【ClearSight：用于缓解多模态大型语言模型中物体幻觉的视觉信号增强技术】

——【ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models】

1 相关资源

pdf: https://arxiv.org/pdf/2503.13107
ppt:
短视频:
数据集:
源码: https://github.com/ustc-hyin/ClearSight
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: 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: Hao Yin, Gunagzong Si, Zilei Wang
单位:

1. University of Science and Technology of 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。
本文的工作旨在缓解 LVLMs 中的幻觉问题，提高模型生成内容的准确性和可靠性，从而增强用户对模型的信任，使其能够更好地应用于实际场景。

(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】
当前，对比解码策略是减少 MLLMs 中物体幻觉的主流方法之一。这些方法通过对比原始视觉输入和扰动后的视觉输入的输出分布，来减少模型对语言先验的过度依赖。例如，视觉对比解码 VCD 方法通过对比原始和扰动后的视觉输入的输出分布，有效地减少了幻觉现象。然而，这些方法存在两个主要问题：一是可能会破坏生成内容的连贯性和准确性；二是增加了推理时间，降低了模型的效率。尽管这些方法在减少幻觉方面取得了一定的进展，但它们的局限性限制了其在实际应用中的广泛使用。

(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如何在不降低多模态大语言模型推理速度和生成内容质量的前提下，有效减少模型生成中的物体幻觉现象。

(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. 发现对比解码方法对生成内容的质量和模型推理速度均存在负面影响。
2. 分析了多模态大语言模型（MLLMs）中的模态融合机制，强调其对视觉信息的关注不足。
3. 提出了 VAF 方法，该方法能有效缓解目标幻觉问题，同时保持推理速度、连贯性和准确性。
4. 展示了 VAF 方法在多个对象幻觉基准测试中的显著性能提升。

(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】
对比解码的局限性
由于对比解码方法不需要训练或外部工具，因此它们具有较高的计算效率和通用性，在学术界引起了广泛关注。然而，这些方法仍存在两个主要缺点：生成内容的质量下降和推理速度较慢。
为了验证这一点，作者将 VCD 方法应用于 LLaVA-v1.5-7B 和 LLaVA-v1.5-13B 模型，评估它们在 ScienceQA 和 NoCaps 基准测试中的表现。研究结果表明，应用 VCD 后，模型在 ScienceQA 上的性能下降了 5%，在 NoCaps 上的性能大幅下降了 45%。这些结果表明，在需要细致自然语言生成的任务中，对比解码方法会显著损害内容质量。
对比解码方法显著降低了推理速度，因为它们需要为额外的对比样本计算输出分布。例如，在 VCD 方法中，每个视觉输入 v 都需要分别计算 $p_\theta(y|v,x)$ 和 $p_\theta(y|v',x)$ 的对数概率。与普通解码相比，这使得推理过程中的计算量增加了一倍。作者在 ScienceQA 上评估了 VCD 与普通解码的推理速度。实验结果显示，VCD 的推理时间几乎是普通解码的两倍。
模态融合中的视觉忽视

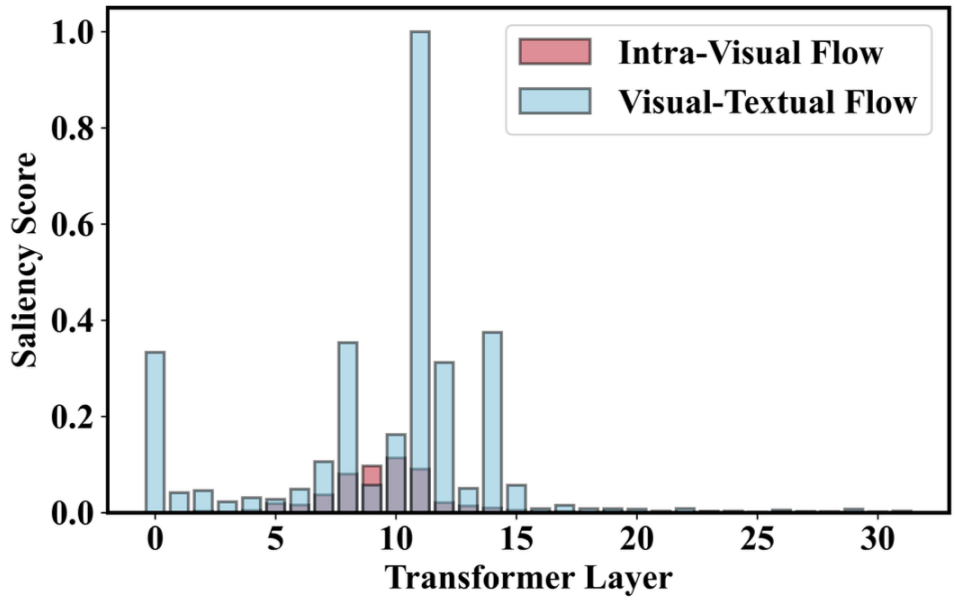

图 1: The importance of intra-visual flow and visual-textual flow across various layers.

本节的主要目的是探究多模态大语言模型在预测时为何倾向于过度依赖语言先验。

首先我们采用显著性技术来突出注意力机制中关键的标记交互，在 POPE 基准下，使用 LLaVA-v1.5-7B 模型在 MS COCO 数据集上进行了实验，选取了 500 个样本用于评估。图中强调了模型中间层（特别是第 8 层到第 15 层）中视觉-文本信息流的关键作用。这一观察结果表明，在这些层中，视觉信息通过注意力机制与文本信息进行深度交互，这对预测结果产生了重大影响。显著性分析表明，图像标记主要通过与中间层中的指令标记相互作用来影响预测结果。

随后比较了不同模态的注意力权重，结果显示，分配给视觉特征的注意力明显低于分配给系统提示和用户指令的注意力。这些发现表明，在模态融合过程中，视觉信息往往未被充分利用，从而导致对语言先验的过度依赖。

视觉增强融合

基于以上提出的见解，作者引入了一种名为视觉增强融合（VAF）的幻觉缓解方法。如图所示，该方法在模态融合过程中增强了对视觉信息的关注，有效减少了对语言先验的过度依赖，并确保生成的内容与视觉输入紧密关联。模型在中间层对视觉和文本模态进行关键融合。然而，在此过程中分配给视觉模态信息的注意力仍然不足。为解决这一问题，作者调整了这些层中的注意力权重，以实现更均衡的关注。令 $A_{l,h}$ 表示第 l 层第 h 个注意力头的注意力矩阵，$Z_{l,h}$ 表示其对应的注意力分数矩阵，定义如下：

$$A_{l,h} = softmax\left(Z_{l,h}\right) \tag{1}$$

在模态融合过程中，目标是增强模型对视觉特征的关注，同时抑制对系统提示的过度强调。这种调整有助于更好地整合视觉信息，并减少对语言先验的过度依赖。为实现这一目标，作者对中间层的注意力分数矩阵（即 $8 < l < 15$）进行了如下修改：

$$\hat{Z}_{l,h} = Z_{l,h} + \alpha \cdot M_{l,h}^{enh} \circ Z_{l,h} - \beta \cdot M_{l,h}^{sup} \circ Z_{l,h} \tag{2}$$

此处， 是增强系数 $((\alpha > 0))$, 其中值越大表明视觉注意力的放大作用越强。抑制系数 $\beta(0 < \beta < 1))$ 决定了针对系统提示的注意力抑制程度。增强和抑制掩码矩阵，分别为 $M_{l,h}^{enh}$ 和 $M_{l,h}^{sup}$，其定义是为了指导注意力元素的调制：

$$\begin{aligned} M_{l,h}^{enh}(i,j) &= \mathbb{I}(i \in \mathcal{T}, j \in \mathcal{V}), \\ M_{l,h}^{sup}(i,j) &= \mathbb{I}(i \in \mathcal{T}, j \in \mathcal{S}). \end{aligned} \tag{3}$$

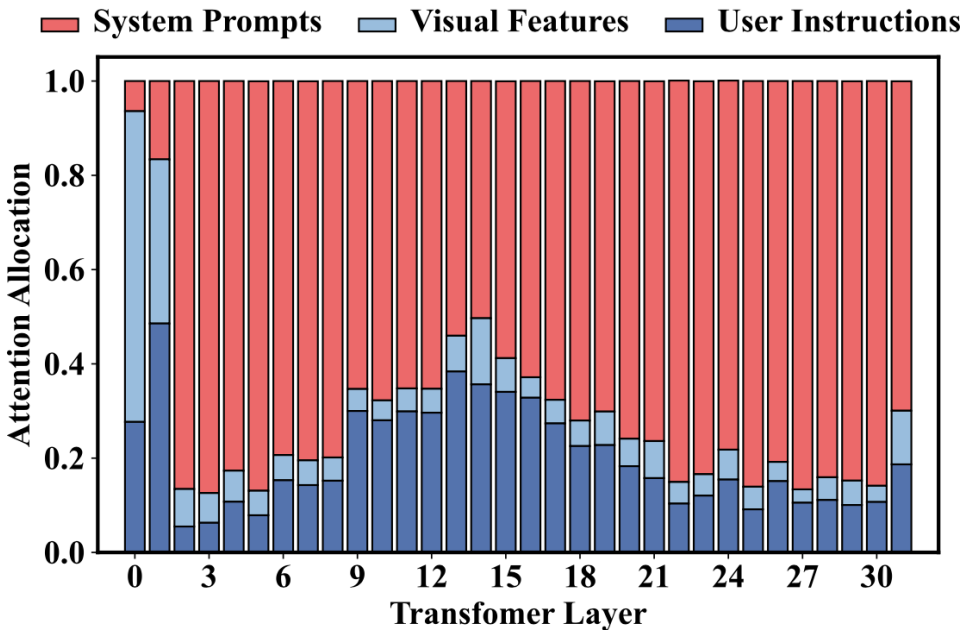

图 2: Attention Distribution of Modal Information Across Model Layers.

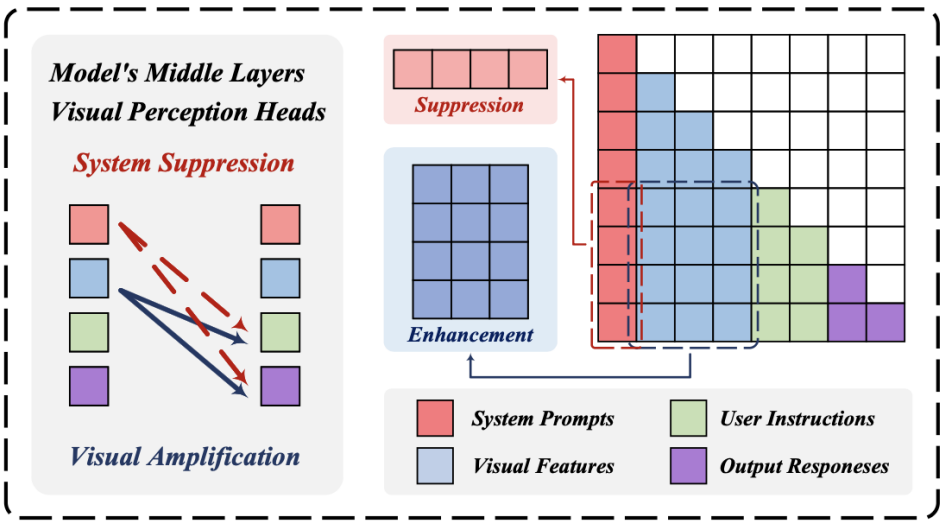

图 3: Illustration of the Visual Amplification Fusion Method.

这些修改通过增强模型在模态融合过程中对视觉特征的关注，并减少对系统提示的多余关注，优化了注意力分配。初步分析表明，这种方法通过促使模型更多地关注视觉信息，有效地缓解了幻觉问题。

在中间层的所有注意力头中增强视觉注意力可能过于激进，并可能对内容生成产生负面影响。为解决这一问题，作者提出了一种选择性增强策略。具体而言，作者识别并分离出对视觉信息表现出更高敏感性的注意力头，将其称为视觉感知头。然后，将视觉注意力增强限制在这些视觉感知头上，以确保在更好地利用视觉信息的同时保持模型的整体性能。

在该模型中，将更多注意力分配给视觉特征的注意力头对视觉信息表现出更高的敏感性。令 $A_{l,h}$ 表示模型第 l 层第 h 个注意力头的注意力矩阵，其对应的视觉注意力分配用 $\lambda_{vis}^{l,h}$ 表示。在每个注意力层中，识别出视觉注意力分配处于前 50% 的注意力头并将它们指定为视觉感知头，随后重新分配它们的注意力。其余注意力头的注意力矩阵保持不变。

(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?】

VAF 目前主要关注中间层的注意力调整，未来可以探索在其他层次或通过其他机制来进一步优化视觉信号的利用。

(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】

本文在多个方面具有借鉴价值。首先，其对 MLLMs 模态融合机制的深入分析，为理解模型如何处理多模态信息提供了新的视角。这种分析方法可以应用于其他多模态任务，帮助研究人员更好地理解模型的行为。其次，VAF 方法的设计思路，即通过增强视觉信号来减少幻觉，为解决类似问题提供了新的方法。这种方法可以启发研究人员探索其他方式来增强模型对特定模态的关注。

(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，通过深入分析模型的内部机制，可以发现潜在的问题并提出有效的解决方案。这不仅适用于多模态领域，也适用于其他人工智能领域的研究。

5 评阅人

姓名:

时间: