【面向多模态大语言模型的幻觉增强对比学习】

——【Hallucination Augmented Contrastive Learning for Multimodal Large Language Model】

1 相关资源

pdf: https://doi.org/10.48550/arXiv.2312.06968
ppt:
短视频:
数据集:
源码: https://github.com/X-PLUG/mPLUG-HalOwl/tree/main/hacl
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: ChaoyaJiang, HaiyangXu, Mengfan Dong, Jiaxing Chen, WeiYe, MingYan, QinghaoYe, JiZhang, Fei Huang, Shikun Zhang
单位:

1. National Engineering ResearchCenter for SoftwareEngineering, Peking University
2. Alibaba Group

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。
本文的工作旨在通过一种新的表示学习方法来减少 MLLMs 中的幻觉现象，从而提高模型在多模态任务中的准确性和可靠性。通过这种方法，本文希望为多模态大语言模型的发展提供一种新的思路，使其能够更好地理解和生成与视觉输入一致的文本信息。

(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，多模态大语言模型（MLLMs）在视觉问答、图像描述生成等任务上取得了显著进展。例如，GPT-4V、LLaVA 等模型通过引入视觉编码器和大语言模型的结合，展示了强大的多模态理解能力。然而，这些模型在生成文本时仍然存在幻觉问题，即生成与视觉输入不一致或完全虚构的信息。现有的方法主要集中在通过限制指令长度或利用人工标注数据进行强化学习来减少幻觉，但这些方法要么牺牲了描述的详细性，要么需要额外的人工标注成本。本文指出，当前 MLLMs 的幻觉问题源于视觉和文本表示之间的语义鸿沟，以及幻觉文本和非幻觉文本表示的纠缠，这使得模型难以区分幻觉和非幻觉文本。

(3) 本文解决的问题
【一句话概括本文解决的核心问题】
通过幻觉增强对比学习缓解多模态大模型的幻觉问题。

(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. 揭示了 MLLMs 中视觉与文本表征间显著的跨模态语义鸿沟，以及含幻觉/非幻觉文本表征意外纠缠的现象，暴露出现有方法在有效桥接视觉与文本表征方面的不足；
2. 提出简单高效的幻觉增强跨模态对比学习（HACL），通过将对比学习引入 MLLMs 并将幻觉文本作为困难负样本，构建跨模态对齐更好、更易区分幻觉的表征空间；
3. 实验证明配备 HACL 的 MLLMs 不仅能减轻幻觉，还能显著提升多项基准评估性能。

(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】
基于上述发现，我们提出幻觉增强跨模态对比学习 HACL，通过增强视觉与文本表征的对齐来缓解幻觉。将含幻觉文本作为图像锚点的困难负样本，自然拉近非幻觉文本与视觉样本的表征距离，同时推远非幻觉文本与幻觉文本的表征。具体而言，我们分别将视觉和文本 token 序列输入 LLMs 以获得各模态的全局表征用于对比学习。使用 GPT-4 生成包含部分对象属性错误或额外虚构信息的幻觉图像描述。如图 1(b) 所示，在 LLaVA[33] 中引入 HACL 迫使视觉表征更接近文本表征，并使正确与幻觉文本表征更易区分。这种有效对齐有助于防止幻觉生成。实验表明，配备 HACL 的 MLLMs 不仅减少幻觉发生，还在多项基准评估中取得提升。

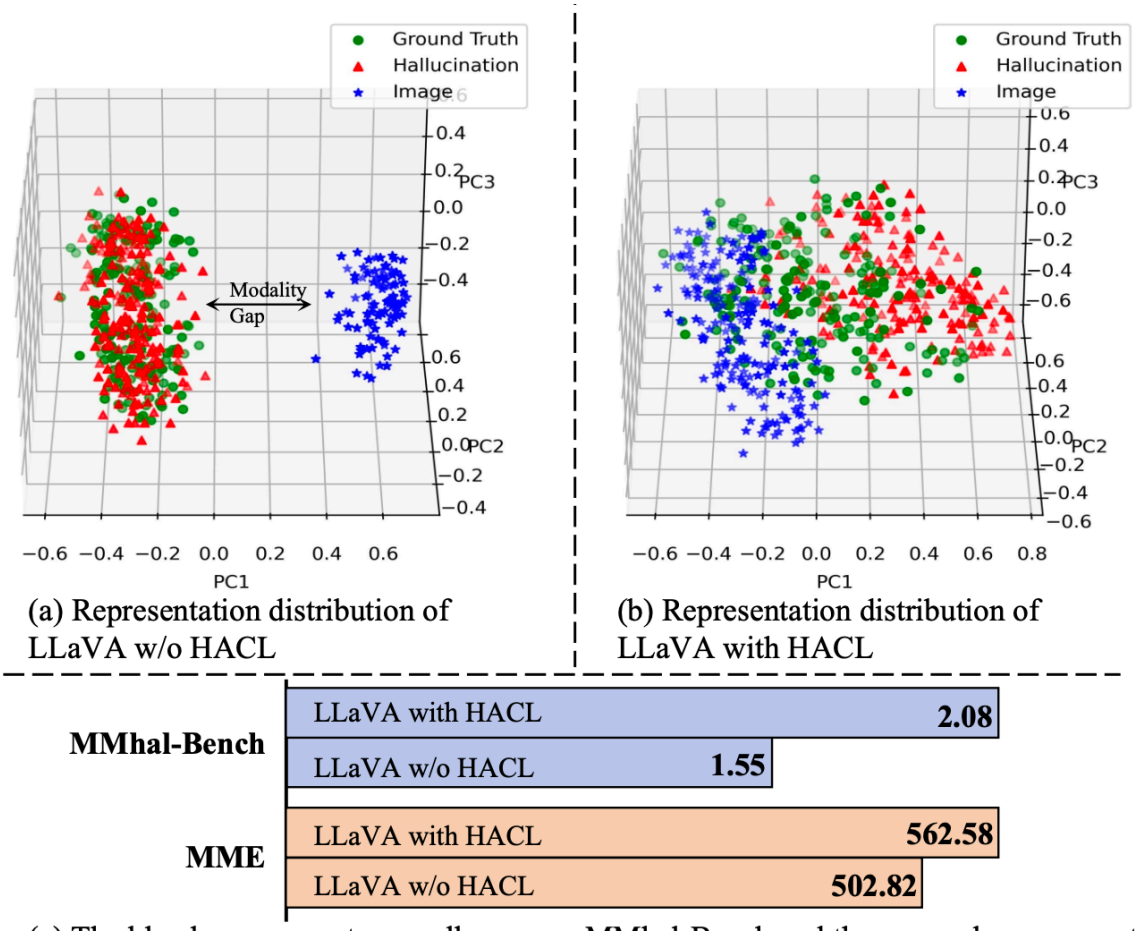

图 1: Subfigure (a) and subfigure (b) show the distributions of the last token's representations yielded by LLM for visual or textualtokensequences.

跨模态对比学习

本方法适用于任何通过可学习接口将视觉信息映射/抽象到文本表征空间的 MLLMs。形式上，假设 MLLM 由视觉编码器 $\mathbf{V}_\theta$、可学习接口 $\mathbf{F}_\alpha$ 和仅解码器架构的大语言模型 $\mathbf{L}_\beta$ 组成，其中 θ, α, β 表示各模块参数。另有无监督预训练数据集包含 N 个图文对，记为 $D = \{I_i, T_i\}, i \in [1, 2, \ldots, N]$

假设图像 I_i 经过视觉编码器 $\mathbf{V}_\theta$ 和可学习接口 $\mathbf{F}_\alpha$ 处理后，被转换为长度为 m 的视觉标记序列。由于大多数 LLM 仅为解码器模型，为获得能捕捉全局语义信息的表征，作者使 $< EOS >$ 标记通过嵌入层 $\mathbf{L}_\beta$ 得到向量表示 $e \in \mathbb{R}^D$，并将其附加至该序列。因此，新的视觉标记序列变为 $S_v^i = [v_1^i, v_2^i, \ldots, v_m^i, e_v^i]$，其中 $v_k^i \in \mathbb{R}^D, k \in [1, 2, \ldots, m]$。

同理，对于与该图像配对的描述文本，我们也在文本标记序列后附加 $< EOS >$ 标记，并通过 LLM 的嵌入层获取文本嵌入序列 $S_t^i = [t_1^i, t_2^i, \ldots, t_n^i, e_t^i]$，其中 $t_k^i \in \mathbb{R}^D, k \in [1, 2, \ldots, n]$。

随后，视觉嵌入序列 S_v 与文本嵌入序列 S_t 分别输入 LLML$\mathbf{L}_\beta$，从 $\mathbf{L}_\beta$ 最后一层获得如下最终输出：

$$H_v^i = \mathbf{L}_\beta(S_v^i) \tag{1}$$

$$H_t^i = \mathbf{L}_\beta(S_t^i) \tag{2}$$

其中 $H_v^i = [\hat{v}_1^i, \hat{v}_2^i, \ldots, \hat{v}_m^i, \hat{e}_v^i]$ 与 $H_t^i = [\hat{t}_1^i, \hat{t}_2^i, \ldots, \hat{t}_n^i, \hat{e}_t^i]$。之后，我们得到捕捉图像 I_i 整体语义信息的全局表征 $\hat{e}_v^i$，以及捕捉真实描述文本 T_i 整体语义信息的全局表征 $\hat{e}_t^i$。随后，类似于视觉语言预训练领域众多现有方法，作者引入以下对比学习策略。假设训练过程中批次大小为 B，文本-图像对

比学习损失计算如下示：

$$\mathcal{L}_{CL}^{t} = -\sum_{i=1:B} \frac{1}{B} \log \left[\frac{f\left(\hat{e}_t^i, \hat{e}_v^i\right)}{f\left(\hat{e}_v^i, \hat{e}_t^i\right) + \sum_{k \neq i} f\left(\hat{e}_t^i, \hat{e}_v^k\right)} \right] \tag{1}$$

其中 $f\left(\hat{e}_t^i, \hat{e}_v^i\right)$ 用于度量 $\hat{e}_t^i$ 与 $\hat{e}_v^i$ 在语义空间中的距离。同理，图像-文本对比学习损失如下：

$$\mathcal{L}_{CL}^{v} = -\sum_{i=1:B} \frac{1}{B} \log \left[\frac{f\left(\hat{e}_v^i, \hat{e}_t^i\right)}{f\left(\hat{e}_v^i, \hat{e}_t^i\right) + \sum_{k \neq i} f\left(\hat{e}_v^i, \hat{e}_t^k\right)} \right] \tag{2}$$

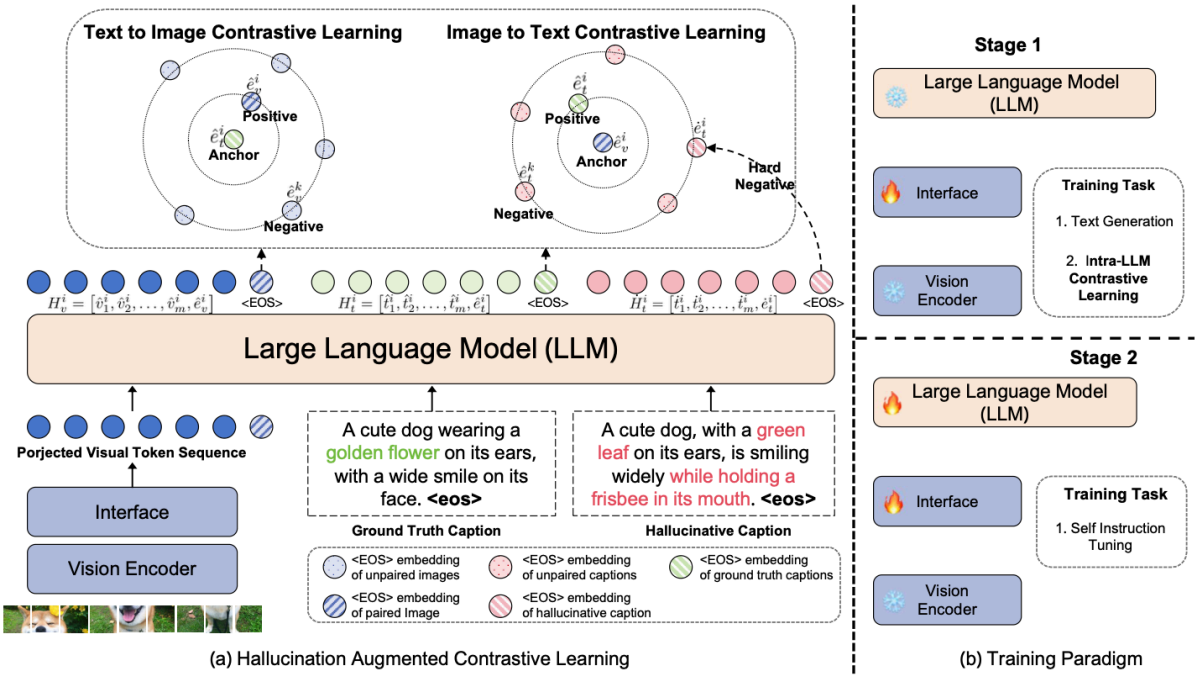

图 2: HACL.

通过幻觉描述提升对比学习效果

本文提出通过引入模拟 MLLM 生成幻觉文本的困难负样本来提升对比学习效果。

作者利用 GPT-4 在真实标注中植入与图像内容不符或完全不存在的元素。这些幻觉既可以是粗粒度的（关注物体存在性），也可以是细粒度的（聚焦数量、属性或位置等具体特征）。以下是输入 GPT-4 的提示模板：Hallucination in Large-scale Visual Language Models (LVLMs) refers to cases where these models generate de- scriptions introducing elements that are inconsistent with the content or completely absent from a provided image. These hallucinations can be coarse-grained, focusing on the mere existence of objects, or fine-grained, focusing on more specific attributes or characteristics such as quantity, properties, and locations. Your task is to revise a given caption-tocreateamirroredversionthatcloselyalignswith the original's content and length but incorporates elements of hallucination. The first step involves identifying the ob- jectsinvolvedandtheirassociate-dattributeswithinthegiven caption. Subsequently, combine thisinsight with the details concerning-hallucinationsprovidedabovetocompleteyour task.

幻觉增强对比学习假设已基于图像 I_i 的原始描述 T_i 生成幻觉描述 $\hat{T}_i$，并获得其全局表征 $\hat{e}_i^i$，可将

其作为图文对比学习中的负样本。因此，图像-文本对比学习的新公式为：

$$\mathcal{L}_{CL}^{v} = -\sum_{i=1:B+1} \frac{1}{B+1} \log \left[\frac{f\left(\hat{e}_v^i, \hat{e}_t^i\right)}{f\left(\hat{e}_v^i, \hat{e}_t^i\right) + f\left(\hat{e}_v^i, \hat{e}_t^i\right) + \sum_{k \neq i} f\left(\hat{e}_v^i, \hat{e}_t^k\right)} \right] \tag{3}$$

训练范式

如图 2(b) 所示，该图展示了如何在多模态大语言模型训练过程中引入 HACL。通常将 HACL 融入模型的第一阶段预训练，以更好地优化接口 $\mathbf{F}_\alpha$。设文本生成任务的损失函数为 $\mathcal{L}_G$，则第一阶段的优化目标可定义为：

$$\mathcal{O}_\alpha = \arg\min_{\alpha} \mathcal{L}_G + \left(\mathcal{L}_{CL}^{v} + \mathcal{L}_{CL}^{t}\right)/2 \tag{4}$$

第二阶段与其他方法保持一致，仅使用指令数据对模型进行微调。

(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?】

本文的局限性在于对比学习的负样本仅依赖于 GPT-4 生成的幻觉文本，这可能限制了负样本的多样性和覆盖范围。此外，本文的训练范式在激活语言模型（LLM）时可能导致性能下降，这表明在训练过程中需要更精细的策略来平衡视觉编码器和语言模型的优化。

(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】

本文的思路和方法在多模态学习领域具有较高的借鉴价值。首先，本文通过分析视觉和文本表示的分布，揭示了幻觉问题的本质，为后续的研究提供了新的视角。其次，对比学习作为一种有效的表示学习方法，在本文中被成功应用于解决幻觉问题，这为其他多模态任务提供了一种新的解决方案。

(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……】

本文展示了对比学习在多模态表示学习中的强大潜力，为解决幻觉问题提供了一种新的思路。此外，本文还强调了在多模态模型的预训练阶段，如何平衡视觉编码器和语言模型的优化是一个重要的研究方向。

5 评阅人

姓名:

时间: