【通过自适应约束信息流缓解大型视觉语言模型中的幻觉现象】

——【Mitigating Hallucinations in Large Vision-Language Models by Adaptively Constraining Information Flow】

1 相关资源

pdf: https://ojs.aaai.org/index.php/AAAI/article/view/34512
ppt:
短视频:
数据集:
源码: https://github.com/jiaqi5598/AdaVIB
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: AAAI 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: Jiaqi Bai, Hongcheng Guo, Zhongyuan Peng, Jian Yang, Zhoujun Li, Mohan Li, Zhihong Tian
单位:

1. Cyberspace Institute of Advanced Technology, Guangzhou University, China
2. Huangpu Research School of Guangzhou University, China
3. CCSE, Beihang University, China
4. University of the Chinese Academy of Sciences, 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 LVLMs 在理解视觉信息时存在物体幻觉的问题，即生成的图像描述中包含图像中不存在的物体。这种幻觉现象严重影响了 LVLMs 的可靠性和适用性，尤其是在需要精确判断的场景中。

本文旨在通过适应性地约束信息流来缓解这种幻觉现象，提高 LVLMs 生成图像描述的准确性和可靠性。

(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】

当前的研究中，许多工作致力于通过减少 LVLMs 对先验知识的过度依赖来缓解物体幻觉问题。这些方法包括对比解码、数据增强以平衡共现模式，以及设计对齐器来缩小视觉和语言之间的模态差距。

然而，这些方法大多没有关注如何精确地压缩视觉模式中的无关特征，同时保留与输入图像相关的视觉信息。此外，尽管现有的视觉编码技术在各种视觉理解任务中表现出色，但在精确表达视觉模式方面仍面临挑战。因此，如何在将视觉特征映射到语言模型的词嵌入空间时避免过度自信，是一个亟待解决的问题。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文通过适应性地约束信息流，缓解 LVLMs 在将软视觉标记映射到语言模型词嵌入空间时对无关视觉特征的过度自信，从而减少物体幻觉现象。

(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. 首次将 VIB 用于缓解目标幻觉问题，降低软视觉标记映射到 LLM 词嵌入时对无关视觉特征的过度自信
2. 提出 ADAVIB 这一基于熵的噪声控制策略，根据与 LLM 词嵌入相似度分布的平滑度，自适应约束视觉标记传递的信息量
3. 综合实验证明 ADAVIB 在缓解目标幻觉方面的有效性，该方法在不同模型架构的两个目标幻觉基准测试上均取得稳定提升

(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】

问题描述

问题为一个通用的图像到文本生成问题，给定数据集 $\mathcal{D} = \{(x_i, q_i, y_i)\}_{i=1}^{N}$，其中第 i 个三元组 (x_i, q_i, y_i) 包含图像 x_i、文本提示 q_i 和图像描述 y_i。目标是学习概率分布 $p(y \mid x, q)$ 的参数 $\mathcal{D}$，从而在图像到文本生成中，给定新样本 (x, q) 时能根据 $p(y \mid x, q)$ 逐标记生成响应 y。通过最大化

以下条件概率来形式化优化 $p(y \mid x, q)$：

$$p(y \mid x, q) = \prod_{t=1}^{|y|} p(y_t \mid y_{<t}, x, q) \tag{1}$$

大型视觉语言模型 LVLM

从图像到文字的过程：给定输入图像，预训练视觉编码器将图像编码为视觉标识，然后用视觉-语言连接器获得中间表示（可选），再使用视觉-语言投影器将视觉标识映射到 LLM 的词嵌入向量生成视觉标记序列。最后将文本提示嵌入，将聚合的视觉标记与其拼接后逐标记生成输出。

损失函数如下：

$$\min \mathcal{L}_{CE} = \mathbb{E}_{(x,q,y)\sim\mathcal{D}, v\sim g_\theta(x)} \left[-\log \left(p_\phi(y \mid v, q) \right) \right] \tag{2}$$

模型通过最小化交叉熵损失来学习如何根据图像和文本提示生成准确的图像描述。这个过程涉及到从数据集中抽取样本，通过视觉编码器处理图像，然后使用模型预测图像描述，并根据预测结果与真实描述之间的差异来更新模型参数。

自适应变分信息瓶颈

ADAVIB 利用信息瓶颈原理，通过最小化损失函数 $\mathcal{L}_{IB}$ 来平衡压缩和预测：

$$\min \mathcal{L}_{IB} = \beta I(v; z) - I(z; y) \tag{3}$$

其中，$I(v; z)$ 表示输入 v 和压缩表示 z 之间的互信息，$I(z; y)$ 表示压缩表示 z 和目标 y 之间的互信息。

这个方法的基本原理是，希望 LLM 在描述图像时，只关注图像中真正重要的部分，忽略那些不重要的细节。这样，它就不会因为一些无关的细节而产生幻觉。

ADAVIB 通过变分方法近似信息瓶颈，引入 KL 散度来衡量后验分布和先验分布之间的差异：

$$\min \mathcal{L}_{VIB} = \beta \mathbb{E}_v \left[\mathrm{KL} \left(p_\theta(z \mid v) \parallel r(z) \right) \right] + \mathbb{E}_{z\sim p_\theta(z|v)} \left[-\log p_\phi(y \mid z, q) \right] \tag{4}$$

其中，$p_\theta(z \mid v)$ 是给定输入 v 时 z 的后验分布，$r(z)$ 是 z 的先验分布，$p_\phi(y \mid z, q)$ 是给定 z 和文本提示 q 时 y 的条件概率分布。

通过最小化变分信息瓶颈损失函数，实际上是在最小化后验分布和先验分布之间的 Kullback-Leibler（KL）散度。这样做的目的是鼓励模型学习到的表示既能够保留对目标变量有用的信息，又能够尽可能地压缩输入数据，从而减少对无关特征的依赖，提高模型的泛化能力。

ADAVIB 通过计算相似性分布的熵来动态调整噪声强度，以适应每个样本的特性：

$$H = -\sum_{i=1}^{|V|} p(E(i)_{\mathrm{LLM}} \mid z) \log p(E(i)_{\mathrm{LLM}} \mid z) \tag{5}$$

$$\beta \leftarrow -\beta \cdot \log \left(\frac{H}{\log(|V|)} \right) \tag{6}$$

其中，H 是相似性分布的熵，$|V|$ 是词汇表的大小。

自适应噪声控制是通过基于熵的机制实现的，该机制能够量化视觉标记与语言模型的词嵌入之间的相似性分布的平滑度。具体来说，该方法利用相似性分布的熵作为度量，以反映视觉标记在映射到 LLM 的词嵌入空间时的不确定性或过度自信的程度。熵 H 越高，表示相似性分布越平滑，即视觉标记对 LLM 词嵌入的映射越不确定。通过调整 ，对于具有低熵（即过度自信）的样本， 值增加，导致更大的噪声注入，从而减少过度自信；对于具有高熵的样本， 值减小，注入的噪声较少，以保留更多的有用信息。

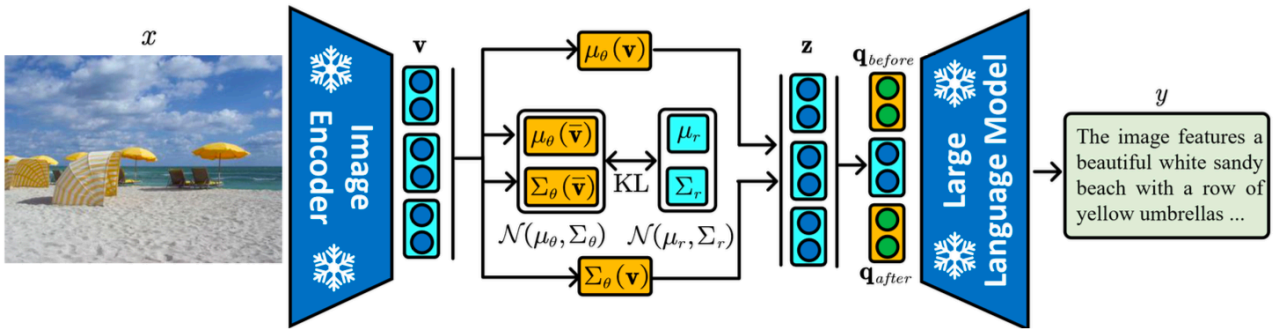

图 1: The model architecture of ADAVIB.

(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 ADAVIB 在缓解物体幻觉方面取得了显著的成果，但本文的工作仍存在一些局限性。例如，ADAVIB 的性能可能受到视觉编码器和语言模型的质量和能力的影响。此外，虽然 ADAVIB 能够有效地缓解对无关特征的过度自信，但在某些情况下，可能需要进一步优化噪声控制策略，以更好地平衡信息压缩和信息保留之间的关系。未来的工作可以探索更复杂的噪声控制机制，或者结合其他正则化技术，以进一步提高模型的性能和鲁棒性。

(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，本文提出的基于 VIB 的噪声注入方法为缓解模型过度自信提供了一种新的视角，这种方法可以应用于其他需要处理模型过度自信问题的领域。其次，本文提出的基于熵的噪声控制策略为自适应地调整模型训练过程中的噪声注入程度提供了一种有效的方法，这种方法可以应用于其他需要动态调整模型训练过程的场景。最后，本文的实验设计和评估方法为评估和比较不同方法在缓解物体幻觉方面的性能提供了有价值的参考。

(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……】

通过本文，我了解到信息瓶颈 IB 和变分信息瓶颈 VIB 是处理模型过度自信和信息压缩的有效工具，这些技术在机器学习和深度学习领域有着广泛的应用。

5 评阅人

姓名:

时间: