【通过解译注意力因果性缓解多模态大语言模型中模态先验诱导的幻觉】

——【MITIGATING MODALITY PRIOR-INDUCED HALLUCINATIONS IN MULTIMODAL LARGE LANGUAGE MODELS VIA DECIPHERING ATTENTION CAUSALITY】

1 相关资源

pdf: https://arxiv.org/pdf/2410.04780
ppt:
短视频:
数据集:
源码: https://github.com/The-Martyr/CausalMM
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: Multi-modal 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: MLLMs, Hallucination Mitigation, Deciphering Attention Causality
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: Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, Xuming Hu
单位:

1. The Hong Kong University of Science and Technology (Guangzhou)
2. The Hong Kong University of Science and Technology
3. Nanyang Technological University
4. Tsinghua 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?】

多模态大模型在工业和学术界的应用日益广泛，但是这些模型常受到视觉和语言先验（模态先验）的偏差影响，导致多模态幻觉问题。模态先验指的是模型初始参数中存在的固有偏差，这些偏差通过注意力机制影响模型输出，导致模型无法准确对去多模态输入。

本文的工作旨在通过因果推断框架来解决这个问题，通过解码注意力因果关系，减少先验知识对模型输出的负面影响，从而提高多模态大模型的性能。

(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】

当前的研究已经取得了一定的进展，例如在多模态大语言模型（MLLMs）的开发和应用方面，像 VITA 和 Cambrian-1 等模型已经在多个基准测试中展现出强大的性能。

然而，现有研究主要集中在如何利用统计相关性来优化模型输出，而忽视了注意力机制与模型输出之间的因果关系。这些方法虽然能够在一定程度上缓解幻觉问题，但效果有限。存在的问题主要是现有方法无法系统地研究视觉注意力、语言注意力、模态先验和模型输出之间的因果关系，导致模型在推理过程中无法准确理解底层依赖关系，从而加剧了偏差并导致多模态幻觉问题。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文构建了因果推理框架，旨在通过解码注意力因果关系来减轻模态先验对模型输出的负面影响，从而提高多模态任务的性能。

(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. 引入了因果推理框架 CAUSALMM，将模态先验视为注意力机制和模型输出之间的混杂因素。
2. 通过后门调整和反事实推理来量化视觉和语言注意力对模型输出的因果效应。

(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】

结构因果模型

作者通过构建结构因果模型 SCM 来描述各组件间的关系，因果图构建如图 1。

这些公式描述了多模态大语言模型中各个组件之间的因果关系。输入图像 I 会影响视觉注意力 A_i 和视觉标记嵌入 T_i，而视觉先验 P_v 也会对这些部分产生影响。同样，语言标记嵌入 T_t 和语言先验 P_l 会影响模型注意力 A_t 和最终输出 O。

$I \rightarrow A_i$: The image input I influences the visual attention layer A_i.
$I \rightarrow T_i$: The image input I directly affects the visual token embeddings T_i.
$P_v \rightarrow A_i$: Visual priors P_v contribute to the attention in the visual attention module.
$P_v \rightarrow T_i$: Visual priors P_v also influence the formation of visual token embeddings T_i.
$A_i \rightarrow T_i$: Visual attention A_i impacts the encoding of visual tokens.
$T_i \rightarrow O$: Visual tokens T_i contribute directly to the model's output.
$T_t \rightarrow A_t$: Language token embeddings T_t influence the MLLM's attention A_t.
$T_t \rightarrow O$: Language token embeddings T_t directly impact the final output.
$P_l \rightarrow A_t$: Language priors P_l inform the MLLM's attention mechanism A_t.
$P_l \rightarrow O$: Language priors P_l directly affect the model output O.
$A_t \rightarrow O$: LLM attention A_t shapes the final output O.

图 1: The Casusal Graph.

多模态注意力干预

作者对视觉和语言组件的注意力层实施特定干预，以研究其对模型输出的因果效应，这些干预通过修改注意力权重生成反事实输出，从而分离各模态的影响。

作者通过随机化注意力权重来生成反事实状态，将原始注意力图替换为反事实状态。对于反事实状态，可以采用随机注意力权重、均匀分布、反向分数或置换注意力图等形式。这样做的目的是通过改变注意力的分布，来观察这些改变对模型输出的影响。

- 随机注意力：用均匀分布的随机值替代原始注意力分数。
- 均匀注意力：为所有注意力分数分配恒定值。
- 反向注意力：通过最大值减去各注意力分数实现注意力图反转。
- 置换注意力：对视觉编码器的空间位置注意力分数进行随机置换。

作者在对不同注意力做消融实验时发现，以随机注意力作为因果效应锚点能带来最显著的模型性能提升。在跨层干预的消融实验中发现，浅层与中层的干预效果最为显著，由此推测语言先验显著影响处理的初始阶段。

反事实注意力状态具体设定如图 2。

在纯视觉反事实推理中，仅干预视觉注意力（即视觉编码器的注意力）；在纯语言反事实推理中，仅干预大语言模型的多头自注意力；在多模态协同反事实推理中，同步干预视觉与语言注意力，并获取其协同因果效应的总和。

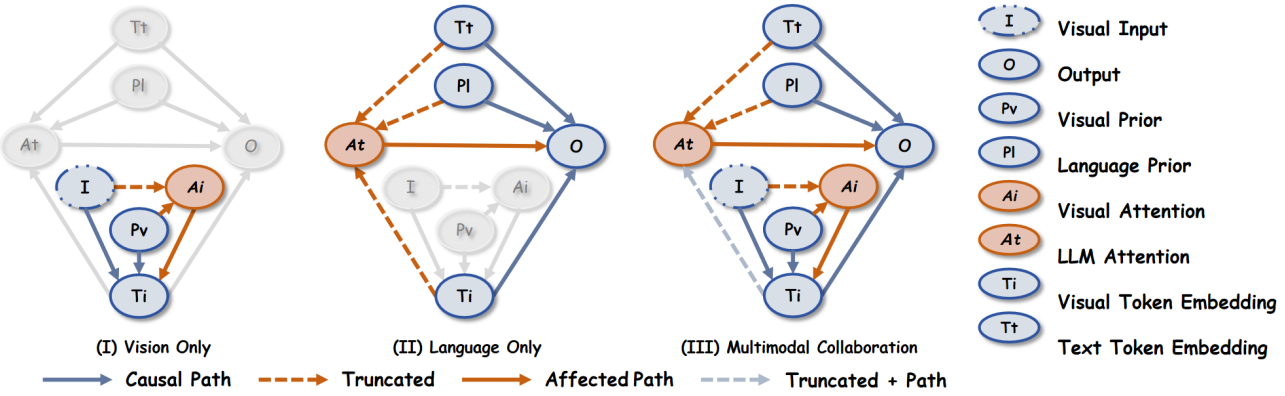

图 2: Causal diagram of counterfactual reasoning.

反事实推理

为量化反事实干预对模型输出的影响，作者基于后门调整准则进行反事实推理。通过后门调整框架，作者能在模态先验混杂因素的影响下有效获取其他变量的因果效应。

视觉注意力的因果效应：

$$P^{V}_{\text{effect}} = E_{A_i \sim \tilde{A}_i}\left[P(O|A_i = A_i, I = I, P_v = P_v) - P(O|\text{do}(A_i = a_i), I = I, P_v = P_v)\right] \tag{1}$$

语言注意力的因果效应：

$$P^{L}_{\text{effect}} = E_{A_t \sim \tilde{A}_t}\left[P(O|A_t = A_t, T_t = T_t, P_l = P_l) - P(O|\text{do}(A_t = a_t), T_t = T_t, P_l = P_l)\right] \tag{2}$$

多模态协同的因果效应：

$$\begin{aligned} P^{M}_{\text{effect}} = E_{A_i, A_t \sim \tilde{A}_i, \tilde{A}_t}\big[&P(O|A_i = A_i, A_t = A_t, I = I, T_t = T_t, P_v = P_v, P_l = P_l) \\ &- P(O|\text{do}(A_i = a_i), \text{do}(A_t = a_t), I = I, T_t = T_t, P_v = P_v, P_l = P_l)\big] \end{aligned} \tag{3}$$

这些公式描述了如何通过反事实推理来估计注意力机制对模型输出的因果效应。具体来说，作者比较了在正常情况下（没有干预）和在干预情况下（改变了注意力分布）模型输出的概率分布。通过这种比较，可以量化注意力机制对模型输出的真实影响，从而调整模型以减少幻觉问题。

(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？】

本文的局限性在于其方法依赖于现有的视觉编码器和语言模型的性能，如果视觉编码器或者语言模型本身存在性能瓶颈的话，CAUSALMM 的效果也会受到限制。

(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】

本文的因果推断框架和反事实推理方法为多模态大语言模型的研究提供了新的思路。特别是如何通过因果关系来优化模型输出，这种方法可以被应用于其他类型的多模态任务，以解决类似的偏差问题。

(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……】

本文展示了因果推断技术在人工智能领域的应用潜力，特别是在解决模型偏差和幻觉问题方面。这让我认识到因果关系在模型优化中的重要性，也为我提供了新的研究方向和技术思路。

5 评阅人

姓名:
时间: