【ICT：通过图像-对象跨层级可信干预来缓解大视觉语言模型中的对象幻觉问题】

——【ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models】

1 相关资源

pdf: https://arxiv.org/pdf/2411.15268
ppt:
短视频:
数据集:
源码: https://github.com/THU-BPM/ICT
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: 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: Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, Xuming Hu
单位:

1. Tsinghua University
2. The Hong Kong University of Science and Technology (Guangzhou)
3. Zhipu AI
4. Chongqing University
5. Shanghai Jiao Tong 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（LVLMs）在理解和响应复杂的视觉文本上下文方面取得了显著进展，但在实际应用中，由于其固有的幻觉倾向，即生成与视觉输入不一致的文本输出，限制了其在需要高精度的现实场景中的应用，例如自动驾驶和医疗手术等。

本文的工作旨在提出一种新的方法来减轻 LVLMs 中的对象幻觉现象，以提高其在现实世界任务中的可靠性和准确性，从而使其能够更广泛地应用于各种需要精确视觉理解的场景。

(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】

目前，针对 LVLMs 幻觉问题的研究已经取得了一定的进展。

现有的方法主要可以分为两类：一类是在训练阶段进行干预，例如通过引入额外的高质量标注数据来更好地对齐模型的行为与人类解释，或者设计新的训练目标来减少幻觉，但这些方法通常需要大量的手动标注和计算资源，限制了其可扩展性；另一类是在推理阶段进行干预，如对比解码方法，通过在解码阶段对与幻觉相关的标记进行惩罚来减轻幻觉的影响，但这些方法可能会不加选择地消除所有语言先验，包括那些可能有益的先验，从而导致性能下降。此外，还有些方法通过操纵注意力权重或利用外部工具和知识来减少幻觉。

然而，这些方法要么需要额外的训练成本，要么在推理时引入额外的延迟，或者无法充分利用模型在推理阶段的激活空间。

(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与训练无关的、轻量级的方法，能够在不消除有用语言先验的情况下，通过在前向传播阶段对模型的注意力头进行干预，增强模型对高级和细粒度视觉细节的关注，从而有效减轻 LVLMs 中的对象幻觉现象。

(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. 提出了一种名为 ICT（Image-Object Cross-Level Trusted Intervention）的新方法，该方法在 LVLMs 的前向传播阶段进行干预，而不是在解码阶段。
2. 这种方法通过计算干预方向，将模型的注意力引导到不同层次的视觉信息上，既增强了对视觉细节的关注，又保留了有益的语言先验。
3. ICT 不会引入额外的推理延迟，并且能够更好地利用语言先验来提高模型的准确性。此外，ICT 还具有模型不可知性和跨数据集泛化的能力，使其能够广泛应用于不同的 LVLMs 和数据集。

(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】

论文提出了一种叫做 ICT 的方法，这个方法的核心思想是：在模型处理图片和生成文本的过程中，通过调整模型对图片的“注意力”，让模型更关注图片中的细节，同时保留有用的“先验知识”。这个方法主要包括两部分：图像级干预和对象级干预。图像级干预使 LVLM 聚焦图像，从而降低对语言先验的过度依赖；对象级干预促使 LVLM 更关注图像对象，有助于缓解关键对象遗漏并减少幻觉生成。

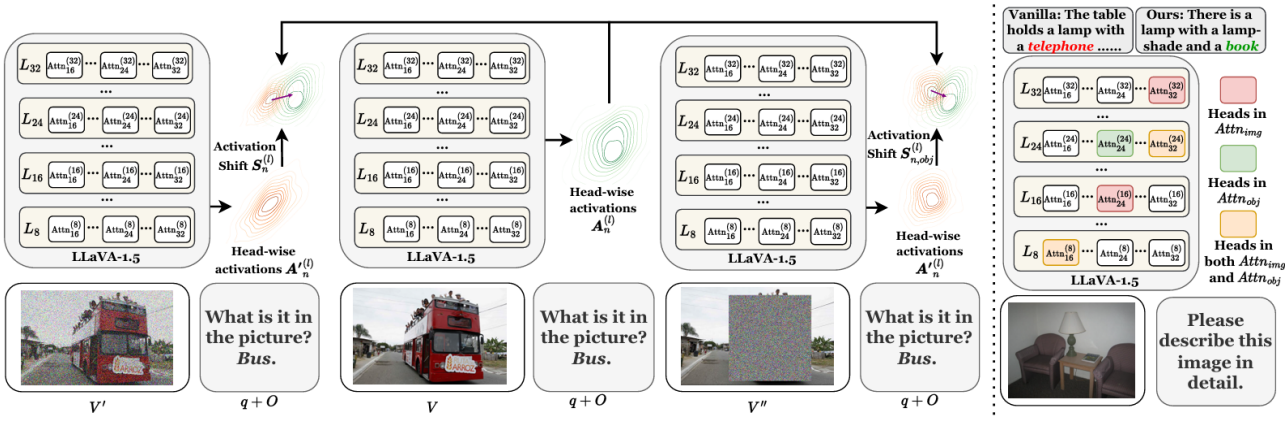

图 1: Overview of proposed ICT method.

图像级干预

该模块旨在识别与整体图像信息相关的注意力头，并施加针对性干预。此方法能增强模型对视觉输入的关注，同时削弱语言先验的影响。

考虑一批图像-问题对 $\{(X_i, V_i)\}_{i=1}^{B}$ 。其中 X_i 采用“图像中是否有 [物体]？”的形式。针对每个问题，提取指定对象 O_i 并将问题重构为 $q =$ “图像中有什么？”。随后对每幅图像 V_i，按照前向扩散过程逐步添加高斯噪声，最终获得模糊图像 V_i'。通过这样的方法生成了一些“可信”和“不可信”的数据对。根据所有样本对获取的可信激活与不可信激活，可计算促使模型更关注视觉信息的激活偏移向量。然后作者为每个注意力头训练二元分类器，以检测哪些头编码了图像级信息——特别是能更好区分可信与不可信样本对差异的头，随后对这些选定头施加激活干预：

$$\mathrm{Attn}_{\mathrm{img}} = \left\{ \mathrm{Attn}_n^{(l)} \mid \mathrm{Attn}_n^{(l)} \in \mathrm{TopK}\left(\mathrm{Accuracy}\left(f_n^{(l)}(\cdot)\right)\right) \right\},$$

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{n=1}^{N} \left(\mathrm{Attn}_n^{(l)}\left(\mathbf{H}^{(l)}\right) + \mathbb{I}_{\mathrm{img},n}^{(l)} \alpha \mathbf{S}_n^{(l)} \right) \cdot \mathbf{W}_o^{(l)},$$

对象级干预

该模块的目的是让模型更关注图片中的具体对象细节，其与图像级干预的做法同理。作者同样生成了一些“可信”和“不可信”的数据对，但这次是针对图片中的具体对象（比如图片中的一只猫或者一辆车）。通过类似的方法，他们调整了那些关注对象细节的注意力头的激活值，让模型在生成文本时能够更准确地描述图片中的具体对象。

(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？】

本文的一个局限性是需要访问模型的权重，因此无法应用于闭源模型。

此外，作者仅使用高斯模糊作为图像变换方法，未来的研究可以探索使用生成方法来变换图像，以进一步提高模型对视觉信息的关注和理解能力

(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】

首先，通过对注意力头的激活模式进行分析，识别出与幻觉相关的头，并对其进行干预，为解决 LVLMs 中的幻觉问题提供了一种新的视角。其次，ICT 方法在前向传播阶段进行干预，而不是在解码阶段，这种干预方式可以避免引入额外的推理延迟，同时保留有用的语言先验，为其他类似问题的解决提供了新的思路。

(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……】

我了解到在解决 LVLMs 幻觉问题时，通常会采用训练阶段干预、推理阶段干预或引入外部信息等技术手段，这些技术之间存在着一定的关联和互补性。例如，训练阶段的干预可以提高模型对视觉信息的关注，但需要额外的训练成本；而推理阶段的干预则可以在不改变模型参数的情况下减轻幻觉，但可能会引入额外的推理延迟。因此，在实际应用中，需要根据具体的需求和场景选择合适的技术手段。

5 评阅人

姓名:

时间: