【通过过度信任惩罚与回溯分配缓解多模态大语言模型中的幻觉现象】

——【OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Alloc】

1 相关资源

pdf: https://arxiv.org/pdf/2311.17911
ppt:
短视频:
数据集:
源码: https://github.com/shikiw/OPERA
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: CVPR 2024
【给出具体会议名称和年限，不要仅仅写 ACM，IEEE】
论文类别: Multi-modal 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: MLLMs, 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: Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu
单位:

1. Anhui Province Key Laboratory of Digital Security University of Science and Technology of China
2. Shanghai AI Laboratory
3. The Chinese University of Hong Kong

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?】

本文主要研究多模态大模型 MLLMs 在实际应用中面临“幻觉”的问题。MLLMs 能够处理图像和文本等多种模态的信息，展现出强大的交互能力，但它们往往会生成与用户提供的图像或提示不符的错误内容。这种幻觉问题严重影响了 MLLMs 在需要精确判断的现实场景中的可信度和可用性，如自动驾驶辅助系统。

因此，本文的工作旨在提出一种新的方法来缓解 MLLMs 的幻觉问题，以提高其在实际应用中的可靠性和准确性。

(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，为减少 MLLMs 的幻觉问题，已经提出了一些方法。这些方法主要包括在训练阶段使用特定设计的数据进行训练，或者在推理阶段从其他来源引入外部知识。

这些方法虽然在一定程度上能够缓解幻觉问题，但都不可避免地引入了额外的成本，如需要额外的标注数据用于训练，或需要整合额外的知识和模型，这些方法需要大量的资源和计算成本，限制了其广泛引用。

此外，这些方法可能无法完全解决幻觉问题，因为它们没有从根本上解决 MLLMs 在生成过程中对某些信息的过渡信任和对其他信息的忽视这一核心问题。因此需要进一步探索更高效、更经济、更有效的方案。

(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 OPERA，该方法基于“过渡信任惩罚”和“回溯分配”策略，在不引入额外数据、知识和训练成本的情况下，缓解多模态大模型在推理阶段的幻觉问题。

(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. 提出了新的解码方法 OPERA，该方法基于“过渡信任惩罚”和“回溯分配”策略。
2. 通过观察 MLLMs 的自注意力矩阵中的知识聚合模式，发现幻觉与模型对某些汇总信息的过渡信任密切相关。
3. 基于以上发现，在束搜索解码过程中引入了一个惩罚项，以降低信任模式的候选词被选择的概率，并通过回溯策略在必要时重新分配词的选择。
4. 这种方法不需要引入额外的训练数据和外部知识，在显著减少生成幻觉内容的同时，保持生成文本的质量和详细性。

(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】

OPERA 方法的核心在于对 MLLMs 的解码过程进行优化，而不是对模型架构本身进行修改。它通过在解码阶段引入惩罚和回溯机制，动态调整生成词的选择，从而减少幻觉内容的生成。这种方法

的优点是它不需要额外的训练数据或模型调整，可以直接应用于现有的 MLLMs。

过渡信任惩罚

过度信任惩罚的核心思想是通过检测模型在生成文本时对某些汇总信息的过度依赖，并对这种依赖施加惩罚，从而降低生成幻觉内容的概率。

在 MLLMs 的生成过程中，模型往往会通过自注意力机制将前面的上下文信息汇总到某些关键的“汇总词”上，这些词在后续生成中起重要的引导作用。但是这种过度依赖会导致模型忽视其他信息，从而产生幻觉。

为了检测这种知识聚合模式，OPERA 在解码过程中分析自注意力矩阵。具体来说，OPERA 会考察当前生成的词与前面生成的词之间的注意力权重，如果某个词的注意力权重在后续多个词中都表现得异常高，则说明模型对该词存在过度信任。

为了减少过度信任带来的幻觉，OPERA 在束搜索解码过程中引入了一个惩罚项。在束搜索中，模型会维护多个候选序列，并在每一步选择最有可能的候选词。OPERA 通过计算一个“过度信任分数”，并将其与模型的原始预测分数 logits 结合，从而调整候选词的优先级。具体公式如下：

$$p(x_t \mid x_{<t}) = \text{Softmax}[H(h_t) - \alpha\phi(w_{\leq t})]_{x_t}, \quad \text{其中 } x_t \in Y \tag{1}$$

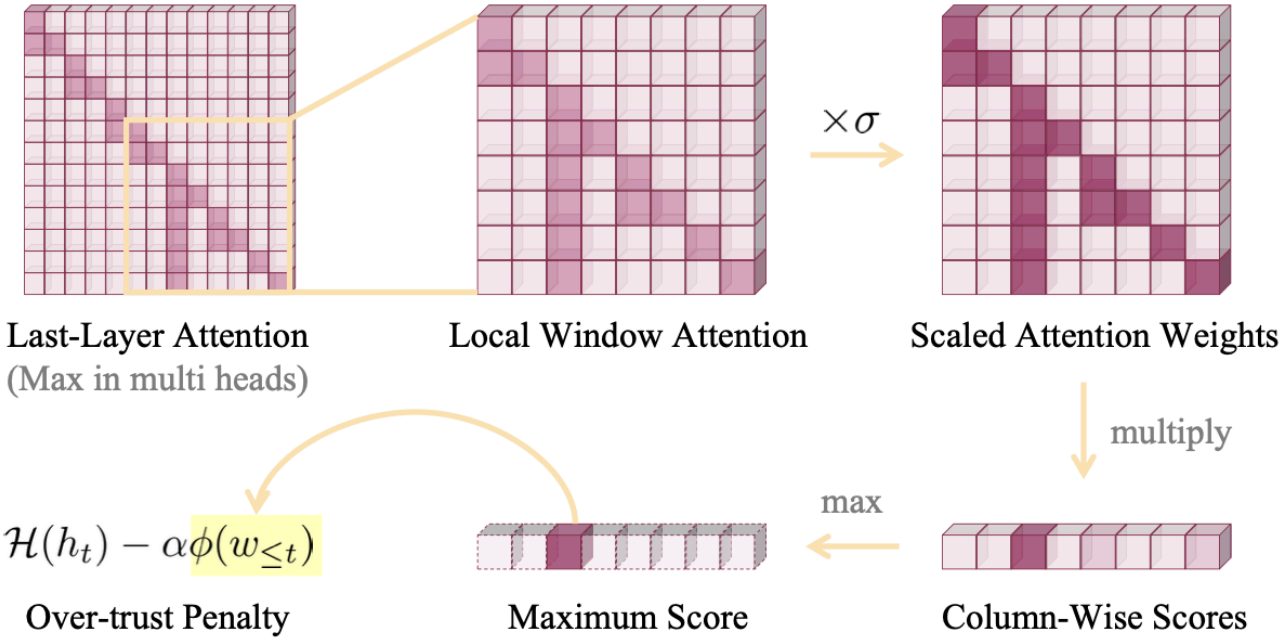

图 1: The scheme of calculating the proposed over-trust penalty term.

回溯-分配策略

尽管过度信任惩罚可以在一定程度上减少幻觉，但在某些情况下，模型可能仍然会生成幻觉内容。为了进一步纠正这种幻觉，OPERA 引入了回溯-分配策略：当检测到过度信任模式时，模型会回溯到生成幻觉内容的起始位置，并重新选择后续的词。

回溯-分配策略的触发条件是基于过度信任分数的位置一致性。具体来说，OPERA 会计算连续多个词的最大过度信任分数的位置坐标，并检查这些位置是否一致。如果这些位置坐标在连续多个词中都指向同一个汇总词，那么模型认为此时生成的幻觉内容已经不可避免，需要触发回顾机制。具体公式如下：

$$N_{\text{overlap}} = \sum_{c \in C} 1_{c=s}, \quad \text{其中 } s = \text{Mode}(C) \tag{2}$$

一旦触发回溯机制，模型会回溯到汇总词的位置，并从候选词集合中重新选择后续的词。为了避免重复选择相同的词，OPERA 会在重新选择时排除之前已经选择过的词。此外，为了避免无限回溯，模型会限制每个位置的回溯次数。具体来说，如果某个位置的回溯次数已经达到最大值 ，则模型会跳过该位置，继续向前回溯。

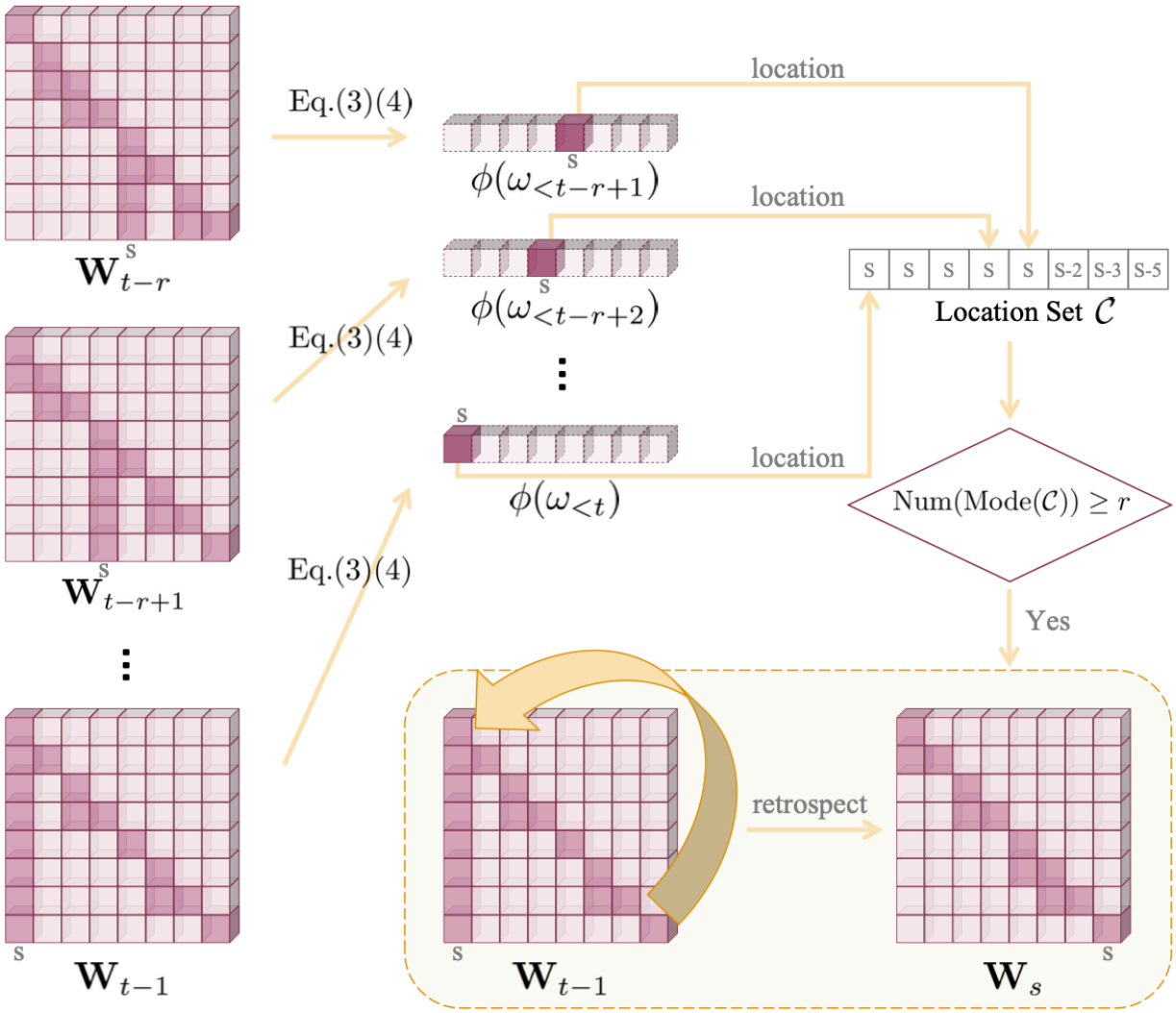

图 2: The scheme of the proposed Retrospection strategy.

(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?】

该方法存在两个主要局限：

1. 无法解决 MLLMs 所有类型的幻觉：例如，当 MLLMs 对某些概念存在强烈的固有偏见时，或者当模型的视觉感知能力不够强大时，OPERA 可能无法有效地纠正这些幻觉。
2. 处理短答案幻觉时收益有限：主要因知识聚合模式的滞后性。OPERA 更擅长处理长序列中的幻觉，改进方向是提升知识聚合模式检测指标的灵敏度。

(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】

本文对 MLLMs 幻觉问题的深入分析和对知识聚合模式的观察为理解模型生成幻觉的原因提供了新

的视角，这可以启发从不同的角度探索模型的生成机制和潜在问题。同时，本文的实验设计和评估方法也提供了一个全面评估模型幻觉问题的框架，可作为对比方法进行比较。

(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，幻觉问题通常是通过训练数据的优化、外部知识的引入或解码策略的改进来解决的，而本文提出的方法为解码策略的改进提供了一个新的方向。

5 评阅人

姓名:

时间: