【通过视觉对比解码缓解大型视觉语言模型中的物体幻觉问题】

——【Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding】

1 相关资源

pdf: https://doi.org/10.48550/arXiv.2311.16922
ppt:
短视频:
数据集:
源码: https://github.com/DAMO-NLP-SG/VCD
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: 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: Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, Lidong Bing
单位:

1. DAMO Academy, Alibaba Group
2. Nanyang Technological University
3. Hupan Lab, 310023, Hangzhou, China

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。

本文旨在提出一种无需额外训练、无需外部工具的方法，在推理阶段即时抑制这种幻觉，使输出内容与图像严格一致，从而提升 LVLM 的可信度与落地价值。

(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】

当前的研究中，已经有一些方法被提出用于减少多模态大语言模型中的幻觉现象。例如，WoodPecker 和 LURE 通过后处理生成的响应来减少幻觉；LRV-Instruction 和 RLHF-V 通过指令微调来缓解幻觉问题。然而，这些方法通常依赖于外部数据或需要额外的模型训练，这在计算资源和数据收集上存在一定的局限性。此外，一些研究通过改进解码策略来减少幻觉，例如 OPERA 通过惩罚不参考视觉信息的生成。尽管这些方法取得了一定的进展，但它们在处理幻觉问题时仍然存在不足，尤其是在不依赖额外数据或模型更新的情况下减少幻觉的能力上。本文认为，现有的方法虽然在一定程度上缓解了幻觉问题，但仍然需要更高效且无需额外训练的解决方案。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文解决的核心问题是减少 LVLMs 在生成文本描述时的对象幻觉现象，同时保持模型的高效性和无需额外训练的特点。

(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. 深入分析视觉不确定性对 LVLMs 物体幻觉的影响机制，重点探讨统计偏差与单模态先验的作用
2. 提出无需训练的 VCD 技术，通过对比原始/失真视觉输入的输出分布校准模型，确保生成内容一致性
3. 实验验证 VCD 在缓解物体幻觉和增强感知能力方面的有效性，该方法无需额外训练或外部工具即可实现显著提升

(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】

视觉不确定性放大幻觉

视觉不确定性会放大语言先验，图 1 表明视觉不确定性会迫使 LVLM 忽视视觉证据，过度依赖语言先验进行决策。这种行为并非完全意外，因为 LLM 本就是基于海量文本语料预测下一词概率的模型。当面对模糊视觉刺激时，LVLM 可能将这些常规的、基于文本的预测误解为” 安全网”。这些先验虽然通常有用，但可能引入与实际视觉内容不符的偏见或假设，尤其在视觉输入不清晰时。

如图，当图像中出现混杂在其他彩色水果中的黑色香蕉时，随着视觉不确定性的增加，LVLM 更倾向于选择常规香蕉颜色（如” 黄色” 和” 绿色”）。真实颜色” 黑色” 的概率随着失真加剧而降低，这使得 LVLM 过度依赖来自 LLM 预训练的语言先验——这些先验通常将香蕉与黄色或绿色相关联。

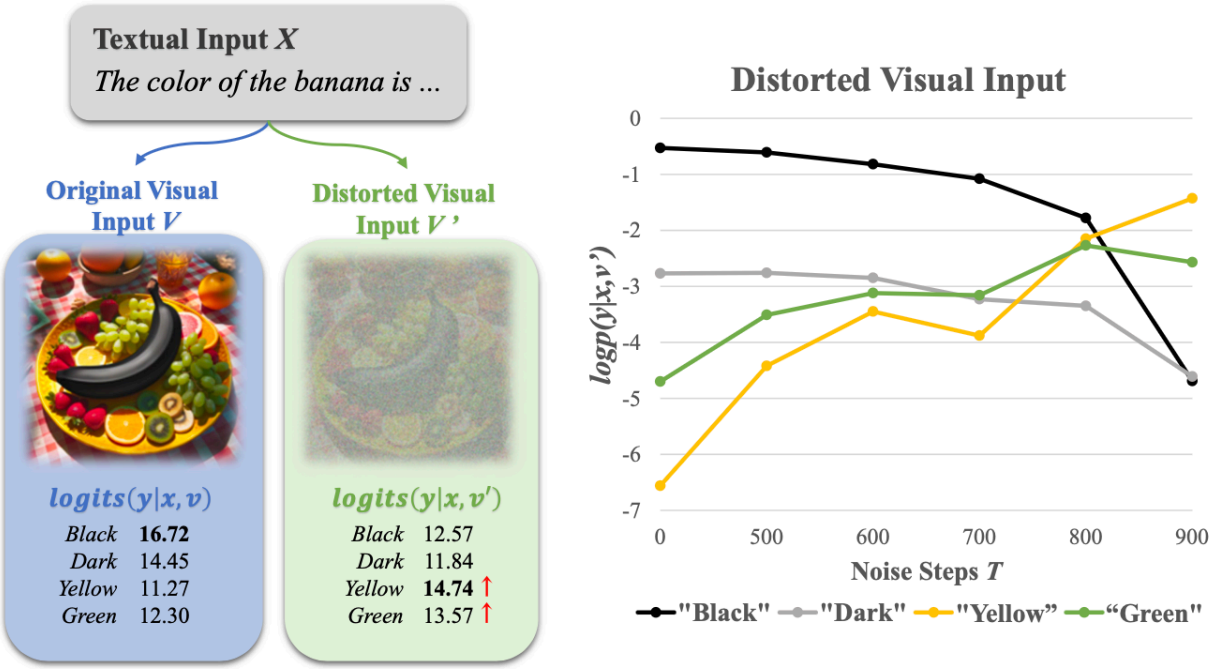

图 1: An illustration of visual uncertainty amplifying lan- guage priors.

为验证视觉不确定性可能放大预训练统计偏差的假设，作者设计了两组针对性实验: (1)LVLM 是否更易对失真视觉输入生成高频幻觉物体; (2)LVLM 是否更易对失真视觉输入生成与图中真实物体高频共现的幻觉物体。图 2 表明，由于训练数据中物体分布不平衡和虚假相关性，LVLM 确实更易产生高频共现物体的幻觉。

视觉对比解码

以上发现表明，视觉不确定性不仅会增强对语言先验的依赖，还使 LVLM 更容易受到预训练数据集中表面物体相关性的影响，导致更严重的幻觉。为此作者提出视觉对比解码 (VCD)。VCD 通过对比原始与失真视觉输入产生的模型输出，旨在抵消 LVLM 中的统计偏差和语言先验。该方法无需额外训练或外部预训练模型，是一种高性价比的高效解决方案。

1. 预测对比

具体而言，给定文本查询 x 和视觉输入 v，模型生成两个不同的输出分布：一个基于原始 v，另一个基于通过预定义失真（即高斯噪声掩码）处理得到的失真视觉输入 v'。随后通过利用两个初始分布间的差异，计算得到新的对比概率分布。该对比分布 p_{vcd} 公式如下：

$$p_{\text{vcd}}\left(y \mid v, v', x\right) = \text{softmax}\left[(1+\alpha)\,\text{logit}_\theta\left(y \mid v, x\right) - \alpha \text{logit}_\theta\left(y \mid v', x\right)\right], \tag{1}$$

其中 α 值越大表示两个分布差异的放大程度越强（当 $\alpha = 0$ 时为常规解码）。根据调整后的输出分布 p_{vcd}，可采用多种采样策略，如核采样和束搜索。

2. 自适应合理性约束

根据公式中对比分布 p_{vcd} 的构成，可能面临一个挑战：该方法会惩罚模型受失真视觉输入影响的全部输出行为。但这并非普遍正确——失真视觉输入下的输出分布仍可保持基本语言规范和常识推理。无差别惩罚可能错误地打击这些有效输出，并促使生成不合理标记。为解决此问题，作者实施基于

原始视觉输入相关置信度的自适应合理性约束：

$$\mathcal{V}_{\text{head}}\left(y_{<t}\right)=\left\{y_t \in \mathcal{V}:\right. \tag{2}$$

$$\left.p_\theta\left(y_t \mid v, x, y_{<t}\right) \geq \beta \max_w p_\theta\left(w \mid v, x, y_{<t}\right)\right\}, \tag{3}$$

$$p_{\text{vcd}}\left(y_t \mid v, v', x\right)=0, \text{if } y_t \notin \mathcal{V}_{\text{head}}\left(y_{<t}\right), \tag{4}$$

其中 $\mathcal{V}$ 表示 LVLM 的输出词汇表，β 是 $[0,1]$ 中控制下一标记分布截断的超参数。β 值越大代表更激进的截断策略，仅保留高概率标记。

引入自适应合理性约束可优化对比分布，增强对明确决策的信心。这确保当模型对原始输入相关输出具有高置信度时，候选池将精简，通常仅保留单一高概率标记。该方法有效中和 VCD 的潜在负面影响，防止其意外促生不合理标记，维护生成内容的完整性。（这个公式表示，从整个词表中只保留那些在原始图像条件下预测概率不低于最大概率的　倍的词。）

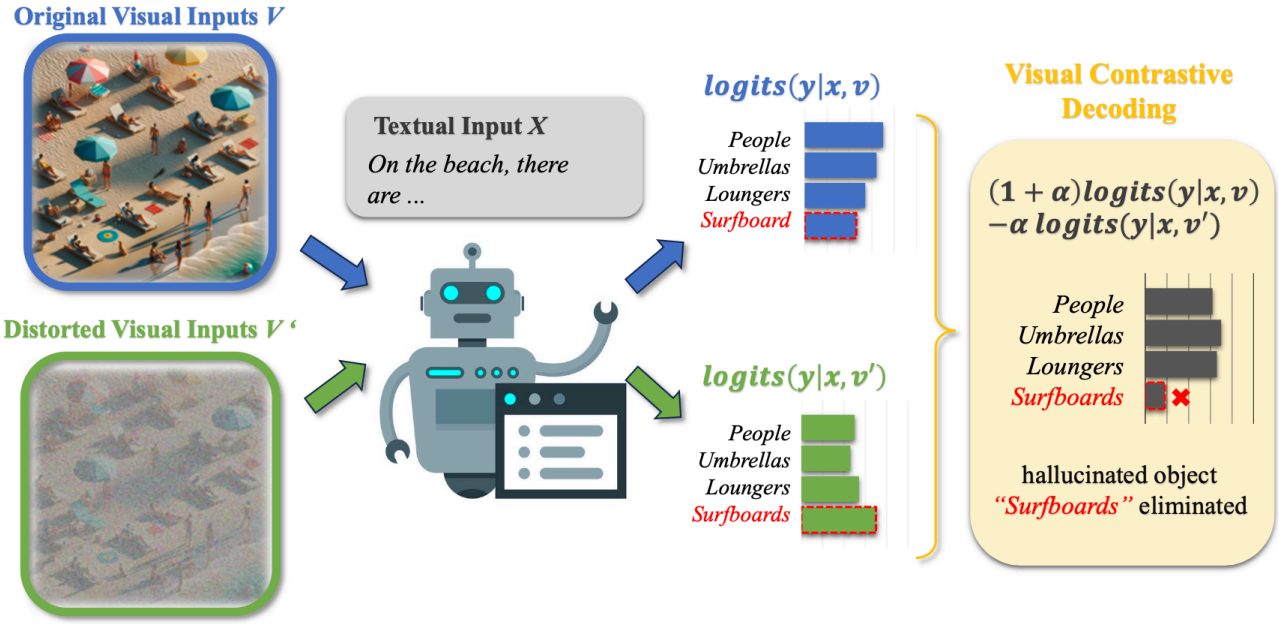

图 2: An illustration of Visual Contrastive Decoding

实现细节：实验默认设置 $\alpha=1, \beta=0.1$ 和 $\gamma=0.1$。

(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？】

VCD 对所有模型的物体级幻觉处理均有提升，并对颜色属性分数产生积极影响。相比之下，位置推理得分较低且 VCD 提升有限，表明 LVLM 在位置推理方面能力较弱。

(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】

本文的 VCD 方法为解决 LVLMs 中的对象幻觉问题提供了一种新的思路，即通过对比原始和失真输入的输出分布来校正模型的输出。这种方法简单高效，且无需额外训练，具有很强的可扩展性。此外，VCD 中引入的自适应合理性约束也为如何在调整模型输出时保持合理性提供了有益的参考。

(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……】

我了解到对比学习和自适应约束等技术在多模态生成任务中的应用潜力。

5 评阅人

姓名:

时间: