【视觉证据提示缓解大型视觉语言模型中的幻觉问题】

——【Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models】

1 相关资源

pdf: https://aclanthology.org/2025.acl-long.205/
ppt:
短视频:
数据集:
源码:
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: Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, Jieping Ye
单位:

1. MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
2. Independent Researcher
3. Xiamen 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。

本文的工作旨在通过一种新颖的方法——视觉证据提示（Visual Evidence Prompting, VEP），利用小型视觉模型的细粒度视觉理解能力来补充 LVLMs，从而减少幻觉现象，提升模型在视觉问答等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，LVLMs 在视觉语义理解方面取得了显著进展，例如在图像描述生成和视觉问答等任务中表现出色。然而，这些模型仍然存在幻觉问题，即生成不存在于图像中的对象、属性或关系。现有的研究尝试通过指令微调和优化解码策略来缓解这一问题，但这些方法并未从根本上增强模型的细粒度视觉理解能力。指令微调虽然可以调整模型的输出风格，但存在过拟合特定数据集的风险，可能导致灾难性遗忘。而优化解码策略虽然可以调整输出分布，但同样无法为模型提供新的视觉知识。因此，现有方法在解决幻觉问题上存在局限性，未能从根本上提升模型对视觉内容的细粒度感知能力。

(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通过视觉证据提示（VEP）方法，利用小型视觉模型的细粒度视觉理解能力来补充 LVLMs，从而减少模型在生成输出时的幻觉现象，提升其在视觉问答等任务中的准确性和可靠性。

(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. 本研究首先通过定量与定性分析探究 LVLM 幻觉的成因，发现幻觉主要源于视觉细粒度理解缺陷，导致模型混淆图像中视觉或语义相似的元素。
2. 本研究探索如何通过参考小型视觉模型的视觉证据来缓解 LVLM 的幻觉问题。

(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】

初步分析

作者通过一个具体的例子展示了 LVLM 在回答问题时产生的幻觉现象。例如，当模型被问及图像中是否存在“运动球”时，模型错误地回答“是”，尽管图像中并没有运动球，而是有一个棒球棒。通过使用图像归因图，作者发现模型在生成回答时错误地关注了棒球棒区域，因为棒球棒与运动球在语义上有较高的相似性，这导致模型产生了幻觉。

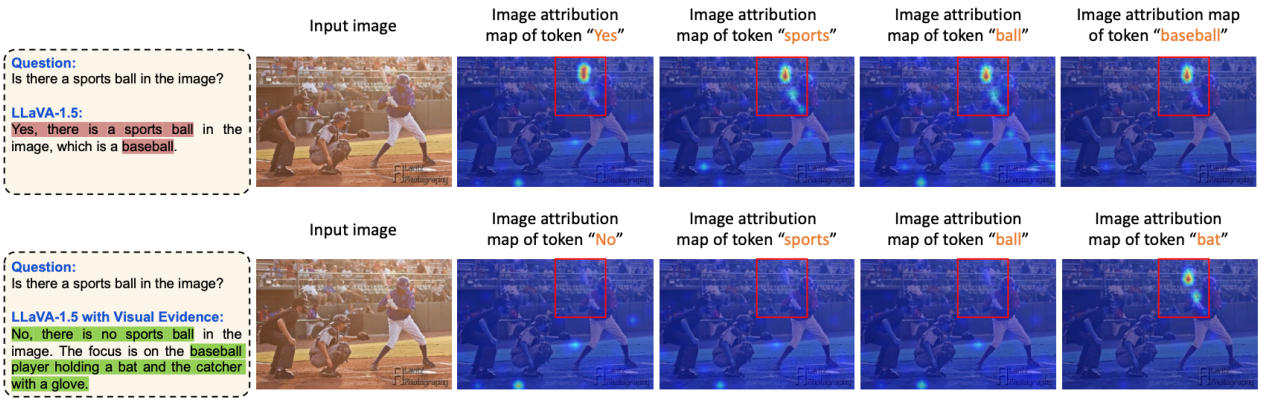

图 1: Visualization of the image attribution map for LLaVA-1.5-7B.

作者统计分析了在幻觉和非幻觉情况下，模型对与查询对象在语义或外观上相似的图像区域的错误激活比例。结果表明，当幻觉发生时，模型错误激活与查询对象相似的区域的比例高达 58.5% 。这进一步证实了模型在处理视觉信息时，容易受到语义或外观相似性的干扰。

作者计算了图像 token 对最终预测的归因分数，发现当幻觉发生时，图像归因分数显著更高。这表明模型在幻觉情况下对图像区域的关注程度更高，但这种关注是错误的。

作者使用 CLIP 模型计算了幻觉对象和非幻觉对象与相应图像之间的特征相似性。结果表明，幻觉对象与图像之间的 CLIPScore 显著高于非幻觉对象，这表明幻觉对象与图像在语义上更接近。这种语义上的接近性导致模型错误地激活了与查询对象相似的图像区域。

作者采用了一种方法来计算 LVLM 内部视觉 token 对对象的置信度分数。结果表明，幻觉对象的置信度分数显著高于非幻觉对象。这表明模型的内部视觉表示倾向于过度编码幻觉内容，进一步揭示了 LVLM 在细粒度视觉感知能力上的不足。

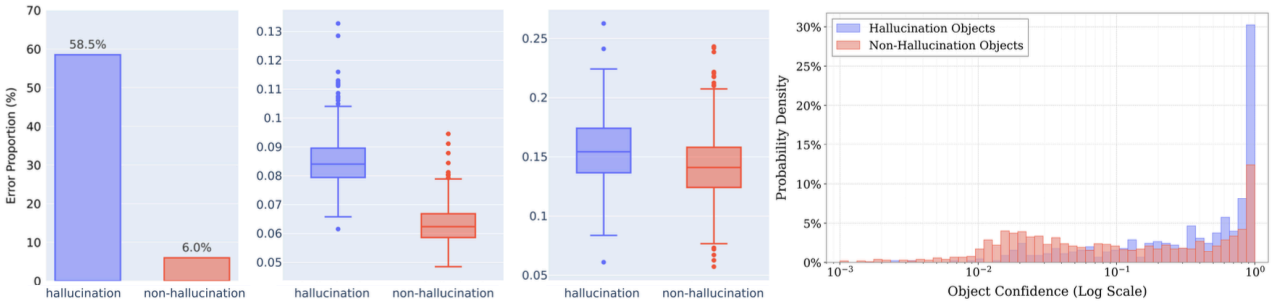

图 2: Hallucination analysis on LLaVA-1.5.

视觉证据提示 VEP

当询问图像中是否存在椅子时，原始 LVLM 给出了错误答案。而小型视觉模型（如目标检测和场景图生成模型）能输出准确物体及关系，如" 狗"、" 杯子"、" 狗靠近杯子"。将这些准确可靠的输出符号化为上下文后，模型便能生成正确响应。作者将这种无需训练的方法称为视觉证据提示（VEP）。本研究旨在通过补充来自小型视觉模型的细粒度视觉知识，减轻 LVLM 的幻觉现象。在概率框架下，基于输入图像 I 和问题 Q 生成答案 A 可表述为估计条件分布 $P(A \mid Q, I)$。视觉证据提示形式

化为 $R(A \mid Q, I, VE)$，其中 VE 表示从图像中提取的视觉证据。

当基于图像内容回答问题时，考虑人的内部认知过程，通常会将问题分解为两个步骤。例如图 3 中关于” 图像中有椅子吗？ ” 的提问：首先识别图像中作为证据的关键要素 (”1 只狗、1 个杯子、1 份报纸，狗靠近杯子，狗在桌上，报纸在桌上”)，然后将证据中的相关内容符号化组合来回答问题，最终生成答案。

提取阶段: 将大视觉语言模型的输入图像馈入小型视觉模型，输出按预定格式组织。对于物体检测模型，输出定义为根据预测类别索引从标签映射中获取的语义标签。若检测到同类多个物体，则合并统计数量，如”3 只狗，1 只猫”。场景图生成模型的输出由，关系，宾语) 三元组构成，每个三元组首先格式化为主语关系宾语，多个三元组用”,” 连接。例如男人在冲浪板上和男人有头发格式化为” 男人在冲浪板上，男人有头发”。

提示阶段: 第二步使用符号化的视觉证据和问题提示从 LVLM 提取最终答案。具体而言，我们将两个元素简单拼接为” 你可以在图像中看到证据。问题？ ”。该步骤的提示具有自增强特性，因其包含视觉模型生成的证据。这是视觉证据的一种简单有效表达形式，更复杂的格式可能带来进一步改进。最终将提示文本和原始图像输入 LVLM 生成答案。

$$A = f_{LVLM}(I, Q, VE), \quad VE = T[f_{SVM}(I)]. \tag{1}$$

此处 SVM 代表小型视觉模型，T 表示将小型视觉模型的结构化输出转换为自然语言的过程。

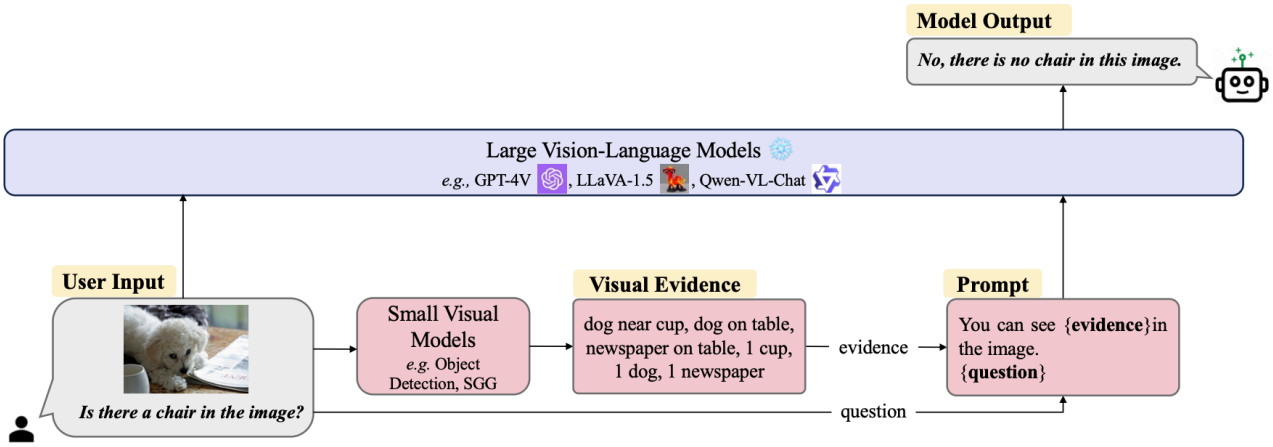

图 3: An overview of visual evidence prompting.

实现细节

首先使用对应视觉小模型（即目标检测模型和场景图生成模型）处理评估数据集中的图像以获取视觉证据。为突出框架有效性，默认采用 detr-resnet-101 提取对象证据，RelTR 获取关系证据，高性能模型和开放词汇模型亦可适用。

(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?】

尽管本研究为缓解幻觉问题提供了启示，但仍存在若干局限性：1) 知识整合有限。与微调不同，基于提示的策略不会将新知识融入模型参数。先前研究表明，过度微调可能导致模型因过拟合训练数据模式而忽视输入问题，从而产生幻觉。相比之下，提示策略保留了模型原始权重，提供更强可控性并保持泛化能力，但也限制了其永久嵌入领域特定知识的能力。

2) 计算开销。引入外部视觉模型必然增加计算成本，尽管本方法产生的开销远低于指令微调

(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】

VEP 方法提供了一种简单且有效的解决方案，展示了如何通过结合小型视觉模型和 LVLMs 的优势来解决复杂的视觉语言问题。这种方法可以应用于其他需要视觉语言融合的任务，如图像描述生成、视觉问答等。其次，本文通过详细的实验和分析，揭示了 LVLMs 幻觉问题的成因，并提出了针对性的解决方案，为后续研究提供了有价值的参考。

(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……】

本文的实验设计和分析方法非常值得学习，尤其是如何通过图像归因图和内部可解释性分析来理解模型的行为。

5 评阅人

姓名:

时间: