【通过潜空间引导减少大型视觉语言模型中的幻觉现象】

——【Reducing Hallucinations in Vision-Language Models via Latent Space Steering】

1 相关资源

pdf: https://arxiv.org/pdf/2410.15778
ppt:
短视频:
数据集:
源码: https://github.com/shengliu66/VTI
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: ICLR 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: Sheng Liu, Haotian Ye, James Zou
单位:

1. Stanford 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。
本文旨在通过研究 LVLMs 中幻觉的产生机制，提出一种新的方法来减少幻觉现象，同时保持模型生成文本的详细性和准确性。

(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，减少 LVLMs 幻觉的研究已经取得了一定的进展。现有的方法主要分为三类：后处理和自我修正技术、基于人类反馈的方法以及解码策略方法。后处理技术通常依赖于额外的数据集和训练，或者需要高性能的外部 LVLMs 来实现。基于人类反馈的方法虽然能够提高模型的可靠性，但需要大量的标注数据，成本较高。解码策略方法则通过对比解码策略来减少幻觉，但这些方法通常需要多轮解码和回滚，显著降低了解码速度。例如，最新的 HALC 方法虽然在减少幻觉方面表现出色，但平均解码速度大幅下降。因此，现有的方法虽然在减少幻觉方面取得了一定的进展，但在效率和实用性方面仍存在不足。本文认为，需要一种更高效的解决方案，能够在减少幻觉的同时保持高效的解码速度。

(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大型视觉-语言模型在生成文本时的幻觉现象，同时保持高效的解码速度和生成文本的详细性。

(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. 通过关注视觉编码器和文本解码器之间的序列关系来研究幻觉机制，发现了视觉特征的稳定性与大型视觉语言模型（LVLMs）幻觉之间的相关性；
2. 提出了一种新方法 VTI，该方法通过在潜在空间中引导大型视觉语言模型，使其输出更少的幻觉内容；
3. 进行了大量实验，将 VTI 与基准方法进行对比评估。在多个指标上的领先结果表明，VTI 在减少幻觉方面具有很高的有效性。

(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】
大视觉语言模型中的幻觉机制
首先强调大型视觉-语言模型中视觉和文本组件之间的一种基本关系：视觉编码器的输出充当语言解码器的输入。这种顺序连接意味着视觉特征的稳定性在模型输出中起着至关重要的作用，并且会影响幻觉的产生。为了研究特征稳定性与对象幻觉之间的联系，用各种类型的噪声干扰原始图像，并分析由此产生的特征分布的方差。理想情况下，如果视觉编码器具有鲁棒性且经过有效捕捉语义信息的训练，那么不会改变图像语义内容的噪声应该对视觉特征或模型输出产生最小影响。
然而，如图左图所示，虽然大多数视觉特征保持稳定，但约 15% 的特征存在显著差异，形成了长尾分布。这些不稳定的特征与幻觉密切相关，因为模型对这些特征过于敏感，导致其输出出现偏差。第二幅图进一步证明了这种关联：对多张注入噪声的图像的视觉特征进行平均，随着扰动次数的增加，无论注入的噪声类型如何，幻觉都会减少。重要的是，如右图所示，这种减少并非源于噪声本身，而是源于对多次扰动的平均过程，而仅添加噪声往往会增加幻觉。

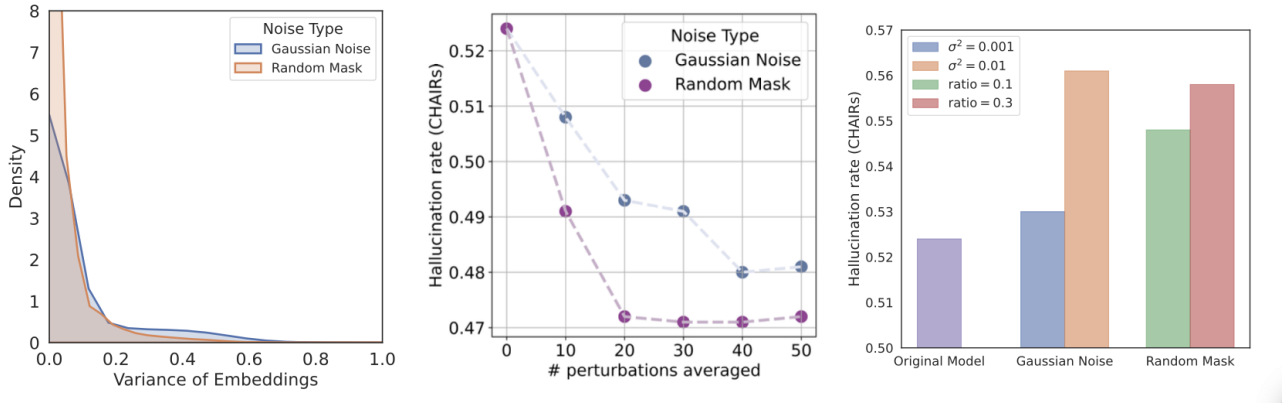

图 1: (Left) Distribution of vision feature stability when different types of noise are injected into the raw images. The x-axis represents the variance of features across 50 perturbations, and the y-axis represents the frequency. (Middle) Illustration of the correlation between object hallucination and vision feature stability. Averaging vision features across multiple perturbed images reduces hallucination (measured by CHAIR, details in Section A) as the number of perturbations averaged increases. (Right) Noise alone tends to increase hallucination, suggesting that the reduction of hallucination is not due to the noise itself but to the averaging process across perturbations.

尽管特征平均化能够减轻幻觉现象，但它也带来了明显的缺点。由于视觉编码器是在干净数据上训练的，对图像进行扰动会影响信息提取，导致细节丢失。此外，对视觉特征进行平均化需要模型进行多次前向传播，这大大增加了计算成本。因此，我们的目标是开发一种方法，在不牺牲信息或产生过多计算开销的情况下，有效增强视觉特征的稳定性并减少幻觉现象。

VIT

上一节中的实验表明，尽管存在副作用，但对轻度损坏图像的特征进行平均后，这些特征具有鲁棒性，并且能有效减少幻觉现象。为避免这种情况，受大语言模型（LLMs）表示工程相关研究的启发——这些研究通过在推理过程中编辑潜在空间中的特征来改变大语言模型的行为，作者开发了一种计算高效的算法，称为视觉和文本干预（VTI），该算法可以提高视觉特征的稳定性以及文本与图像的依赖性，从而减少视觉-语言模型（LVLMs）的幻觉现象。具体而言，预先计算了更稳定特征的“方向”，然后在推理过程中，将它们一致地应用于所有查询示例，以减少幻觉，且不会增加额外的训练或推理成本。由于幻觉有时源于文本解码器（即大语言模型），进一步获取了一个文本方向，并将其应用于文本解码器，以最大限度地提高性能。所提出方法的概述如图所示。给定视觉输入 v，令视觉编码器针对 v 的潜在状态表示为 $h^v_{l,t}$，其中 $l \in 1, 2, ..., L$ 表示编码器中的层索引，$t \in 1, 2, ..., T$ 表示视觉令牌的索引。

为了增强潜在表示的鲁棒性，作者利用随机掩码创建 v 的多个扰动版本。具体来说，将 m 个不同的随机掩码 c_i、$i = 1, ..., m$ 应用于 v，生成 m 个损坏版本 $C_i(v)$。对于每个扰动输入 $C_i(v)$，视觉编码器会生成相应的潜在状态 $h^{C_i(v)}_{l,t}$。

直观地说，v 的稳健潜在嵌入可以通过对从扰动输入中获得的嵌入进行平均来近似。这个平均嵌入的计算方式如下：

$$\overline{h}^v_{l,t} = \frac{1}{m} \sum_{i=1}^{m} h^{\mathcal{C}_i(v)}_{l,t} \tag{1}$$

视觉偏移向量定义为鲁棒平均嵌入与原始嵌入之间的差值：

$$\Delta_{l,t}^{v} = \overline{h}_{l,t}^{v} - h_{l,t}^{v}. \tag{2}$$

为了使这个偏移向量适用于新的图像查询，旨在从 $\Delta_{l,t}^{v}$ 中去除特定于图像的信息，只保留由特征平均引入的通用效果。为实现这一目标，为一组 N 个示例图像 $v_1, v_2, ..., v_N$ 计算偏移向量 $\Delta_{l,t}^{v_i}$。通过将这些向量堆叠成一个矩阵，

$$\left[\Delta_{l,t}^{v_1}, \Delta_{l,t}^{v_2}, ..., \Delta_{l,t}^{v_N}\right] \tag{3}$$

提取该矩阵的第一个主方向，记为 $d_{l,t}^{vision}$。这个主方向捕捉了由特征平均所引入的主要变化模式。值得注意的是，对于每张图像，用于生成扰动输入 $C_i(v)$ 的随机掩码是独立采样的，这确保了扰动的多样性，并提高了计算出的偏移向量的通用性。

除了视觉令牌偏移之外，进一步引入了文本偏移向量，该向量在生成模型输出时会引导文本解码器的潜在状态。获取文本偏移向量的方法与先前研究中提出的对齐大型语言模型风格的方法一样简单：借鉴前人做法，作者精心挑选了 N 个无幻觉的图像标题，记为 x，并采用 GPT 模型生成其幻觉版本 $\bar{x}$。由此，得到了成对的有幻觉和无幻觉的标题。我们直接使用原始对应的图像 v 作为视觉输入。然后，为每个样本计算文本方向，公式如下：

$$\Delta_{l,t}^{x_i,v_i} = h_{l,t}^{x_i,v_i} - h_{l,t}^{\overline{x}_i,v_i} \tag{4}$$

其中，$h_{l,t}$ 表示生成输出的最后一个令牌时，第 l 层中第 t 个文本令牌的隐藏状态。特别地，由于文本解码器采用因果建模，仅使用最后一个令牌的潜在状态，即 t = 最后一个 token。同样，通过主成分分析（PCA）来消除特定示例选择带来的额外噪声，以获得整体方向 $d_{l,t}^{text}$。

分别将视觉方向和文本方向应用于干预视觉编码器和文本解码器。由于视觉编码器不是因果建模的，在正向传播中对视觉编码器所有层在所有标记位置的潜在状态进行偏移：

$$h_{l,t}^{v} := h_{l,t}^{v} + \alpha \cdot d_{l,t}^{vision} \tag{5}$$

对于文本，使用文本方向对文本解码器的潜在状态进行偏移：

$$h_{l,t}^{x,v} := h_{l,t}^{x,v} + \beta \cdot d_{l,t=lasttoken}^{text} \tag{6}$$

(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?】

虽然它在减少幻觉方面表现出色，但在某些情况下可能会过度干预，导致生成的文本过于简短，缺乏细节。

(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】

VTI 从视觉特征稳定性的角度出发，为减少幻觉提供了一种新的视角。这种视角可以应用于其他生成任务中，如文本摘要、对话系统等，帮助减少生成文本中的幻觉现象。其次，VTI 提出的视觉和文本干预方法为优化生成策略提供了一种新的思路，可以用于其他需要优化生成策略的任务中，如强化学习、策略优化等。

(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……】

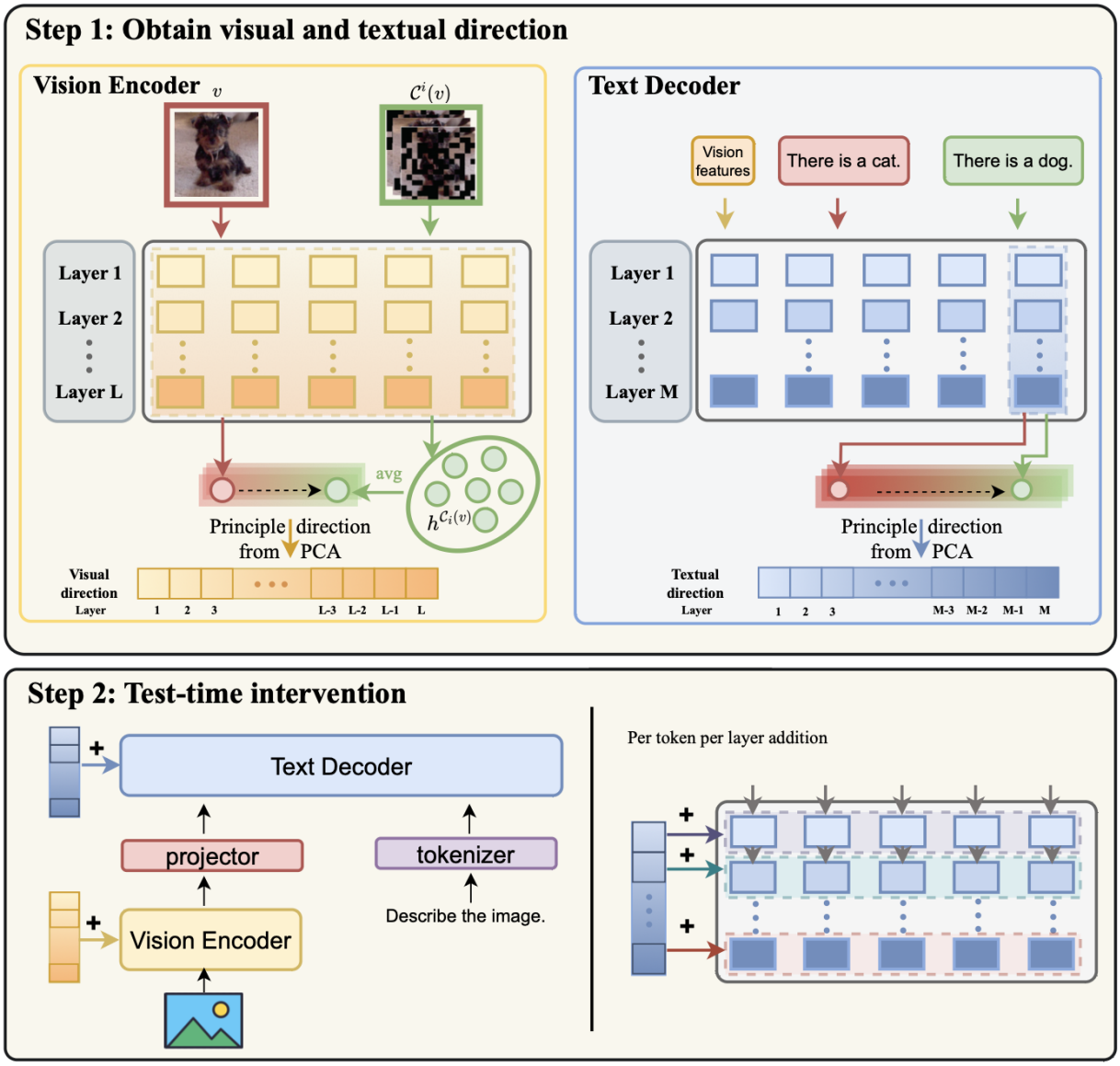

图 2: Overview of the proposed algorithm visual and textual test-time intervention (VTI).

本文展示了特征工程在生成任务中的重要性，通过在推理过程中调整潜在空间表示，可以有效减少幻觉现象，提高生成文本的准确性和可靠性。这种思路可以应用于其他生成任务中，为减少幻觉现象提供一种新的视角。

5 评阅人

姓名:
时间: