【VASparse：通过视觉感知 token 稀疏化实现高效视觉幻觉缓解】

——【VASparse: Towards Efficient Visual Hallucination Mitigation via Visual-Aware Token Sparsification】

1 相关资源

pdf: https://arxiv.org/pdf/2501.06553
ppt:
短视频:
数据集:
源码: https://github.com/mengchuang123/VASparse-github.
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: Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, Yuexian Zou
单位:

1. Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University
2. School of Electronic and Computer Engineering, Peking 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。

为了解决这一问题，本文提出了一种高效的解码算法——VASparse（Visual-Aware Sparsification），旨在通过视觉感知的稀疏化方法减少视觉幻觉，同时保持高效的解码速度。该方法能够在不牺牲语言流畅性的情况下，显著提高模型在视觉-语言任务中的表现，使其更适合实际应用。

(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，减少视觉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视觉幻觉，同时保持高效的解码速度。

(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. 从解码过程中令牌稀疏化的角度探索 VH 缓解方法，并提出一种新颖、高效的一种高效、即插即用的方法，它同时实现了模型保真度和效率，将令牌稀疏性和视觉感知增强统一为一个优化问题。
2. 提出了一种新颖的视觉感知令牌选择策略，以及一种基于稀疏性的视觉对比解码方法，以缓解视觉幻觉（VH）。该方法利用嵌入来实现对比性的对数几率，并避免了多轮解码。
3. 全面的实验和评估表明，VASparse 在性能和解码速度上均显著优于现有的 VH 缓解方法。

(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】

观察

1. 大型视觉语言模型注意力中的稀疏激活：

如图 a 所示，作者发现，注意力分数呈现出明显的长尾分布，在解码过程中只有一小部分令牌被高度激活。图 a 中的结果表明，仅保留注意力分数最高的前 1% 令牌，就能召回超过 98% 的总注意力

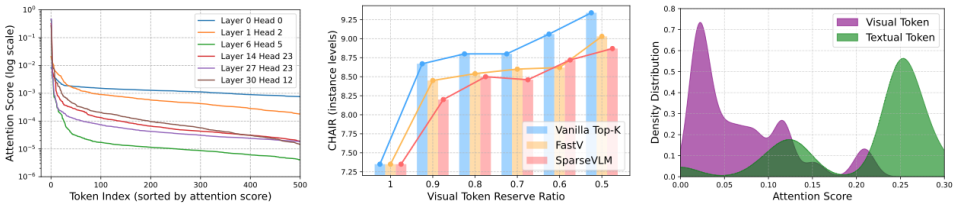

图 1: VH evaluation and attention analysis using LLaVA-1.5 on the CHAIR benchmark: (a) token sorting by attention score; (b) token sparsification effects observed with Vanilla Top-K, FastV, and SparseVLM on sampled 500 images from the MSCOCO validation set, where Vanilla Top-K denotes keeping tokens with top-K scores in 1-th layer; and (c) attention density distribution across various tokens.

分数。这意味着大型视觉语言模型解码器的大多数层中的注意力是稀疏的。

研究结果证实，大语言视觉模型解码器的大多数层中的自注意力是稀疏的。这一发现表明，通过剪枝相应的令牌有可能降低解码过程中的计算成本。

2. 与视觉无关的稀疏化加剧视觉幻觉：

采用与视觉无关（解码过程中不调整 token 选择）的 token 稀疏化方法来评估视觉幻觉，包括常规的 Top-K 策略、FastV 和 SparseVLM。如图 b 所示，发现随着稀疏化程度的提高，模型更易产生视觉幻觉。

实证研究结果表明，这些与视觉无关的稀疏化技术会加剧大型视觉语言模型中的视觉幻觉，这意味着仅仅应用此类方法来加速解码可能会损害输出的可信度。

3. 图像和文本标记的独特分布：

作者分析了视觉和文本标记的注意力分布，结果如 c 所示。分布上存在明显差异：图像标记主要占据低注意力区域，而文本标记则集中在高注意力区域。

这些研究结果表明，在解码过程中，大型视觉语言模型往往会优先处理文本标记而非图像标记。这解释了为什么与视觉无关的标记稀疏化策略可能会加剧幻觉现象：它们更有可能修剪低注意力的图像标记，而这可能会包含关键的视觉信息。这一见解凸显了在稀疏化过程中提高模型对图像标记的感知度可能带来的好处。

4. 文本标记上的注意力下沉：

观察：作者进一步分析了大型视觉语言模型中的注意力模式，发现在某些文本标记中存在显著的注意力“sink ”效应。这种现象类似于在大型语言模型中观察到的摘要标记和注意力偏差效应。然而，与大型语言模型不同的是，研究结果表明，在大型视觉语言模型中，注意力 sink 标记主要集中在文本标记中，即使文本标记的数量远少于图像标记。值得注意的是，这些注意力 sink 标记通常语义含量较低，例如 <.> 和 <s>。

在大型视觉语言模型中，存在注意力下沉现象的标记表现出高关注度但语义信息低的特点。这种模式表明大型视觉语言模型存在内在偏差。然而，过度关注低语义标记可能会导致模型严重依赖语言先验知识，而忽视视觉信息。因此，对这些下沉标记施加惩罚可能会增强大型视觉语言模型对视觉标记的感知能力。

问题描述

定义 1（统一目标）：作者将大视觉语言模型中可信度与效率的联合目标定义为以下约束优化问题的解：

$$\begin{aligned} min_M \quad & \mathcal{E}(M) = \left\| qK^\top - q(M \odot K)^\top \right\|^2 - \lambda P \cdot M \\ & = \sum_{i=1}^{L} \left(\langle q, K_i \rangle - M_i \langle q, K_i \rangle \right)^2 - \lambda P_i \cdot M_i \end{aligned} \tag{1}$$

其中，$q \in \mathbb{R}^{1\times D}$、$K_i \in K$ 和 $K_i \in \mathbb{R}^{1\times D}$，$\|\cdot\|^2$ 表示 L_2 范数。$< \cdot, \cdot >$ 表示内积，S 是稀疏率，是用于平衡视觉感知和注意力召回的权衡参数。

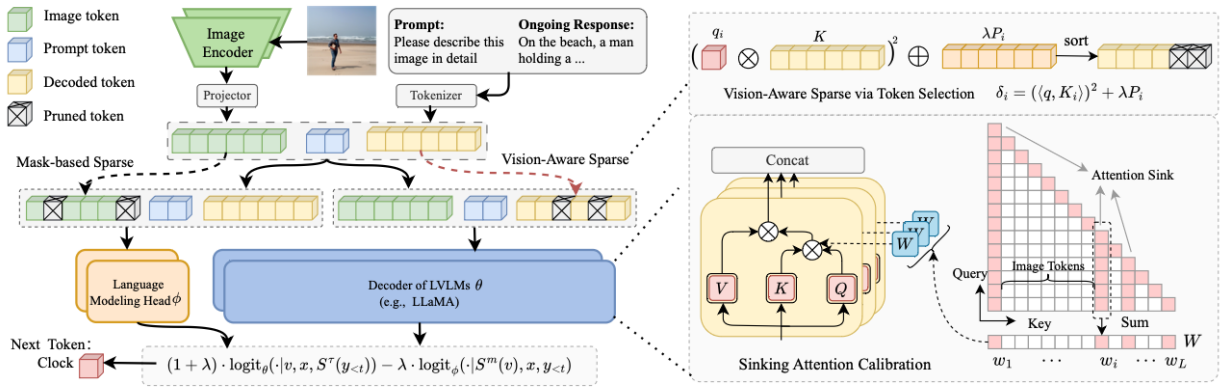

图 2: The illustration of the proposed VASparse framework, which consists of (1) the visual-aware token selection designed to prune the generated tokens during decoding; (2) a sparse-based visual contrastive decoding method to recalibrate the distribution of hallucinated outputs; and (3) the calibration strategy for punishing sinking attention.

目标 1 本身包含以下约束条件：（1）稀疏性约束：$\sum_{i=1}^{L} M_i = S$，其中 S 表示 M 中非零元素的数量，且 $M_i \in 0,1$；（2）视觉显著性约束：$P = P_{i=1}^{L}$ 代表视觉感知分数。为了高效解决该问题 1，作者提出了一种新颖的视觉感知令牌选择策略，以实现高效的 VH 缓解，整体框架如图所示。

视觉感知令牌选择

为了解决统一目标（定义 1）并有效缓解 VH，作者提出了一种视觉感知的令牌选择策略。具体而言，对于每个注意力头，基于聚合分数 δ_i 对令牌进行降序排序，为前 S 个令牌设置 $M_i = 1$，为其余令牌设置 $M_i = 0$。为每个令牌提出的聚合分数 δ_i 定义为：

$$\delta_i = (\langle q, K_i\rangle)^2 + \lambda P_i, \tag{2}$$

其中，$< \cdot, \cdot >$ 表示内积，分数 δ_i 结合了注意力分数 $< q, K >$ 和视觉显著性 P_i，确保在保持计算效率的同时保留视觉相关的令牌。

为了获得视觉感知分数，利用每个生成标记和图像标记的注意力分数，这些分数被视为相应标记的视觉显著性分数。具体来说，计算通过保留大型视觉语言模型（LVLM）历史计算中最后一个注意力头的权重，得到视觉显著性分数 P：

$$P_i = \frac{exp\left(\sum_{k\in\mathcal{I}(v)} a_{i,k}\right)}{\sum_j exp\left(\sum_{k\in\mathcal{I}(v)} a_{j,k}\right)} \tag{3}$$

其中，$I(v)$ 表示图像标记集，$a_{i,j}$ 是标记 $\dot{z}$ 和 j 之间的注意力分数。通过将图像令牌的注意力分数用作重要性度量，可以有效利用已计算的注意力权重，同时避免引入额外的计算开销。对于丢弃的令牌集 $T = K_i|M_i = 0$，采用 k 近邻密度峰值聚合算法 [34] 来实现自适应令牌聚合。同一聚类中的令牌会被求和并保留为单个聚合令牌。

基于稀疏的视觉对比解码

基于实证观察，作者利用与视觉无关的 token 稀疏化会加剧视觉-语言偏差（VH）这一发现，来减轻输出分布中的语言偏差。作者创新性地提出，通过对比视觉感知和与视觉无关（基于掩码）的稀疏化（S^τ 和 S^m）的解码概率分布，对输出中的 logit 进行重新分配，从而增强视觉语境中的信息对比度。

然而，直接使用视觉-语言大模型（LVLMs）的输出分布来获取对比性 logit 分布，会因二次解码过

程不可避免地带来显著的开销。为解决这一问题，作者提出仅将与视觉无关的 token 的嵌入作为输入，传入大语言模型解码器的语言解码头 以获取 logit 分布，而无需经过完整的文本解码器。具体而言，采用所提出的视觉感知稀疏化策略来获取 logit 分布 $logit_\theta$。然后，随机掩码视觉 token，并将其嵌入直接输入大语言模型的语言解码头，以获取对比性 logit 分布 $logit_\phi$。最后，对 token 的 logit 分布进行分配，得到最终结果：

$$\begin{aligned} y_t \sim (1+\alpha) \cdot logit_\theta \left(\cdot | v, x, S^\top(y_{<t})\right) \\ -\alpha \cdot logit_\phi \left(\cdot | S^m(v), x, y_{<t}\right), \end{aligned} \tag{4}$$

其中， 是一个权衡参数。需要注意的是，该解码策略绕过了视觉-语言大模型的解码器，从而避免了二次计算开销。同时作者将自适应合理性约束应用于基于稀疏化的视觉对比解码中。

下沉注意力惩罚

观察结果表明，在大型视觉语言模型中存在明显的注意力下沉现象，即某些标记尽管语义信息较少，却获得了不成比例的高注意力分数。在解码过程中，过度关注这类标记可能会模糊视觉信息。因此，有针对性的应对具有异常高注意力分数的标记应施加惩罚。作者定义了一个惩罚权重矩阵 $W = w_1, \cdots, w_L$，其中每个 w_i 都作为异常注意力分数的惩罚因子。为了有效地实施针对下沉注意力的惩罚，将每个标记的注意力分数与后续查询累积起来，以评估下沉程度。然后，应用 softmax 归一化来获得下沉注意力的校准权重：

$$w_j = \frac{exp\left(\sum_{i=j}^{L} a_{i,j}\right)}{\sum_{k=1}^{L} exp\left(\sum_{i=k}^{L} a_{i,k}\right)} \tag{5}$$

其中，$a_{i,j}$ 表示注意力矩阵第 i 行第 j 列的元素，w_j 表示应用 softmax 操作后权重向量 W 的第 j 个元素。这种方法确保在后续查询中逐步评估下沉注意力，并且在解码过程中，W 将作为 $(1+\beta)qK^\top - \beta W \odot qK^\top$ 那样的权重被使用。

(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稀疏化，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】

首先，VASparse 从视觉感知的稀疏化角度出发，为减少视觉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，VASparse 提出的视觉对比解码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……】

本文展示了稀疏化在生成任务中的重要性，通过在解码过程中动态调整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: