【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】

——【Multi-Modal Hallucination Control by Visual Information Grounding】

1 相关资源

pdf: https://arxiv.org/pdf/2403.14003
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: 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: Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, Stefano Soatto
单位:

1. AWS AI Labs

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。
本文提出了一种新的方法，通过在生成过程中最大化文本与视觉提示之间的互信息，来减少幻觉现象，从而提高模型在视觉-语言任务中的表现。

(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】
当前的最好研究已经在视觉-语言模型的幻觉问题上取得了一定的进展。例如，一些研究通过复杂且昂贵的对齐算法，涉及直接的人类监督来减少幻觉现象。还有一些研究通过改进解码算法，如上下文感知解码和点互信息（PMI）解码，来提高模型对上下文的依赖，从而减少幻觉。然而，这些方法存在一些问题。首先，依赖人类监督的方法不仅成本高，而且难以扩展。其次，现有的解码算法虽然在一定程度上提高了生成文本的准确性，但它们通常需要额外的训练或对模型结构进行修改，这增加了实现的复杂性和计算成本。此外，这些方法在处理长文本生成时效果不佳，容易出现主题漂移或幻觉现象。本文认为，现有的方法虽然在减少幻觉方面取得了一定进展，但仍然存在改进空间，尤其是在如何更有效地利用视觉信息来指导文本生成方面。

(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生成式视觉-语言模型在生成文本时的幻觉现象，通过增强模型对视觉提示的依赖，提高生成文本的准确性和可靠性。

(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. 提出视觉提示依赖度指标 PDM，用于判断模型输出是否与视觉输入相关联。
2. 提出 M3ID 方法，这是一种在自回归视觉 - 语言模型生成分布上无需训练的干预手段。
3. 实验表明，应用 M3ID 或 DPO 方法后，在图像描述任务中，幻觉物体的比例分别降低了 25% 和 28%；在 POPE 视觉问答幻觉基准测试中，相较于基础模型，准确率分别提升了 21% 和 24%。

(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】
幻觉分析
作者首先研究视觉语言模型 VLMs 中的幻觉现象。作者引入了一种视觉提示依赖度测量（PDM）方法，主要是通过比较模型输出（如一个标记）的可能性与在没有相关条件因素（即图像）的情况下生成相同输出的可能性，来判断该模型输出是否属于幻觉。
用 p 表示基于文本标记词汇 v 的概率生成式视觉语言模型，并将给定文本提示 x 和图像上下文 c 时，标记 $y \in V$ 的概率表示为 $p(y|x,c)$。视觉语言模型的目标是利用输入图像 c 中包含的信息，对输入提示 x 进行续写 $y = [y_0, ..., y_T]$，例如“详细描述这张图像”。有效序列 y 的概率可计算为 $p(y|x,c) = \prod_{t=1}^{T} p(y_t|y_{<t}, x, c)$，其中 $y_{<t} \triangleq [y_0, ..., y_{t-1}]$。

使用如下定义的 PDM 来研究视觉语言模型（VLM）的幻觉问题

$$PDM\left(y_{<t};c|x\right) \triangleq dist\left(p(\cdot|y_{<t},x,c),p\left(\cdot|y_{<t},x\right)\right) \tag{1}$$

其中，dist 是概率分布之间的任何距离度量，这里使用赫尔利距离。PDM 用于量化语言模型输出的通用性或特定语境性。具体而言，较高的 $PDM(y_{<t};c|x)$ 表明标记 y_t 与特定的输入提示密切相关，而较低的 PDM 则表明该标记更具有提示中立性或与提示无关性。根据距离函数的选择，PDM 会突出生成分布的不同方面。

发现一语境压力：(1) 像 “in”（介词）、“and”（连词）、“the”（冠词）这类保证句子语法正确、表达流畅的 token，无论模型是否参考图像（即 “含图像约束的 VLM” vs “不含图像的纯语言模型 LLM”），都能以极高概率预测到。(2) 当模型生成由多个 token 组成的 “细粒度物体名称”（如 “Peanut butter”）时，即使屏蔽图像，仅通过前半部分文本（如 “Peanut bu”），VLM 和纯语言模型 LLM 都能准确预测出最后一个 token（“butter”）。

发现二条件稀释/记忆衰减效应：观察到 PDM-H 随生成 token 数量的增加而持续下降。这一趋势直接反映：模型在生成文本的全过程中，对视觉信息的关注度不断降低，视觉信息的约束作用被逐渐 “稀释”，甚至被模型忽略。并且，随着生成 token 数量增加，PDM-H 逐渐减小，幻觉物体的数量则随之显著增加，说明模型因过度依赖语言先验，开始生成图像中不存在的物体。

M3ID

防止条件稀释：在记忆衰减假设下，作者对条件对数概率进行建模：

$$l(y_t|y_{<t},x,c) = \gamma_t l^*\left(y_t|y_{<t},x,c\right) + (1-\gamma_t)\, l\left(y_t|y_{<t},x\right) \tag{2}$$

其中 $\gamma_t \in [0,1]$ 是一个随时间单调递减的混合系数。当 γ_t 较小时，条件分布在不提供输入图像的情况下也能得到大部分解释，且条件分布会 “忘记” 该图像，而当 γ_t 较大时，输入图像则变得更为相关。取 $\gamma_t \triangleq \exp(-\lambda t)$， 模拟 γ_t 的变化率，并定义了记忆衰减率（遗忘率 [14, 30]）

根据来自条件分布 $l_c \triangleq l(y_t|y_{<t},x,c)$ 和无条件分布 $l_u \triangleq$、$l(y_t|y_{<t},x)$ 的观测结果，目标是找到潜在生成分布 l^* 的估计值 $\hat{l}^*$，该估计值不会遗忘过去的信息。

为此，假设 l^* 是条件分布 $l^* = l_c + \Delta$ 的一个扰动，其中 Δ 可以被假设为一个均值为零且方差有界的随机变量。因此，将式（2）中的模型重写为：

$$(1-\gamma_t)(l_c - l_u) = \gamma_t \Delta. \tag{3}$$

在时间索引 t 上，式（3）是一个跨标记的随机过程，其方差随时间减小。因此，可以估计对 l_c 的最佳干预，以抵消记忆衰减效应，并在 $\hat{l}^* = l_c + \hat{\Delta}$ 时更接近 l^*。我们使用式（3）的测量结果来估计校正项 $\hat{\Delta}$。这使得到最佳干预：

$$\hat{l}^* = l_c + \frac{1-\gamma_t}{\gamma_t}\left(l_c - l_u\right) \tag{4}$$

适应语境压力：然而，无论生成的标记数量如何，语境压力都迫使 l_c 和 l_u 变得相似。因此，通过不加区分地放大 $l_c - l_u$ 来惩罚语言先验 l_u，有时也会惩罚正确但不需要对输入图像进行推理的明显标记（如介词或连词）。为避免这种情况，当上下文压力较大时，会抑制对式（4）的干预。具体而言，如果条件模型 l_c 对下一个标记有很高的置信度（即其最大概率高于阈值 ），就不应用校正项 $l_c - l_u$。多模态互信息解码 M3ID：综合所有因素，用于生成的多模态互信息解码算法 1 如下：

$$y_t = argmax_{y\in\mathcal{V}} l^*\left(y|y_{<t},x,c\right) \tag{5}$$

其中 $\hat{l}^* = l_c + 1[max_k(l_c)_k < log\alpha]\frac{1-\gamma_t}{\gamma_t}(l_c - l_u)$ M3ID 可应用于不同的搜索算法，如贪婪搜索或波束搜索。

M3ID+DPO

有了计算资源和模型权重的访问权限，可以对模型进行优化，使其输出的续写内容更贴合图像内容。在本节中，将这一目标重新表述为一个偏好优化问题，其目标是偏好贴合图像的续写内容而非不贴合的，并利用这一目标对视觉语言模型进行微调。

多模态偏好优化：考虑针对图像 c 的同一提示词 x 的两个不同续篇 y_w 和 y_l，其中 y_w 比 y_l 更受青睐。直接偏好优化（DPO）是一种对齐技术，旨在学习一种生成策略，该策略更有可能生成类似于 y_w 而非 y_l 的续篇。具体而言，给定包含偏好数据对 (y_w, y_l) 的数据集 D，DPO 最小化以下损失函数：

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(c,x,y_w,y_l)\sim\mathcal{D}}\left[log\sigma\left(\beta log\frac{p_\theta\left(y_w|c,x\right)}{p_{ref}\left(y_w|c,x\right)} - \beta log\frac{p_\theta\left(y_l|c,x\right)}{p_{ref}\left(y_l|c,x\right)}\right)\right], \tag{6}$$

其中，p_{ref} 表示 DPO 训练前的视觉语言模型（VLM），p_θ 表示正在训练的模型。简而言之，这一目标背后的思路是，相对于基础模型 p_{ref}，提高续篇 y_w 的可能性，同时相对于基础模型降低生成 y_l 的可能性。

生成多模态偏好数据 DPO 的成功取决于数据集 D 中配对的质量：因此，建议使用 DPO 目标对预训练的 VLM 进行微调，同时确保更优的续写内容与视觉信息有更强的关联性。为此，通过从经 M3ID y_w $p^*(y|x,c)$ 改进的预训练条件分布中采样来生成更优的续写内容，而从无条件分布中采样来生成负面的续写内容 y_l $p(y|x)$ 。然而，请注意，使用无条件模型从 x（例如，“描述这张图像。”）生成文本续篇会产生负续篇 y_l，这些负续篇往往与图像标题大相径庭，因此在 DPO 目标中提供的信号很少。为了克服这一限制，在 x 后附加了条件 VLM 生成的第一句话，以限制可能的续篇范围。

(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?】

本文的局限性在于 M3ID 方法需要在生成过程中进行两次前向传播，一次用于条件预测，一次用于无条件预测，这可能会增加推理时间。虽然可以通过使用批量查询来减少推理时间，但这会增加内存消耗。

(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】

本文提出的视觉提示依赖度量（PDM）为评估模型输出与视觉提示的相关性提供了一种量化的方法，可以用于其他视觉-语言任务中，帮助研究人员更好地理解和分析模型的行为。

(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……】

本文展示了互信息在生成任务中的重要性，通过最大化文本与视觉提示之间的互信息，可以有效减少幻觉现象，提高生成文本的准确性和可靠性。

5 评阅人

姓名:

时间: