{mole}} = (q_F + q_{SO} + q_{PR}) \tag{9}$$

该框架中，最终专家作为主要预测来源，整合了最全面的信息综合。第二意见专家针对关键标记提供批判性视角，确保模型在必要时考虑替代性解释。而提示保留专家则确保模型始终忠实于原始提示，尤其在序列延长时。这种协作式解码方法通过整合 LVLM 内多个层级的专业知识，有效减少幻觉现象。此外该方法计算高效，仅需单次前向传播，且无需额外训练或外部工具。因此 MoLE 为提升 LVLM 输出可靠性提供了实用而强大的解决方案，适用于各类复杂多模态任务。**实验细节**

MoLE 实现中，最终专家选定为模型末层 (N=32)，第二意见专家从未三层 ($L \in \{29, 30, 31\}$)

k　　$\alpha = 0.5$ 作为第二意见专家的缩放因子。提示保留专家的温度系数 λ 设为 100。

(6) 可改进的地方

【本文工作的局限性是什么？你觉得可以从哪些方面改进工作?】

MoLE 目前主要针对 LVLMs 的解码过程进行优化，未来可以探索如何将这种方法扩展到模型的训练阶段，以进一步提高模型的整体性能。

(7) 可借鉴的地方

【你觉得本文哪些方面可以借鉴？比如思路、方法、技术等】

从思路层面，本文提出的利用模型内部不同层的专长进行协同解码的思路，为解决类似问题提供了一种新的视角。这种方法可以启发研究者在其他领域中探索如何充分利用模型内部的结构和信息。

(8) 其他收获

【你有什么其他收获吗？比如了解了哪些团队和大牛在某领域做得很好，某类问题通常用什么技术解决，某些技术之间存在什么样的关联，某些会议和期刊在某领域很知名……】

我了解到对比解码技术和模型内部层结构在解决幻觉问题中的重要作用，这些技术之间的关联和相互补充，为未来的研究提供了新的方向。

5　评阅人

姓名:

时间: