【通过全局与局部注意力组合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 with Assembly of Global and Local Attention】

1 相关资源

pdf: https://arxiv.org/abs/2406.12718
ppt:
短视频:
数据集:
源码: https://github.com/Lackel/AGLA.
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: Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, Shijian Lu
单位:

1. Xi'an Jiaotong University
2. National Engineering Laboratory for Big Data Analytics
3. Nanyang Technological University
4. Lenovo Research
5. University of Massachusetts Boston

6. Zhejiang University of Technology

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。

本文提出全局与局部注意力组合 (AGLA)——一种无需训练即插即用的方法，通过同时整合用于响应生成的全局特征和用于视觉判别的局部特征来缓解幻觉。。

(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 在图像描述、视觉问答等多模态任务中取得了显著进展，但对象幻觉问题仍然是一个关键挑战。现有的研究已经从多个角度探讨了对象幻觉的成因，包括预训练数据的统计偏差、模型对参数化知识的过度依赖以及特征学习的偏差等。然而，这些研究大多未能从根本上解决模型对图像中提示相关局部特征的忽视问题。现有的缓解方法，如指令微调、后处理修正器和解码策略改进等，虽然在一定程度上有效，但仍然存在局限性，尤其是在处理复杂图像和多对象查询时，模型的视觉定位能力仍然不足。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

通过引入全局和局部注意力的组合（AGLA），本文旨在同时捕捉图像的全局特征和局部特征，从而提高模型的视觉定位能力和生成文本的准确性。

(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. 系统研究 LVLM 中的物体幻觉现象，首次将注意力缺陷确定为关键诱因，为理解与缓解幻觉提供新视角
2. 提出 AGLA 方法：这种无需训练的即插即用解码方案通过集成提示无关的全局注意力与提示相关的局部注意力，同步捕捉生成性与判别性图像特征，显著提升 LVLM 感知能力
3. 在多种生成与判别基准测试中验证 AGLA 在幻觉缓解方面的卓越性能

(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】

为使 LVLMs 聚焦于与提示相关的图像区域并屏蔽干扰，作者提出图像-提示匹配（IPM）技术来生成输入图像的增强视图。在此基础上，构建全局与局部注意力组装机制（AGLA），通过逻辑融合将原始图像的生成性全局特征与增强图像的判别性局部特征相结合，从而推导出更精确的解码校准分

布。

图像-提示匹配 IPM

大型视觉语言模型（LVLMs）常因注意力缺陷而无法聚焦于与输入提示相关的图像区域。为解决该问题，作者引入图像-提示匹配（IPM）模块，通过识别并遮蔽无关图像区域，确保 LVLMs 专注于提示相关内容。

该模块首先基于匹配模型计算图像 v 与文本提示 t 的整体相似度得分 $\mathrm{sim}(v,t)$。借助可解释性的进展，将 GradCAM 应用于匹配模型的交叉注意力层，得到每个图像块相对于输入提示的关联分数。具体而言，令 $X \in \mathbb{R}^{M\times D_t}$ 表示提示词特征，$Y \in \mathbb{R}^{K\times D_v}$ 表示图像块特征（M 和 K 分别代表提示词数量与图像块数量），交叉注意力矩阵 $C \in \mathbb{R}^{M\times K}$ 可通过如下方式计算：

$$C = \mathrm{softmax}\left(\frac{XW_T W_V^\top Y^\top}{\sqrt{D_t}}\right) \tag{1}$$

其中 $W_T \in \mathbb{R}^{D_t\times D_t}$ 和 $W_V \in \mathbb{R}^{D_v\times D_v}$ 表示交叉注意力头的参数，D_t 和 D_v 分别表示提示词特征与图像块特征的维度。交叉注意力矩阵 C 量化了每个提示词对图像块的注意力分配，C_{ij} 表示第 i 个提示词与第 j 个图像块间的注意力权重。因此，第 j 个图像块相对于整个文本提示的关联分数可计算如下：

$$\mathrm{cor}(j) = \frac{1}{H}\sum_{i=1}^{M}\sum_{h=1}^{H}\max\left(0, \frac{\partial \mathrm{sim}(v,t)}{\partial C_{ij}^{(h)}}\right) C_{ij}^{(h)} \tag{2}$$

式中 H 表示交叉注意力头数量，$C^{(h)}$ 为第 h 个注意力头的交叉注意力矩阵。该偏导数项衡量相似度得分对交叉注意力分数的敏感度，反映各注意力权重的重要性。此过程可计算每个图像块的关联分数，并识别与输入提示最相关的区域。

为确保 LVLMs 聚焦于提示相关图像内容并屏蔽干扰对象，作者提出基于关联分数的自适应遮蔽策略。该策略会遮蔽低关联分数区域，同时保留高分数区域。我们根据整体相似度得分 $\mathrm{sim}(v,t)$ 动态确定遮蔽比例，而非对所有图像和提示采用固定比例。

具体实现中，直接采用 $\mathrm{sim}(v,t)/2$ 作为遮蔽比例，使遮蔽过程能适应不同程度的图像-提示匹配。对于相似度较高的图像-提示对（表明二者相关性更强），将遮蔽更大比例的图像区域以实现有效干扰抑制和幻觉缓解。

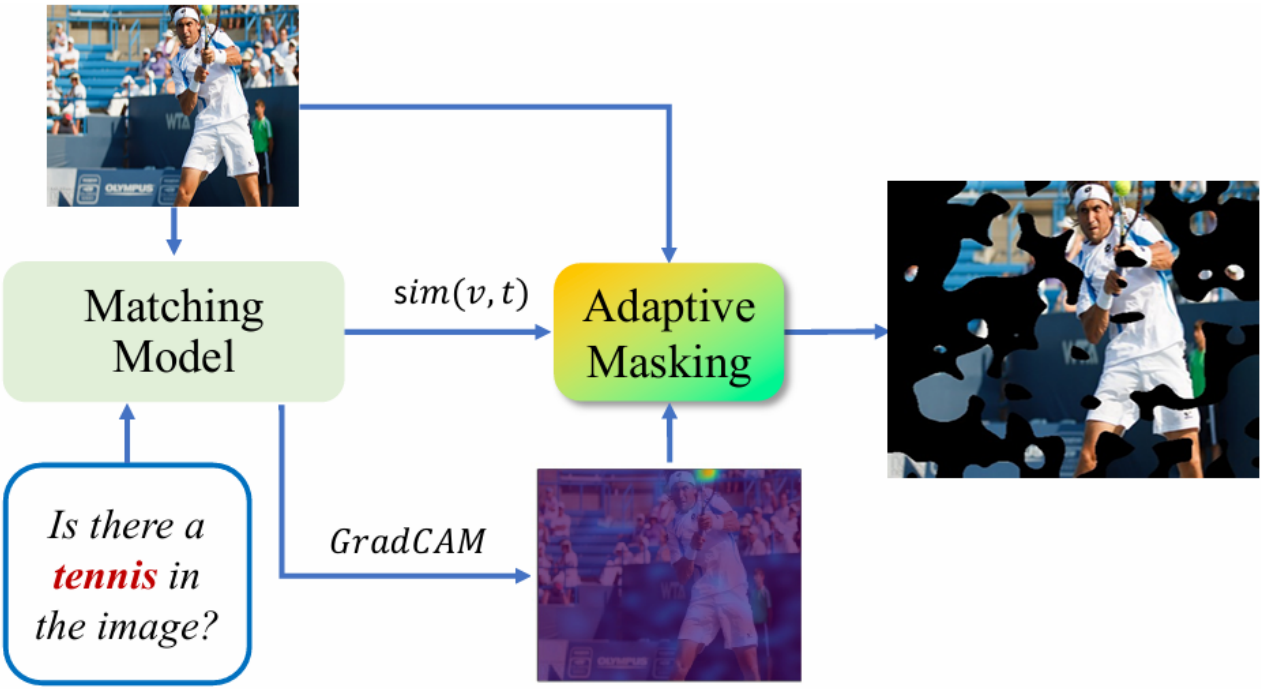

图 1: An illustration of the proposed Image-Prompt Matching.

全局与局部注意力的整合 AGLA

虽然 IPM（局部感知模块）能有效屏蔽干扰并捕获对视觉判别至关重要的局部特征，但不可避免地会因屏蔽生成任务所需的全局特征而导致信息损失。为此，作者设计了 AGLA（自适应全局-局部注意力）机制，通过融合局部与全局注意力来同时捕获判别性和生成性特征，从而学习更全面的图像表征。在每个解码步骤 i，整合原始图像与增强图像产生的逻辑值，获得校准后的解码分布。该策略可表述为：

$$P_{AGLA}(y_i \mid v, v^{\text{aug}}, t, y_{<i}) \sim \text{softmax}\left[\text{logit}_\theta\left(y_i \mid v, t, y_{<i}\right) + \alpha \text{logit}_\theta\left(y_i \mid v^{\text{aug}}, t, y_{<i}\right)\right] \tag{3}$$

其中 y_i 表示解码步骤 i 的标记，$y_{<i}$ 代表步骤 i 前已生成的标记序列。v, v^{aug}、t 分别表示原始图像、增强图像和输入提示。参数 θ 指代 LVLM（大规模视觉语言模型）的模型参数，α 是平衡局部与全局逻辑值贡献的加权系数。与 VCD 通过建模噪声分布并从原始分布中减去的做法不同，作者的模型生成关注局部兴趣区域的有效分布，并将其作为补充加入原始分布以缓解注意力缺失问题。这两种方法具有正交互补性。

自适应合理性约束：先前研究指出，如公式 3 所示校准整个输出分布可能抑制原始分布中的有效输出，同时增强增强分布中的不合理输出。为此，作者采用自适应合理性约束 [**?**]，选择性保留高原始概率的标记并截断其他标记：

$$\mathcal{V}_{\text{token}}(y_{<i}) = \{y_i \in \mathcal{V} : p_\theta(y_i \mid v, t, y_{<i}) \geq \beta \max_w p_\theta(w \mid v, t, y_{<i})\} \tag{4}$$

$$P_{AGLA}(y_i \mid v, v^{\text{aug}}, t, y_{<i}) = 0, \text{if } y_i \notin \mathcal{V}_{\text{token}}(y_{<i}) \tag{5}$$

其中 $\mathcal{V}_{\text{token}}$ 为选定标记集合，$\mathcal{V}$ 是输出词表。β 是控制截断强度的超参数，值越大则仅保留高概率标记。

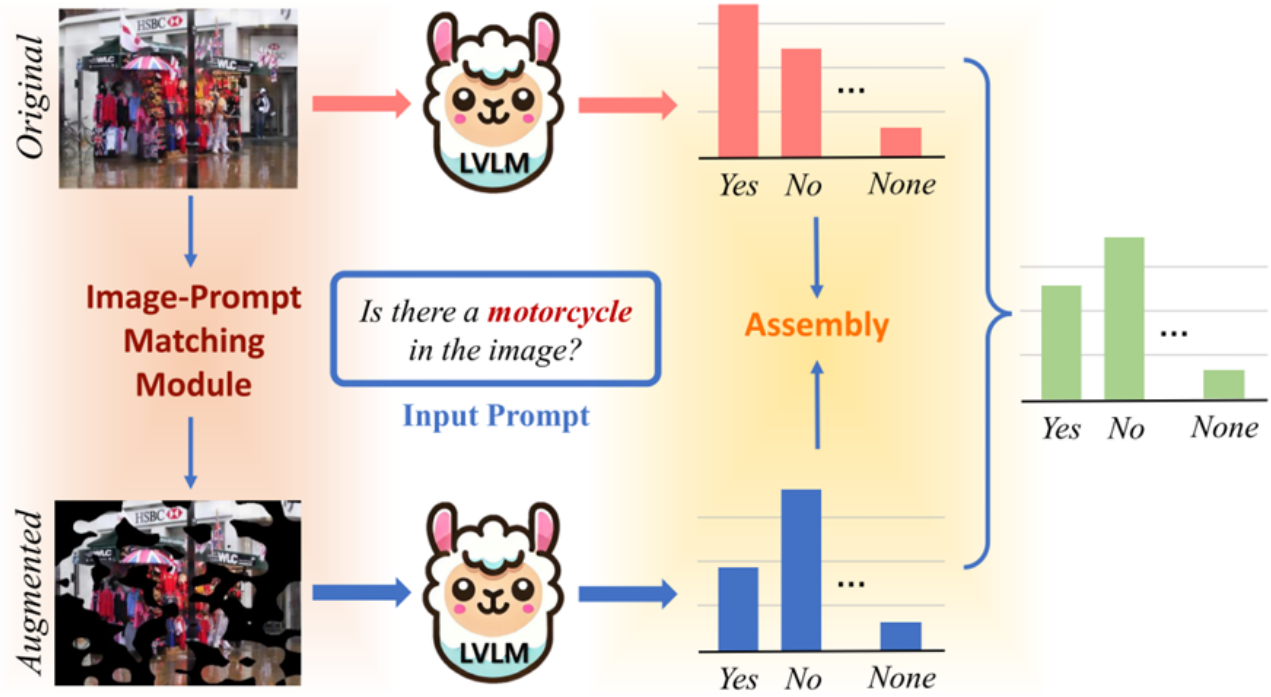

图 2: Decoding with Assembly of Global and Local Attention.

(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 AGLA 在减少对象幻觉方面取得了显著进展，但仍有改进空间。例如，当前的 IPM 模块依赖于 GradCAM 技术来计算图像区域的相关性，这种方法虽然有效，但可能对某些复杂的图像和提示组合不够敏感。未来的研究可以探索更先进的图像-提示匹配技术，以进一步提高模型对局部特征的捕捉能力。

(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】

首先，从注意力不足的角度分析对象幻觉问题，为研究其他类型的幻觉提供了新的思路。其次，AGLA 的即插即用特性使其可以轻松集成到现有的 LVLMs 中，而无需额外的训练，这为快速改进现有模型提供了实用的方法。此外，图像-提示匹配方案和全局与局部特征融合的技术也为其他多模态任务提供了有价值的参考，例如在图像描述、视觉问答和多模态对话系统中，这些技术可以用来提高模型的视觉定位能力和生成文本的质量。

(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……】

我了解到多模态任务中幻觉问题的解决通常依赖于对模型注意力机制的改进，以及对图像和文本特征的更精细处理。

5 评阅人

姓名:
时间: