【通过注意力因果解码减轻多模态大型语言模型的幻觉现象】

——【Seeing Far and Clearly: Mitigating Hallucinations in MLLMs with Attention Causal Decoding】

1 相关资源

pdf: https://arxiv.org/pdf/2505.16652
ppt:
短视频:
数据集:
源码: https://mllms-farsight.github.io/
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: CVPR 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: Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zelin Peng, Zhiwei Yang, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, Zongyuan Ge
单位:

1. Monash University
2. MBZUAI
3. XJTLU
4. Shanghai Jiaotong University
5. Fudan University
6. University of Minnesota
7. Cornell University

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。

本文的工作旨在缓解 LVLMs 中的幻觉问题，提高模型生成内容的准确性和可靠性，从而增强用户对模型的信任，使其能够更好地应用于实际场景。

(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】

当前的研究中，一些方法通过外部知识检索或鲁棒的指令微调来减少幻觉，但这些方法往往需要额外的训练成本。另一些方法则关注无训练的解码策略，如对比解码和自校准注意力，旨在减少对语言先验的过度依赖。然而，这些方法并没有深入分析多模态 token 之间的交互过程以及幻觉产生的原因。例如，尽管一些方法在减少初始幻觉方面取得了一定效果，但在视频字幕生成等任务中，雪球幻觉（即基于初始幻觉进一步扩展的幻觉）的比例仍然很高。这表明现有的方法在处理多模态 token 交互和幻觉成因方面仍有不足。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文解决的核心问题是多模态大语言模型中的幻觉现象，通过优化因果掩码来减轻幻觉，同时提高模型对上下文信息的利用效率。

(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. 分析自注意力 token 传播模式，揭示多模态大型语言模型中幻觉现象的两大成因：注意力坍缩与位置信息衰减。
2. 提出即插即用型解码策略 FarSight，仅通过调整因果掩码，即可有效缓解由上述问题引发的幻觉现象。
3. 在图像和视频任务上的大量评估表明，FarSight 性能优越，为减轻幻觉现象提供了有效的解决方案。

(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】

在本研究中，作者假设 token 之间的交互不足可能导致对异常值 token 的过度依赖，从而忽略了密集且信息丰富的上下文线索。在这项工作中，作者认为，对 token 交互过程进行有效干预能够增强上下文内推理。此外，现有的因果掩码优化方法主要是改善 token 交互，并以单模态文本外推为目标。相比之下，作者的 FarSight 通过增强多模态大语言模型（MLLMs）中的视觉-语言 token 交互，明确地解决了多模态幻觉问题。

为了更深入地探究这一现象，作者对解码过程中的注意力图进行了分析，并发现了导致幻觉的两个

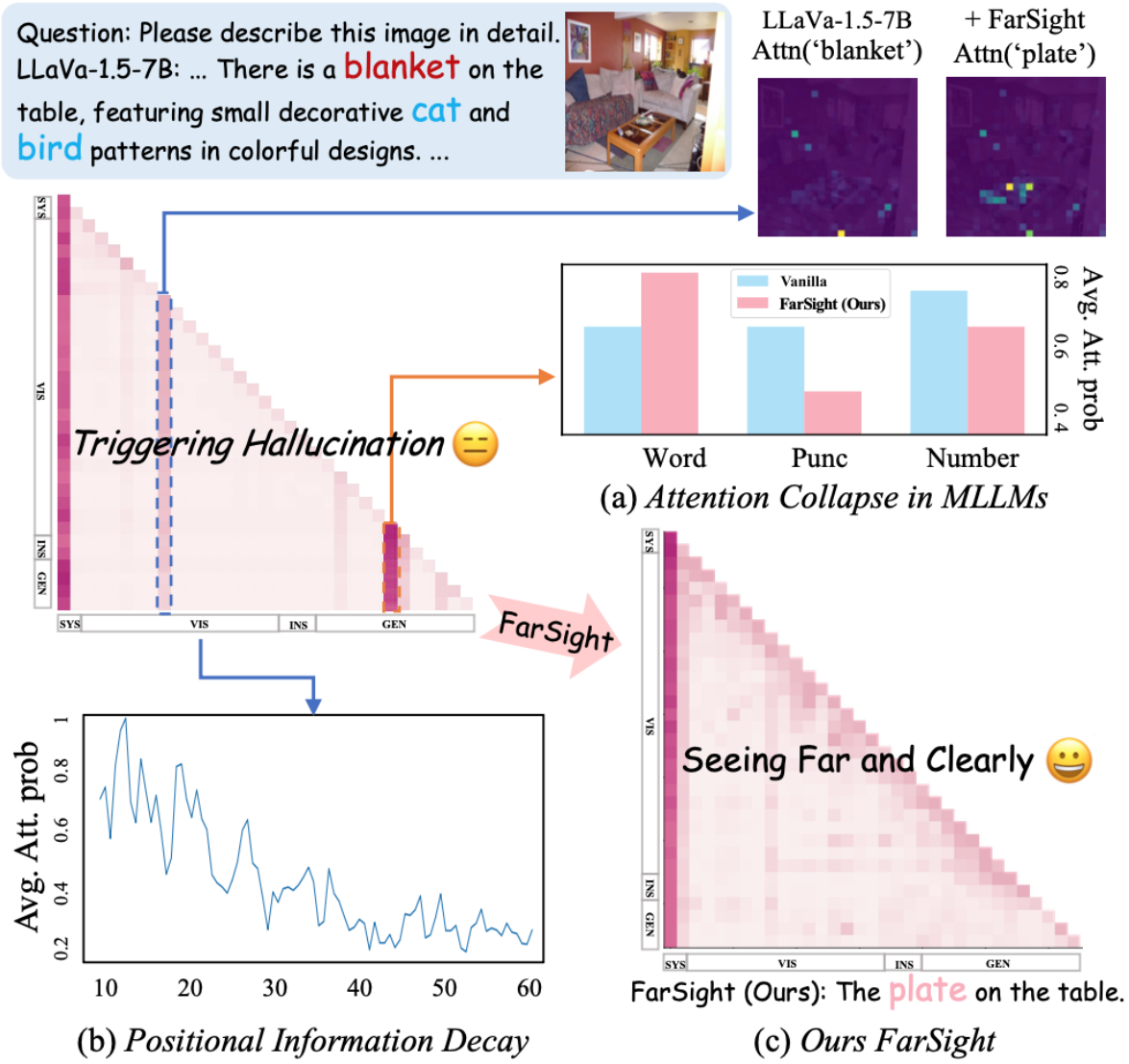

图 1: (a): Attention Collapse in MLLMs (b): Positional Information Decay (c): Our FarSight.

问题。

(i) 多模态大语言模型（MLLMs）中的注意力坍缩：作者观察到模型往往会将过多的注意力分配给信息含量有限的标记。这些信息含量低但关注度高的异常标记，例如视觉背景和文本符号，会干扰相关信息的有效传播。这一问题的产生是因为 softmax 注意力机制要求所有注意力分数非零且总和为 1，这使得即使是信息含量低或非优先的标记也会获得过多的注意力。注意力坍缩类似于 Opera 中关于“总结标记”的研究结果，会随着生成文本的延长导致视觉和文本信息的传递逐渐减弱。

(ii) 位置信息衰减：如图 (b) 所示，发现在整个生成过程中，对密集视觉信息的注意力逐渐下降。这是由于旋转位置编码（RoPE）的长期衰减特性，无法提供足够的位置信息来确保视觉标记和文本标记之间的充分交互。随着相对距离的增加，视觉标记信息的流动信息会逐渐减少，从而可能导致幻觉。

因此，研究结果表明，维持平衡的信息传播并改进位置编码，能够缓解注意力坍缩和位置信息衰减，这两者都会引发幻觉。

在这项研究中，作者提出了 FarSight，这是一种通用的即插即用解码策略，通过优化因果掩码来减少异常值标记带来的注意力干扰。具体而言，作者在因果掩码的上三角矩阵中初始化了一组注意力

寄存器，以捕捉转向异常值标记的注意力。这些注意力寄存器保留了因果解码特性，确保不会过早获取来自未来标记的信息。此外，作者设计了一种动态寄存器-注意力分配机制，在每个解码步骤中显式优化注意力分配，以实现稳健的上下文内推理。

方法的核心是优化标记的有效传播。对具有多模态信息内容的标记的注意力分布进行调制，以改善标记传播。此外，RoPE 编码的相对位置限制导致在上下文交互过程中视觉到文本标记信息的传输不足，这削弱了位置感知能力。因此，在因果掩码中引入了逐渐降低的掩码率来编码绝对位置信息，使模型能够关注更远距离的先前标记，特别是在视频序列任务中。

注意力寄存器

传统因果掩码是“下三角矩阵”（只有当前 token 及之前的位置能被关注，未来位置全是 0 或 -∞），FarSight 在这个矩阵的“上三角部分”（原本被屏蔽的未来位置）加了一个“注意力寄存器”。定义注意力寄存器：

$$\mathcal{P}_i = [0, 0, \cdots, 0, \underbrace{\mathcal{P}_{i,i+1}, \mathcal{P}_{i,i+2}, \cdots, \mathcal{P}_{i,n}}_{n-i}]_n, \tag{1}$$

新的注意力矩阵：

$$\mathbf{W} = \omega \cdot \mathbf{C} + \mathbf{P}, \quad \text{where} \quad \mathbf{C} = \text{tril}(\mathbf{1}_{n\times n}), \tag{2}$$

P 矩阵的规则定义：

$$\mathcal{P}_{i,j} = -(j-i) \cdot \sigma, \quad \forall j > i, \tag{3}$$

距离越远，格子能“吸收”的多余注意力越少，符合“近未来信息关联性稍强，远未来关联性弱”的逻辑。

位置感知编码

这一步是“先存多余注意力，再确保因果性”，核心是让 softmax 后的注意力分数“不强制总和为 1”（只对历史 token 的分数做归一化，仓库里的分数不算进来），具体分两步：

第一步：先把原始注意力分数里的“多余部分”存到 P 仓库（解决注意力坍缩）

第二步：先对括号里的结果做 softmax（此时历史 token 的注意力分数被优化，低信息 token 的分数降低），再乘一次 C——这一步是“双重保险”，确保未来位置的注意力分数全为 0，不破坏因果性，同时让早期 token 的注意力分数“不被过度稀释”。

$$\bar{\mathbf{W}} = \text{SoftMax}(\omega \cdot (\mathbf{C} + \mathbf{P}) \cdot \mathbf{C}). \tag{4}$$

(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?】

虽然 FarSight 通过优化因果掩码减少了异常 token 的注意力干扰，但在处理更复杂的多模态交互时，可能需要进一步优化注意力机制以更好地处理不同模态之间的信息流动。

(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】

本文的思路和方法在多方面值得借鉴。首先，通过分析 token 传播模式来揭示幻觉成因的方法为理解 MLLMs 的行为提供了新的视角。其次，FarSight 中提出的注意力寄存器和位置感知编码方法为优化因果掩码提供了新的思路，这些方法可以应用于其他需要减少幻觉或增强上下文推理的场景。

(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……】

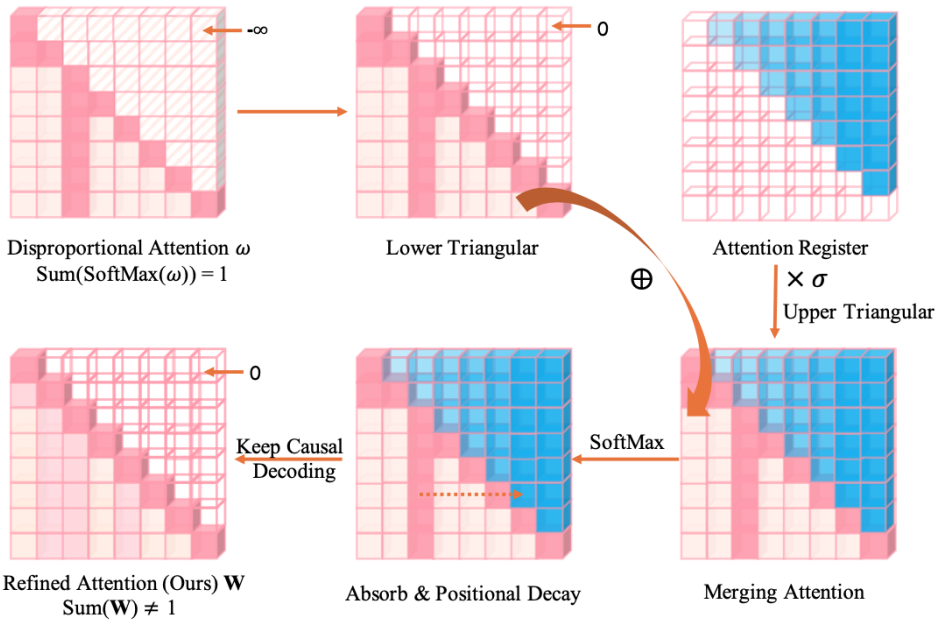

图 2: The scheme of the proposed FarSight strategy, which integrates with the softmax operation, replacing the traditional causal mask.

本文展示了因果掩码优化在减少幻觉方面的潜力，这为未来的研究提供了新的方向。

5 评阅人

姓名:
时间: