【ConVis：基于幻觉可视化的对比解码方法缓解多模态大语言模型中的幻觉问题】

——【ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models】

1 相关资源

pdf: https://ojs.aaai.org/index.php/AAAI/article/view/32689
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: AAAI 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: Yeji Park, Deokyeong Lee, Junsuk Choe, Buru Chang
单位:

1. Sogang University
2. Korea Universit

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?】

本文的研究背景是多模态大语言模型在生成与图像相关的文本时存在的“幻觉”问题，即生成的文本内容与给定图像不一致，严重影响了模型的可靠性。这种问题在需要高可靠性的领域（如医疗诊断、自动驾驶等）尤为突出。本文的目标是提出一种新的训练无关的对比解码方法（ConVis），通过在解码过程中引入视觉对比信号来减少幻觉现象，从而提高多模态大语言模型在生成任务中的可靠性和准确性。

(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，VCD 利用扭曲图像来减少模型对统计偏差的依赖。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】

本文提出了一种利用 T2I（文本到图像）模型的新型解码方法。具体而言，该方法首先通过 T2I 模型将初始生成描述中的幻觉内容可视化，随后对比重构图像与原始图像生成的响应。通过这一过程，作者对比了幻觉标记的分布特征，从而有效抑制了幻觉现象。

(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. 提出了一种利用 T2I 模型的新型解码方法。
2. 在五个主流基准测试上的实验表明，ConVis 能有效降低各类 MLLMs 的幻觉现象，展现了提升模型可靠性的潜力。

(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】

ConVis 利用文本生成图像模型（特别是 Hyper-SDXL）捕捉视觉对比信号：首先生成输入图像的描述文本，再由 T2I 模型重建图像。若文本含幻觉，重建图像将出现视觉差异。ConVis 通过对比原始与重建图像的概率分布，捕捉凸显幻觉的视觉对比信号，进而在解码过程中抑制幻觉生成。**幻觉可视化**

作者首先假设 T2I 模型可通过在解码过程中提供视觉对比信号来缓解幻觉。若 T2I 模型接收到包含幻觉的 MLLM 生成描述，将在生成图像中如实呈现这些幻觉内容。将此过程称为幻觉可视化。

具体实现时，ConvVis 首先使用指示 MLLM 描述图像的简单指令文本，为原始图像 v 生成初始描述 c。T2I 模型随后以该描述 c 作为查询生成图像 v'。若描述包含幻觉，生成图像 v' 将忠实呈现这些内容；反之若初始描述准确无幻觉，生成图像将与原始图像保持语义相似性。

鉴于当前 T2I 模型可能无法生成完全匹配描述的图像，作者通过以下方法提升生成图像的多样性：(1) 采用核采样解码而非贪婪解码，首先生成 n 组多样化描述；(2)T2I 模型利用这些 n 描述生成 n 对应图像。该方法通过描述多样化覆盖 MLLM 可能产生的各类潜在幻觉，同时使用多幅图像（而非单幅）增强了方法对 T2I 模型性能缺陷导致的图文失配的鲁棒性。

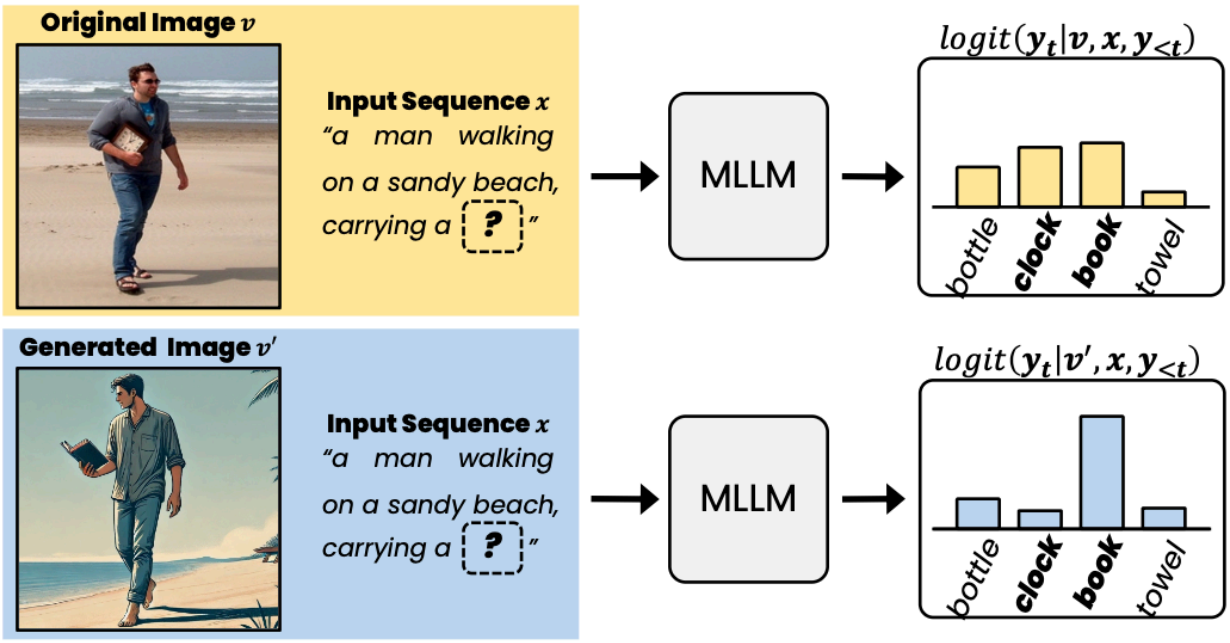

图 1: The original and reconstructed image generate the contrastive logit distribution for the hallucinated tokens (e.g., 'book').

对比解码

描述中的幻觉会导致原始图像 v 与生成图像 v' 间的视觉差异。我们通过捕捉这些差异产生的视觉对比信号来缓解幻觉。具体实现时，在解码过程中同时利用原始图像 v 和 n 幅生成图像产生各图像的逻辑分布，最终对比逻辑分布 $\hat{f}_\theta$ 通过原始图像与各生成图像的对比逻辑分布平均值计算得出：

$$\hat{f}_\theta = \frac{1}{n}\sum_{i=1}^{n}\left((1+\alpha) f_\theta\left(\cdot \mid v, x, y_{<t}\right) - \alpha f_\theta\left(\cdot \mid v_i', x, y_{<t}\right)\right), \tag{1}$$

其中 α 是控制原始图像与生成图像间逻辑分布差异强度的超参数。对比逻辑分布 $\hat{f}_\theta$ 用于生成响应 y。对于与幻觉相关的标记，其对比逻辑分布会相较于其他标记显著放大，从而实现对这类标记的惩罚并减少幻觉现象。

该公式与 VCD 和 HALC 使用的对比解码方法类似。但本方法的创新点在于直接从 T2I 生成模型可视化的幻觉中捕捉视觉对比信号，以此区别于现有方法。

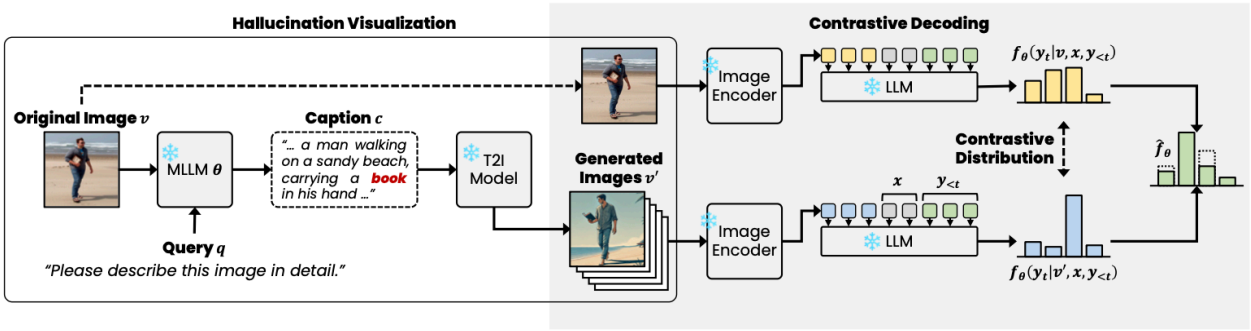

图 2: The original and generated image produce the contrastive distribution for the hallucinated tokens (e.g., 'book').

实验细节

实现细节。图像生成采用 Hyper-SDXL T2I 模型。默认使用该模型 Step 1 生成结果。

T2I 模型最大文本查询长度为 77 个标记，为处理 MLLM 生成的长描述，采用 Compel 工具支持超长文本处理。

描述生成最大标记数设为 256，使用温度 0.7、top-p 值 0.9 的核采样生成图像。

查询语句为"请详细描述此图像"。

生成图像数量 n 设为 4，基于不同随机种子生成四组图像。

对比解码采用的自适应合理性约束，仅对比有效标记。

合理性约束超参数 λ 设为 0.1。

对比强调程度控制参数 α 在描述类指标（如 CHAIR 和 LLaVA-Bench）中设为 1，在 VQA 类指标（包括 POPE、HallusionBench 和 MME）中设为 0.1。所有方法响应生成均采用贪婪解码策略。对于 CHAIR 测试，使用三组不同随机种子采样的图像集进行评估，结果取均值与标准差。

(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?】

本方法主要局限在于高度依赖 T2I 生成模型。这种依赖性可能影响在 VQA 等任务中的效果——当生成描述包含与问题严重偏离的幻觉时。在 POPE 基准测试中，这种局限尤为明显，性能提升未达预期。针对特定物体存在性的提问，若该物体与描述生成的幻觉无关，T2I 模型的可视化可能无法充分反映 VQA 任务所需信息。

(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】

利用 T2I 模型可视化幻觉内容的思路为处理多模态模型中的幻觉问题提供了一个新的视角。我觉得作者的想法很新颖，之前的文章都是利用本身已有的信息修改概率分布，而本文是将生成的图片描述再去生成图片，对比两张图的不同来判断是否有幻觉，很有意思。

(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 T2I 模型在多模态任务中的潜力。

5 评阅人

姓名:
时间: