【DAMRO：深入探索大型视觉语言模型的注意力机制以减少目标幻觉】

——【DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination】

1 相关资源

pdf: https://arxiv.org/pdf/2410.04514
ppt:
短视频:
数据集:
源码: https://github.com/coder-gx/DAMRO.
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: EMNLP 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: Xuan Gong, Tianshi Ming, Xinpeng Wang, Zhihua Wei
单位:

1. Department of Computer Science and Technology, Tongji 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。
本文的工作旨在深入分析 LVLMs 中视觉编码器与语言解码器的注意力机制之间的关系，并探索这种关系对对象幻觉的影响，从而提出一种新的方法来减少 LVLMs 中的对象幻觉现象，提高模型生成文本与输入图像的一致性。

(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，针对 LVLMs 中对象幻觉问题的研究已经取得了一定进展。早期工作主要通过优化训练方法、引入外部信息或模型（如 DETR）、提供幻觉信息反馈等方式来改善模型性能。近期，对比解码等新型解码方法也被引入，用于减少幻觉现象。然而，这些方法大多集中在改进模型架构或特定模块（如视觉编码器或语言解码器），而忽略了 Vision Transformer（ViT）结构本身对幻觉生成机制的影响。现有的研究虽然在一定程度上缓解了幻觉问题，但仍然存在局限性，例如对模型结构的依赖性较强，或者需要额外的训练过程。此外，现有方法在处理细粒度语义幻觉方面效果有限，尤其是在复杂场景下，模型仍然容易生成与图像不匹配的对象描述。

(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 LVLMs 中视觉编码器与语言解码器之间的注意力机制一致性导致的对象幻觉现象。具体来说，研究发现视觉编码器中的高注意力异常值在解码阶段被语言模型过度关注，从而导致模型生成与图像不一致的对象描述。本文提出了一种新的方法来过滤这些异常值，减少其对解码阶段的影响，从而有效缓解对象幻觉问题。

(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. 对视觉编码器和大语言模型解码器的注意力图之间的关系进行了深入分析，揭示了它们异常值标记的分布具有高度一致性。
2. 分析了一致性对对象幻觉的影响，并设计了 DAMRO 方法以减轻大型视觉语言模型中的幻觉问题。
3. 通过在各种模型和基准上进行的大量实验，证明了我们方法的有效性。此外，我们的无训练方法适用于大多数大型视觉语言模型，且不需要外部知识或模型。

(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 Transformer 的缺点
视觉 Transformer 凭借其卓越的视觉表征能力，已成为所有大型视觉语言模型（LVLM）广泛青睐的骨干视觉编码器。然而，前人发现，ViT 中始终存在高范数离群令牌，这些令牌往往出现在具有冗余补丁信息的背景区域，包含极少的局部信息，但有少量全局信息。
如图所示，大语言视觉模型视觉编码器的注意力图也会聚焦于少数高范数异常值令牌。作者推测，这些异常值令牌体现了 ViT 中的负面视觉先验。当图像令牌被投影并发送至大语言模型时，由于这

些令牌在视觉编码器中具有较高的注意力值，大语言模型也倾向于关注它们，从而忽略了其他图像块中包含的局部信息。这可能会导致模型的细粒度视觉能力下降。
研究结果证实，这些少量标记确实包含大量信息，但不够准确。

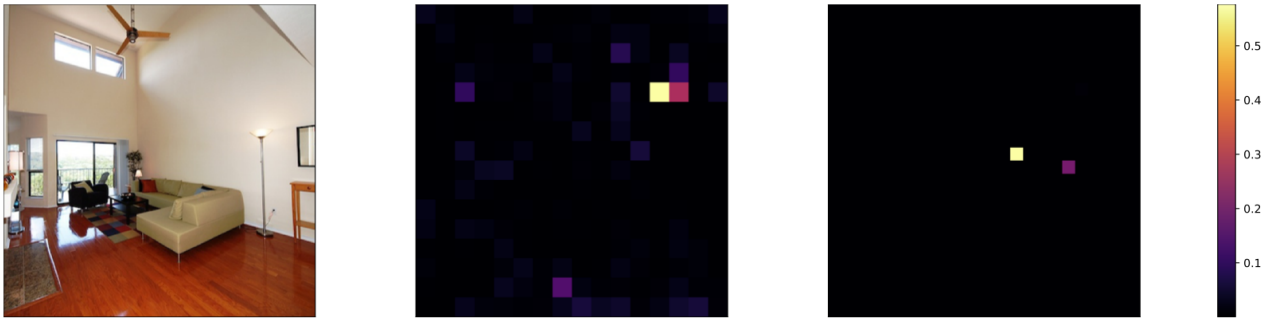

图 1: Attention map of visual encoder. Left: original image. Middle: attention map of InstructBLIP ViT (16x16). Right: attention map of LLaVA-1.5 ViT (24x24).

异常值标记导致幻觉

基于 VIT 中上述问题，作者尝试观察大语言模型解码阶段图像 token 的注意力图。发现大语言模型解码器的注意力图也具有一个与视觉编码器相同位置的少数异常值标记获得了比其他标记更多的注意力。
作者假设这种一致性与幻觉的发生有关，即 LLM 解码器更关注视觉编码阶段识别出的异常值标记。然后选取了一个例子来展示这种相关性。

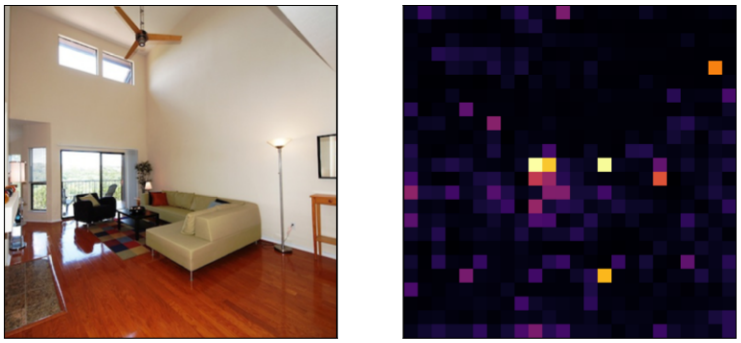

图 2: LLM decoder attention map of ”plant” to- ken (non-hallucinatory). It is evident that attention can accurately locate the position of the plotted plant.

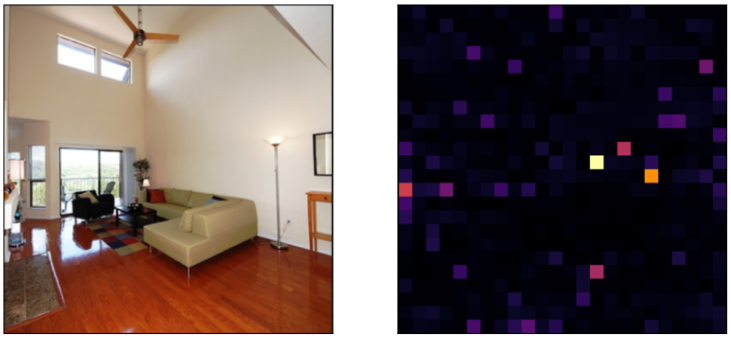

图 3: LLM decoder attention map of ”clock” to- ken (hallucinatory). The attention mainly focus on the outlier tokens in the background, whose positions are the same in visual encoder attention map in the right sub-image of Figure 2.

为了定量描述这种一致性，提出了一个评估指标 H_i，其中 $S_v(i)$ 表示从视觉编码器的注意力图中提取的注意力值排名前 i 的标记集合，而 $S_l(i)$ 表示从 LLM 解码器的注意力图中提取的排名前 i 的标记集合。在这个公式中，$|S|$ 表示集合 S 的基数，即集合 S 中包含的元素数量。

$$H_i = \frac{|S_v(i) \cap S_l(i)|}{i}. \tag{1}$$

从 MSCOCO 数据集的 val2014 子集中随机选择 1000 张图像，并使用提示 “在这张图像中你能看到什么” 向 LLaVA-1.5 进行查询，以获取模型生成的描述。将生成的标题和目标词作为两种单元，并采用 CHAIR 来识别幻觉内容。然后，我们利用指标 H_i 进行分析幻觉的出现与其分布一致性之间的关系，如图所示。

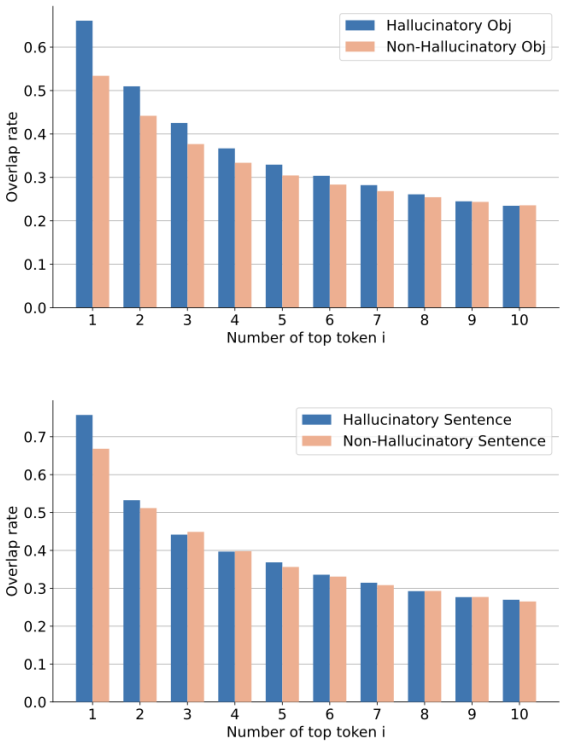

图 4: Top 1-10 outlier tokens overlap rate between visual encoder and LLM decoder. Both of object-level and sentence-level results show that hallucination tends to happen when overlap rate is higher, especially con- sidering the top tokens.

此外，作者发现视觉编码阶段注意力分数最高的前三个标记占了超过 99% 的注意力，。为了进一步验证这些标记的影响，分析了这三个相同标记在大语言模型解码器注意力图中的占比。影响的评估指标记为 F，定义为

$$F = \frac{\sum_{j=1}^{3} ATT\left(L_v(j)\right)}{\sum_{i=0}^{n-1} ATT(i)} \tag{2}$$

其中，$L_v(i)$ 表示视觉编码器注意力图中具有第 i 高注意力值的标记的位置，而 $ATT(i)$ 表示位置 i 处标记的大语言模型解码器注意力值。

同样，使用生成的标题和对象词作为单位来识别幻觉，可以观察到，视觉编码阶段的异常标记确实会影响后续的大模型解码阶段，这与幻觉的产生密切相关。

Granularity	HA	Non-HA
sentence-level	0.0554	0.0539
object-level	0.0605	0.0551

图 5: F Value results. HA: hallucinatory, Non-HA: non-hallucinatory. It is easily observed that at both the sentence level and the object level, the influence of outlier tokens from the visual encoder is greater when hallucinations occur.

异常值标记选择

在 ViT 的自注意力最后一层中，类别标记 [CLS] 通常用于分类。[CLS] 标记在注意力计算中用作查询向量，其他视觉标记则作为关键向量：其中 Q_{cls} 是 [CLS] 标记的查询向量乘以相应权重后的结果；K^T 是所有其他图像标记的关键向量乘以其相应权重后的结果，d 是 Q_{cls} 的维度。

$$A_{\text{cls}} = \text{softmax}\left(\frac{Q_{\text{cls}}K^T}{\sqrt{d}}\right) \tag{3}$$

作者基于倒数第二层的类别标记 [CLS] 和空间视觉标记之间的注意力值，对前 k 个异常标记进行采样，记为：

$$\text{token}_{\text{outlier}} = \arg\max_{\text{token}_i}\left(A_{\text{cls}}(\text{token}_i)\right) \tag{4}$$

（在 ViT 架构中，输入图像会被分割为多个空间 token（如 LLaVA-1.5 的 24×24=576 个 token），同时会额外引入一个分类 token（[CLS]）。这个 [CLS] token 不对应任何图像区域，而是通过与所有空间 token 进行自注意力计算，聚合全局图像信息，最终用于分类任务。

核心思路是：[CLS] token 对空间 token 的注意力值，能反映空间 token 的 “全局信息重要性” —— 而异常 token（背景中含冗余信息的 token）往往会获得 [CLS] token 的高注意力（因冗余背景信息易被误判为 “全局相关”），因此可通过 [CLS] 与空间 token 的注意力值筛选异常 token。）

对比解码

作者使用对比解码来减轻视觉编码器中视觉异常标记对后续文本生成的影响。在大型视觉语言模型中，对比解码通常在大型语言模型解码的采样过程中进行，其中下一个标记是根据 logits 空间中的概率分布来确定的。

$$p_t = \text{softmax}\left((1+\alpha)\text{logits}_\theta(y_t|y_{<t}, v, x) - \alpha\text{logits}_\theta(y_t|y_{<t}, v_{\text{cls}}, x)\right) \tag{5}$$

logits 空间中的概率分布减弱了先前异常值标记对解码的影响。这使得模型能够更专注于细粒度的语义信息，并消除包含视觉编码器先验的冗余信息，从而减轻 LVLM 中的幻觉现象。为解决全局信

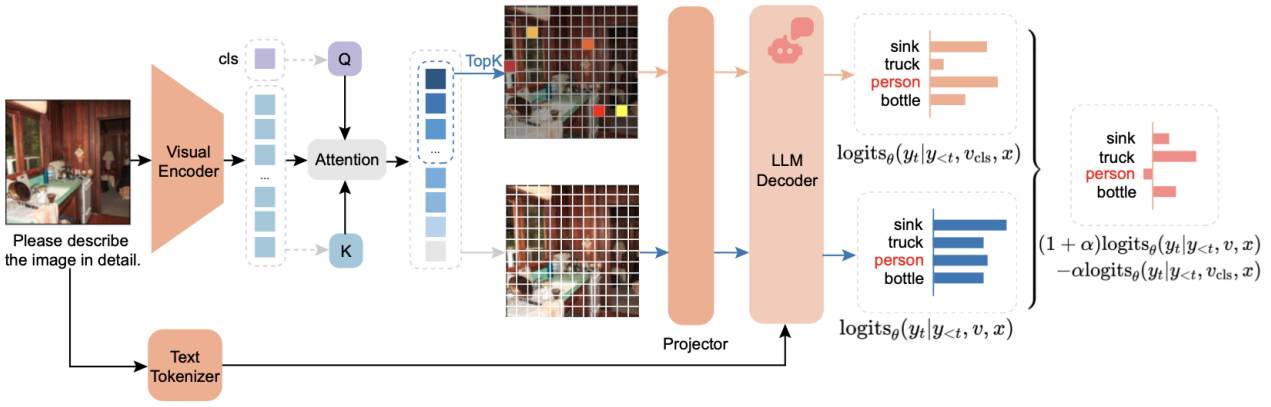

图 6: An overview of DAMRO.

息过度移除的问题，引入了自适应合理性约束。在对比解码阶段，设定了一个阈值 来截断基于原始模型预测置信度的新概率分布：

$$\mathcal{V}_{head}(y_{<t}) = \{y_t \in \mathcal{V} : p_\theta(y_t|v, x, y_{<t}) \geq \beta max_w p_\theta(w|v, x, y_{<t})\}. \tag{6}$$

V_{head} 作为采样的过滤约束条件。

(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?】

本文主要关注了 ViT 结构的视觉编码器，而对于其他复杂的视觉编码器结构（如 QFormer）的适用性尚未充分验证。

(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】

对 LVLMs 中视觉编码器与语言解码器注意力机制的深入分析为理解模型幻觉现象提供了新的视角，这种分析方法可以应用于其他多模态模型的研究中，帮助揭示不同模态间信息交互的潜在问题。

(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……】

本文让我意识到，尽管现有的技术在一定程度上缓解了幻觉问题，但仍然存在许多挑战，尤其是在处理复杂场景和细粒度语义时。这表明未来的研究需要在模型架构、训练方法和解码策略等方面进行更深入的探索和创新。

5 评阅人

姓名:

时间: