【基于视觉感知注意力头差异解码 LVLM 幻觉机制】

——【Cracking the Code of Hallucination in LVLMs with Vision-aware Head Divergence】

1 相关资源

pdf: https://arxiv.org/pdf/2412.13949
ppt:
短视频:
数据集:
源码: https://github.com/jinghan1he/VHR.
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: ACL 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: Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, Jinqiao Wang
单位:

1. Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences
2. School of Artificial Intelligence, University of Chinese Academy of Sciences
3. National University of Singapore 4Southeast University 5Wuhan AI Research

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。

本文旨在研究 LVLMs 内部导致幻觉的机制，特别是多头注意力模块，并提出一种新的方法来有效缓解幻觉问题，从而提高 LVLMs 的性能和可靠性，使其在实际应用中更具价值。

(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】

目前，针对 LVLMs 幻觉问题的研究主要分为三类：训练对齐、后处理和解码策略。训练对齐方法通过引入额外的信息或模型来进行对齐训练，但这种方法会增加训练成本；后处理方法则需要在模型生成输出后再进行处理，增加了推理成本。而解码策略主要是在推理过程中调整 logits 分布，如对比解码、束搜索等方法，这些方法虽然能够在一定程度上改善幻觉问题，但它们只是在输出层面进行干预，无法从根本上解决模型内部导致幻觉的问题。

此外，现有研究还发现 LVLMs 倾向于优先使用语言模式，这可能导致生成流畅但不准确的内容，而这种语言偏差与幻觉现象密切相关。尽管已有研究对语言偏差进行了探讨，但大多数方法都是通过操纵输出 logits 来解决，缺乏对模型内部机制的深入分析。

因此，现有研究在解决 LVLMs 幻觉问题上虽然取得了一定进展，但在深入理解和调整模型内部机制方面仍存在不足，无法从根本上有效解决幻觉问题。

(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深入探究 LVLMs 内部导致幻觉的机制，并提出一种有效的方法来缓解幻觉问题，从而提高 LVLMs 生成文本与视觉内容的一致性和可靠性。

(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. 提出 VHD 指标探测 LVLM 注意力头的语言偏见倾向，并通过 T-VHD 指标分析语言偏见生成与 LVLM 幻觉的关联
2. 提出无需训练的 VHR 方法，通过在生成过程中自适应识别并强化关键注意力头来主动缓解幻觉
3. 大量实验表明，VHR 在广泛采用的幻觉基准测试中优于现有解码方法，且额外时间成本可忽略不计

(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】

视觉感知注意力头识别

视觉感知注意力头分歧度 (VHD): 受模型中上下文注意力头与记忆注意力头存在的启发，我们研究

不同注意力头是否对视觉内容表现出显著差异的敏感度。具体而言，我们提出视觉感知注意力头分歧度指标，用于衡量移除图像上下文时生成步骤注意力头 t 输出的变化：

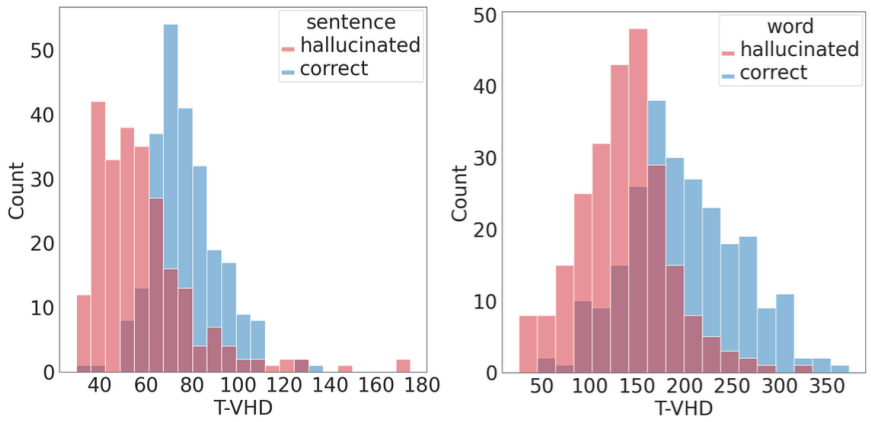
$$\mathrm{VHD}_{l,i} = \left(A_{l,i}\left(y_t \mid y_{<t}, x_V, x_T\right), A_{l,i}\left(y_t \mid y_{<t,x_T}\right)\right) \tag{1}$$

具体而言，作者使用图像和指令" 请详细描述该图像" 提示 LLaVA-1.5 生成描述，计算预测首个标记时的 VHD 分数。结果显示少数注意力头表现出显著较高的 VHD 分数，其余则敏感度较低，这表明存在对视觉信息更敏感的视觉感知注意力头。

标记级 VHD: 除模型内部视觉感知程度差异外，我们进一步探究 VHD 分数是否随标记生成步骤变化。为此，将模型每层最显著注意力头的 VHD 分数聚合为 Token-VHD 指标：

$$\text{T-VHD} = \mathrm{VHD}_{l,i} = \sum_{l}\sum_{i} \mathrm{topk}_i\left(\mathrm{VHD}_{l,i,k}\right). \tag{2}$$

利用该指标，可定量分析 LVLM 幻觉现象与语言偏差在不同粒度层级（句子级和词语级）的关联。为此，在 CHAIR 基准测试中随机选取 500 张图像进行实验，追踪每个生成步骤的 T-VHD 分数。生成描述中物体相关词语根据是否出现在图像标注物体集合中被分类为幻觉或正确，句子则根据是否包含幻觉词语进行标注。图 1 展示实验结果，凸显幻觉实例与正确实例 T-VHD 分数的分布差异，这些发现为语言偏差与 LVLM 幻觉的密切关联提供了统计证据。



(a) Histogram of sentence-level T-VHD scores

(b) Histogram of word-level T-VHD scores

图 1: Relationship between T-VHD scores and hal- lucinations in LVLMs.

视觉感知注意力头强化

由于模型中仅少量注意力头对视觉信息敏感，可在生成过程中放大其贡献以增强模型对视觉线索的依赖并抵消语言偏差。如前所述，VHD 指标能有效捕捉注意力头对视觉信息的敏感度，适合作为选择关键注意力头进行强化的依据。但作者发现部分高 VHD 值源于移除视觉上下文时注意力头激活的激增，表明存在负面视觉敏感性。放大此类注意力头的贡献将背离目标，因此作者将这些异常值置零，即 $\mathrm{VHD}_{l,i} = 0$, 当满足以下条件时：

$$\begin{cases} VHD_{l,i} > \mu\left(VHD_{l,*}\right) + \sigma\left(VHD_{l,*}\right), \\ TA_{l,i} > \mu\left(TA_{l,*}\right) + \sigma\left(TA_{l,*}\right), \end{cases} \quad \text{where} TA_{l,i} = \left\|A_{l,i}\left(y_t \mid y_{<t}, x_T\right)\right\|^2. \tag{3}$$

μ 和 σ 分别表示均值和标准差。接着，在模型给定层的多头注意力模块中，作者根据 VHD 分数筛选前一半注意力头，并将其输出直接放大 α 倍：

$$\tilde{A}_{l,i} = \begin{cases} \alpha \cdot A_{l,i}, & \text{if } i \in H_l, \\ A_{l,i}, & \text{otherwise}, \end{cases} \quad \text{where } H_l = \left\{i \mid VHD_{l,i} > \text{median}\left(VHD_{l,*}\right)\right\}. \tag{4}$$

这种实现方式可在单次前向传播中完成注意力头的筛选与强化，而非先全层筛选再分两次强化。此外，当多层 VHR 同步实施时，底层强化会影响上层 VHD 分数计算。逐层策略通过确保计算当前层 VHD 时底层已完成强化，有效避免了此类不一致性。

虽然可以每步计算 VHD 分数并筛选关键头，但不同步骤强化不同头会产生负面影响。具体而言，LVLM 依赖 KV 缓存加速推理，这意味着先前 token 的键值在后续生成步骤中不会重新计算。因此应在生成伊始确定重要头，确保注意力模块中所有 token 的 Q、K 和 V 一致性。实验表明该方法足以缓解幻觉现象。

注意力输出重定向

放大层内特定注意力头的输出以增强其贡献是直观操作，我们通过理论分析论证其合理性。设第 l 层 MHA 模块后的 FFN 输入可表示为：

$$Z_l = \mathrm{RMSNorm}\left(\widehat{X}_l + \mathrm{MHA}_l(X_l)\right) = \hat{g}_l \cdot \frac{\widehat{X}_l + \mathrm{MHA}_l(X_l)}{\widehat{X}_l + \mathrm{MHA}_l(X_l)}, \tag{5}$$

命题 1: 设 LVLM 中第 l 层，h 为待强化注意力头索引。$\tilde{Z}_l$ 为常规 FFN 输入，$\tilde{A}_{l,h} = \alpha \cdot A_{l,h}$ $(\alpha > 1)$，Z_l 为原始输入，$A_{l,h}, Z_{l,h}$ 为仅含 $A_{l,h}$ 分量的伪输入，则有

$$\cos\left(\tilde{Z}_l, Z_{l,h}\right) > \cos\left(Z_l, Z_{l,h}\right)$$

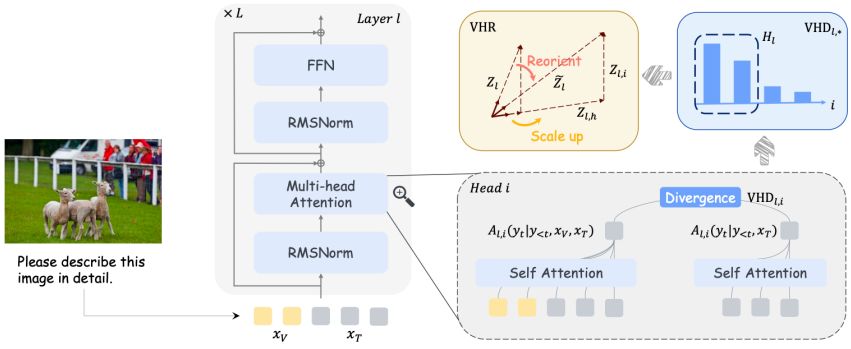


图 2: The illustration of the proposed VHD metric and the VHR approach to mitigate hallucinations in LVLM.

实验细节

设定 2 以平衡幻觉修正效果与隐状态干预强度。VHR 应用于 LLaVA 系列的第 2 层及最后 14 层，InstructBLIP 的最后 18 层。严格复现各基线方法并采用论文报告的参数配置，所有方法在相同基础模型、提示语和生成参数下测试以确保公平性。最大生成长度设为 512 标记，束搜索方法的束宽统一为 5。

(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?】

本文的分析和干预主要集中在多头注意力模块，而 LVLMs 的架构中可能还有其他因素，如视觉编码器和 LLMs 中的前馈网络（FFN）模块等，也会对幻觉现象产生影响，但本文并未对这些部分进行深入探讨和干预。

(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】

作者通过引入 VHD 指标来量化注意力头对视觉上下文的敏感性，为深入理解 LVLMs 的内部机制提供了一种新的视角和方法。这种方法可以启发其他研究者在研究类似问题时，从模型的内部结构和机制入手，

寻找关键的影响因素，并通过量化分析来深入理解问题的本质。

(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……】

在解决 LVLMs 幻觉问题时，通常会采用训练对齐、后处理和解码策略等技术手段，但这些方法各有优缺点，无法从根本上解决问题。而本文提出的从模型内部机制入手，通过分析和调整多头注意力模块来解决问题的新思路，为解决类似问题提供了新的方向。

5 评阅人

姓名:

时间: