【动态校正解码缓解多模态大模型的幻觉现象】

——【MLLM CAN SEE? DYNAMIC CORRECTION DECODING FOR HALLUCINATION MITIGATION】

1 相关资源

pdf: https://arxiv.org/pdf/2410.11779
ppt:
短视频:
数据集:
源码: https://github.com/zjunlp/DeCo
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: ICLR 2025
【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，Dynamic Correction Decoding
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: Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, Huajun Chen
单位:

1. Zhejiang University
2. Nanjing University of Aeronautics and Astronautics
3. National University of Singapore, NUS-NCS Joint Lab, Singapore
4. Zhejiang Key Laboratory of Big Data Intelligent Computing

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?】

近期多模态大模型 (MLLMs) 的快速发展展现了通用人工智能 (AGI) 的潜在路径，然而在实践中，MLLMs 的发展受困于幻觉现象，即模型常生成关于不存在的图像的描述，却忽略可见物体，本质上是在自我欺骗。这种现象在医学领域、自动驾驶等高风险领域可能造成不可逆的后果。

因此，本文旨在探索 MLLMs 产生幻觉的内部机制，并提出一种有效的解决方案，以减少幻觉现象，提高模型在多模态任务中的准确性和可靠性。

(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】

当前的研究已经对 MLLMs 中的幻觉现象进行了一定的探讨。一些研究发现，MLLMs 在早期层能够更好地处理视觉信息，但在深层中，语言模型的先验知识可能会抑制视觉信息的表达，从而导致幻觉。例如，Grad-CAM 可视化显示，图像-文本交互在早期层存在，但在深层中消失。OPERA 提出“聚合模式”会导致幻觉，即视觉信息在早期层被逐渐聚合到锚点标记上，而在预测时仅关注这些标记，忽略其他视觉信息，从而导致生成序列中幻觉的概率增加。

然而，现有研究对 MLLMs 幻觉机制的理解仍然有限，尤其是关于视觉信息是否被正确识别，以及是否被后续信息流抑制的问题尚未明确。此外，现有的幻觉缓解方法大多依赖于数据增强、对齐训练或后处理策略，这些方法要么需要大量的标注数据和计算资源，要么会增加推理成本和延迟。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文探究了多模态大模型幻觉产生的成因，并且提出了动态校正解码方法，以此来通过高效且无需额外训练的方法，减少 MLLMs 在生成过程中产生的幻觉现象，同时保持模型的生成性能和多样性。

(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. 本文通过实验证明 MLLMs 在前置层能够识别视觉对象，但这种识别在后续层中被语言模型的先验知识所抑制。
2. 基于以上发现，作者提出动态校正解码方法 DeCo。DeCo 通过动态选择前置层，并将其知识融入最终层的输出中，调整输出的 logits，从而减少幻觉。这种方法不仅简单高效，而且与现有的解码策略（如贪婪搜索、核采样和束搜索）兼容，并且可以无缝应用于不同的 MLLMs。

(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】

发现 1：MLLMs 在一定程度上知道对象是否存在

作者将探究 MLLMs 在图像描述任务中对物体的理解机制简化为判定函数 isexist(obj)，即判定图像中是否存在某物体，然后在有 7B 参数、32 层的语言模型的组件各层末端进行探测试验。

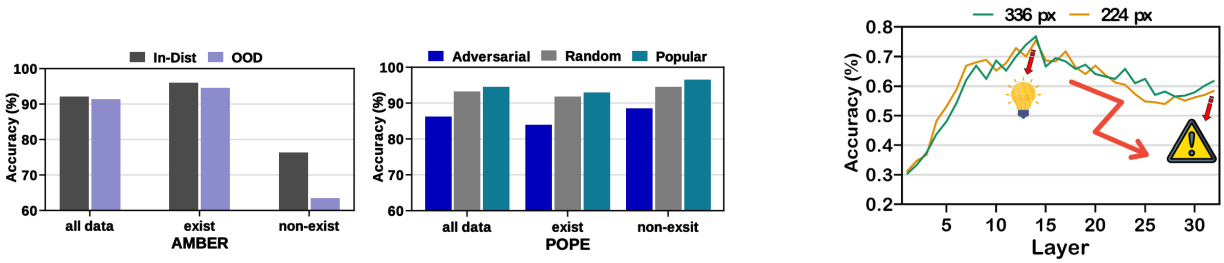
作者采用 prompt 模版：“USER: <image> Describe the image. ASSISTANT: The image contains obj.”，将训练集和测试集格式化后输入到 MLLMs 中，在每层隐藏状态输出到末位训练探测分类器，

共得到 32 个分类器。

实验结果如图 1 所示，通过比较所有物体/存在物体/不存在物体的准确率，结果显示：MLLM 对图像描述中正确生成的物体具有高准确率，尽管会产生大量不存在物体，探测实验仍保持约的准确率，这表明 MLLMs 对图像中物体的存在性具有一定程度的认知。

同时，探测试验还表明 MLLMs 在前置层的识别能力更强，准确率随着层数的增加而逐渐下降。这表明 MLLMs 在前置层能够更好地捕捉视觉信息，但在深层中这种能力被削弱。

通过进一步实验，作者还发现提高视觉编码器的分辨率（从 224px 增加到 336px）可以增强 MLLMs 在前置层对不存在对象的识别能力。这进一步证明了前置层的视觉信息对幻觉现象的影响。



(a) Object probing results.

(b) Different resolution results.

图 1: Overall results of the probing experiment with MLLMs.

发现 2：语言模型的先验知识抑制了 MLLMs 已经看到的视觉信息

作者假设了 MLLMs 在前置层能够有效捕捉视觉信息，但由于语言模型的先验知识，这些视觉信息在深层中被抑制。为了验证这一假设，作者通过分析 MLLMs 在不同层的输出概率分布，研究了真实标记和幻觉标记的概率变化。

作者从 MSCOCO 数据集中随机选择了 500 张图像，并使用随机提示生成原始响应。然后，他提取了所有不存在的对象及其对应的前文，并重新输入到 MLLM 中，观察模型在预测下一个标记时的概率分布变化。具体来说，作者计算了每个层的输出概率分布，并通过截断词汇表（类似于 Top-p 采样）来获取候选标记。

实验结果如图 2，结果显示，真实标记在前置层具有较高的概率，但在深层概率显著下降，尤其在最终层，幻觉标记的概率超过了真实标记。

为了进一步验证先验知识的影响，作者在没有输入图像的情况下生成了候选标记。结果表明，即使没有视觉信息，MLLMs 仍然倾向于生成幻觉标记。这表明 MLLMs 的先验知识可能会抑制真实标记的概率，从而导致幻觉现象。

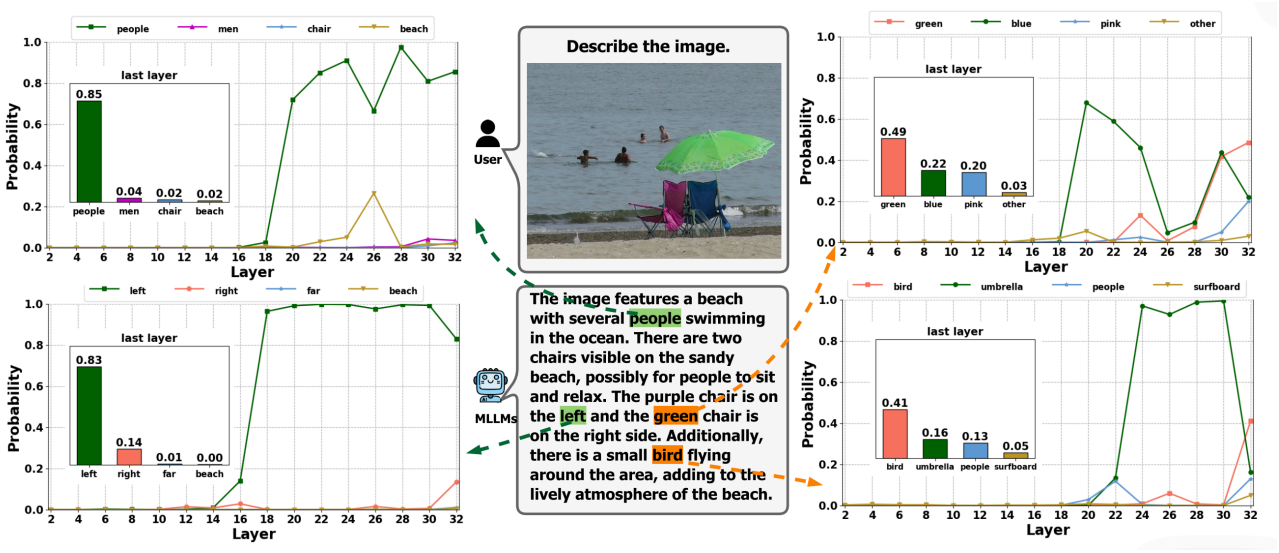


图 2: Illustration of token probabilities across transformer layers.

动态校正解码 DeCo

基于之前探究 MLLMs 生成幻觉的原因，作者提出了基于前置层知识的动态校正解码 (DeCo)，可在推理阶段缓解幻觉现象。

图 3 为 DeCo 的模型结构图，它主要分为两部分：动态前置层选择和基于前置层知识的解码校正。首先是动态前置层选择，该部分主要分为两部分：

- 候选标记获取：由于词汇空间庞大，所以作者仅追踪不同层间排名靠前的标记作为候选标记，采用 Top-P 截断策略选择候选标记。

$$\mathcal{V}_{\text{candidate}}\left(x_T \mid x_{<T}\right)=\left\{x_T \in \mathcal{V}: \sum_{v \in \mathcal{V}_p} P_\tau\left(x_T=v \mid x_0, x_1, \ldots, x_{T-1}\right) \leq p\right\} \tag{1}$$

- 前置层选择：DeCo 的核心在于动态选择一个“锚点层”，该层的标记概率最高。作者发现，真实标记在早期层的概率通常高于幻觉标记。因此，模型在指定的层区间 [a,b] 内，计算每个候选标记在每个层的概率，并选择概率最高的层作为锚点层。

$$\mathcal{A}=\operatorname{argmax}_i\left\{x_T \in \mathcal{V}_{\text{candidate}}: \operatorname{softmax}\left(\phi\left(h_{T-1}^i\right)\right)_{x_T}, i \in[a, b]\right\}, \tag{2}$$

第二部分为基于前置层知识的解码校正，主要分为两部分：

- 动态软调节：作者引入了一个动态调节系数，具体如下，该系数有助于防止逻辑值的剧烈波动，尤其是在前置层中候选标记间的概率差异很小的时候。

$$\max \operatorname{prob}=\max\left(\operatorname{softmax}\left(\phi\left(h_{T-1}^{\mathcal{A}}\right)\right)\right). \tag{3}$$

- 基于前置层知识的解码：选定某个前置层后，作者将该层和最终层的信息整合起来以修正逻辑值的分布，通过一个超参数来控制前置层信息的融合比例，并采用动态软调节技术保留原始模型的生成风格。

$$\widehat{p}\left(x_T \mid x_{<T}\right)=\operatorname{softmax}(\text{ logits })_{x_T}, \tag{4}$$

$$\text{logits }=\phi\left(h_{T-1}^N\right)+\alpha \times \text{ max\_prob } \times \phi\left(h_{T-1}^{\mathcal{A}}\right) \text{ , } \tag{5}$$

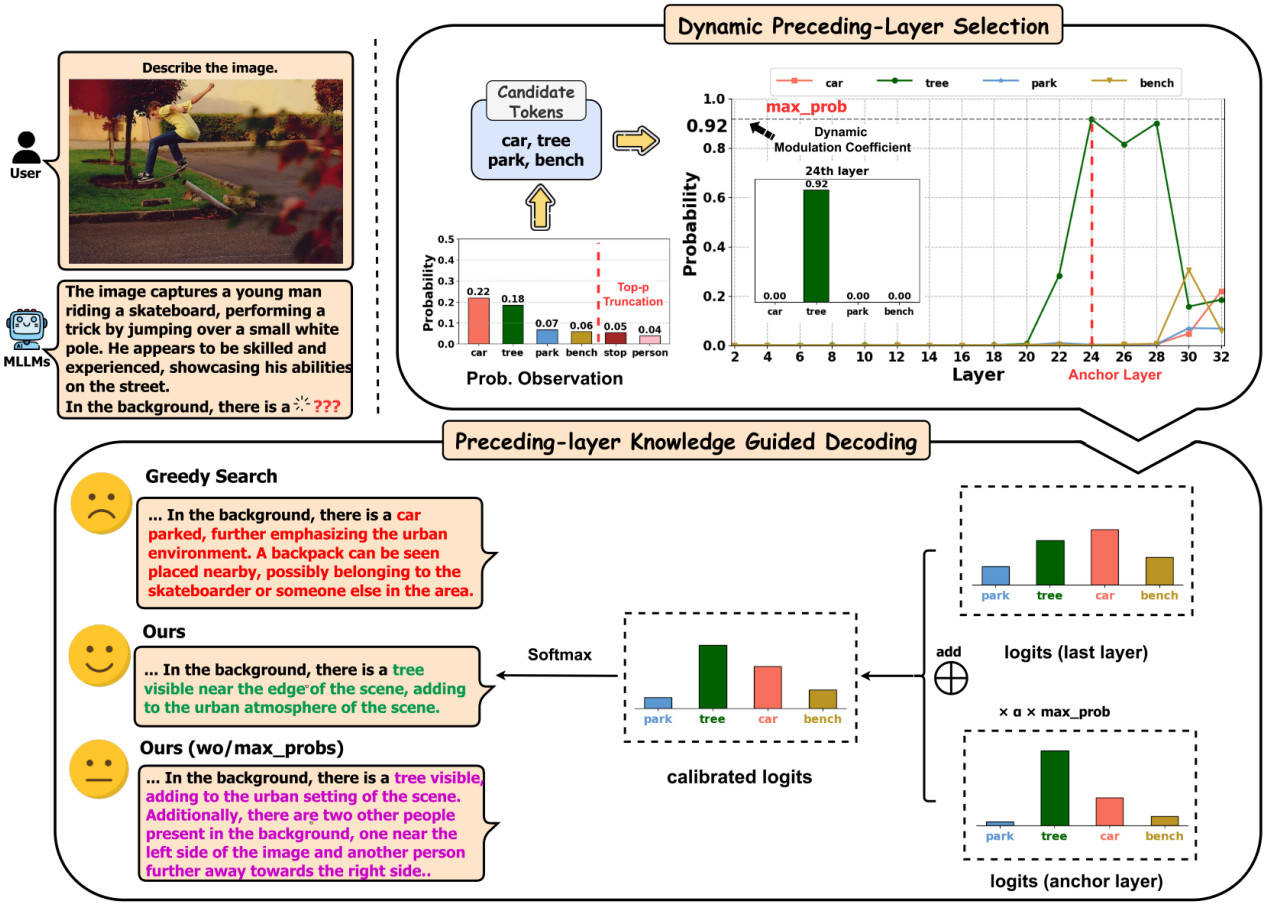


图 3: Framework of DeCo.

作者将 DeCo 与多种解码方法结合，并与多个幻觉缓解的基线对比，发现 DeCo 的方法均优于其他方法，且在推理效率和简洁性上更具有优势。

(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?】

DeCo 中的超参数 和锚点层的选择是基于实验确定的，用在不同的任务和模型时需要进行调整，如果可以自动化调整的话会更好。

虽然 DeCo 有效地减少了幻觉现象，但对于 MLLMs 产生幻觉的深层次原因感觉可以有更深入的研究。

(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】

DeCo 方法提供了一种有效的策略来利用模型前置层的知识来校正最终层的输出，这种思想可以应用于其他类型的模型和任务。

作者通过许多精心设计的实验来验证方法的有效性，这种严谨的实验设计和评估方法也值得学习

作者将复杂问题分解为多个可管理的子问题（如候选标记获取、动态前置层选择等），这种分解问题的方法我认为有助于更清晰地理解和解决问题，值得借鉴。

(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……】

通过这篇论文，我了解到 Transformer 架构在处理长序列时可能存在的信息丢失问题，以及如何通过动态解码策略来缓解这一问题。

同时，我也学习到如何通过实验分析来深入理解模型的内部机制，以及如何基于这些理解来设计和改进模型。

5 评阅人

姓名:

时间: