【RUR: 通过推理不确定性引导的优化减轻大型视觉-语言模型中的幻觉问题】

——【Mitigating Hallucinations in Large Vision-Language Models via Reasoning Uncertainty-Guided Refinement】

1 相关资源

pdf: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=11125489
ppt:
短视频:
数据集:
源码: https://github.com/Mrshenshen/RUR?tab=readme-ov-file
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: IEEE Transactions on Multimedia 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: Shenshen Li, Xing Xu, Wenxin Meng, Jingkuan Song, Chong Peng, Heng Tao Shen
单位:

1. Computer Science and Technology, Tongji University, Shanghai, 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 (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。
因此，本文的工作旨在通过一种新的框架来缓解 LVLMs 中的幻觉问题，提升模型生成内容的准确性和与输入数据的一致性，从而增强其在复杂多模态任务中的实用性。

(2) 研究现状
【当前的最好研究做到什么程度了】
当前，针对 LVLMs 中的幻觉问题，已有研究主要通过引入特定设计的数据集或采用对比解码技术来缓解。例如，一些方法通过对比学习的方式，利用负样本引导模型生成更准确的内容。然而，这些方法存在一些局限性。一方面，它们高度依赖于数据集的质量和负样本的构建，一旦数据质量不佳或负样本设计不合理，模型的性能就会大打折扣。另一方面，这些方法忽略了推理过程中固有的不确定性，即由于语言先验和数据复杂性导致的推理模糊性。这种不确定性使得模型难以准确识别生成内容中每个词或句子背后的因果关系，从而增加了幻觉的风险。从论文的描述来看，现有的方法虽然在一定程度上缓解了幻觉问题，但仍然无法从根本上解决由于推理不确定性导致的幻觉现象，这为本文的研究提供了切入点。

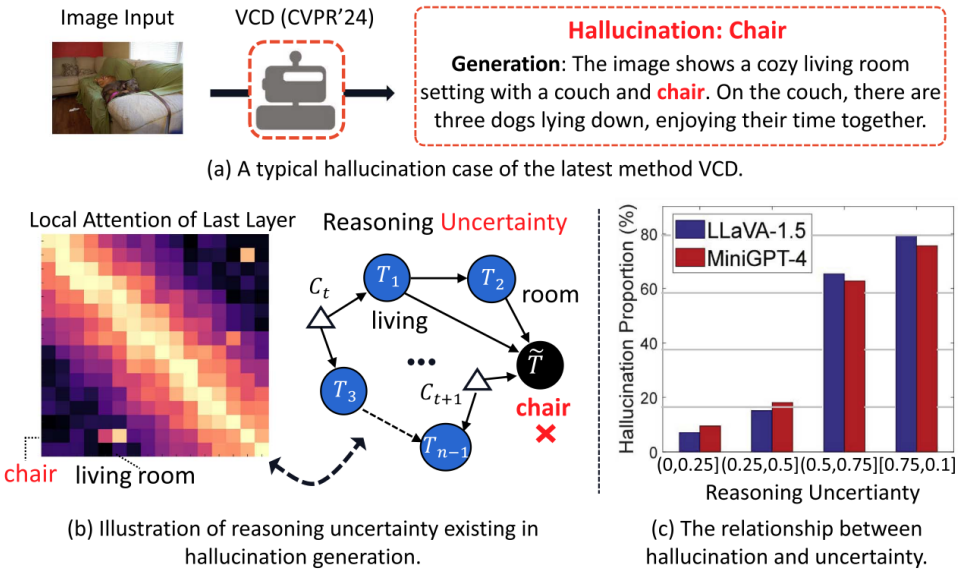
(3) 本文解决的问题
【一句话概括本文解决的核心问题】
本文解决的核心问题是通过建模推理不确定性来缓解大视觉 - 语言模型中的幻觉现象，从而提高模型生成内容的准确性和可靠性。

(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. 提出了一种推理相关性发现机制，该机制利用结构因果模型与大型视觉语言模型中 Transformer 之间的联系，来揭示标记之间的因果推理关系。这有助于模型理解和评估大型视觉语言模型中的复杂推理过程。
2. 开发了一个推理不确定性建模模块，以明确地对局部和全局层面的推理不确定性进行建模。它能够熟练地引导模型识别可能过度自信的生成内容。
3. 提出了一种基于多级不确定性的调整方法，以减轻层级幻觉，该方法能有效消除不可靠的标记和句子。

(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】
发现
基于对比解码的方法过于关注所构建的噪声样本的质量。这些方法通过在解码过程中利用噪声图像来放大幻觉，生成对比概率分布。虽然在某些场景下有效，但当噪声校准不理想时，其固定的噪声增强策略会带来挑战，可能导致幻觉的产生。如图 1(a) 所示，最新的 VCD 模型存在幻觉问题，例如会生成原始图像中不存在的椅子，尤其是在噪声条件不理想的情况下。作者推测，这一问题可能源于忽视了与语言先验偏差和输入复杂性相关的固有推理不确定性。具体而言，图 1(b) 显示，由于存在“客厅”和“椅子”之间存在强相关性的语言先验偏差，VCD 在推理过程中会误解标记之间的上下文关系。

此外，为了验证这一假设，作者在图 1(c) 中探究了幻觉比例与推理不确定性之间的正相关性，其中推理的不确定性对应着更高的幻觉频率。这些发现表明：（1）忽视推理不确定性会使模型无法准确识别标记之间的因果关系，从而增加幻觉出现的可能性。（2）推理不确定性可以表明模型输出生成的不可靠程度。



(a) A typical hallucination case of the latest method VCD.

(b) Illustration of reasoning uncertainty existing in hallucination generation.

(c) The relationship between hallucination and uncertainty.

图 1: Illustrative examples of existing problems.

初步研究

问题表述：模型的输入既包含图像也包含文本。RUR 模型采用视觉编码器从输入图像中提取视觉令牌，随后将这些令牌转换到大型语言模型的输入空间中。这些视觉令牌用 $T_v = t_0, ..., t_{m-1}$ 表示，其中 m 代表视觉令牌的数量，它们与经过令牌化的文本数据一同作为 LLM 输入的一部分。文本输入通过文本编码器进行令牌化，用 $T_t = t_m, ..., t_n$ 表示。视觉令牌和文本令牌被拼接起来，形成模型的综合输入序列，用 $T = t_i{}_{i=0}^{n}$ 表示。

结构因果模型：SCM 旨在捕捉某个领域中的因果机制。一个 SCM 由 $U, X, F, P(U)$ 组成，其中 $U = U_1, ..., U_m$ 是潜在的外生随机变量，$X = X_1, ..., X_n$ 是内生随机变量，$F = f_1, ..., f_n$ 是一组确定性函数，用于根据其直接原因定义 x，而 $P(U)$ 是关于 U 的分布。每个内生变量 X_i 都有一个唯一的外生原因 $U_i(m = n)$，其中 X_i 的值由下式给出：

$$X_i \leftarrow f_i(Pa_i, U_i) \tag{1}$$

其中 Pa_i 表示 X_i 的直接原因集。令 G 为权重矩阵，其中 $G(i, j)$ 是决定子节点 X_i 的父节点 x_j 的权重，Λ 为对角矩阵，其中 $\Lambda(i, i)$ 是外生节点 U_i 的权重。那么，矩阵形式 x 可以定义为：

$$X = GX + \Lambda U \tag{2}$$

推理相关性发现

自注意力与结构因果模型之间的联系：作者首先将自注意力描述为一种在未知结构因果模型中编码符号间相关性的机制。具体而言，我们表明自注意力输出的协方差类似于结构因果模型（SCM）中观测节点的协方差。在结构因果模型中，内生节点的值以矩阵形式定义为式（2）。X 可表示为：

$$X = (I - G)^{-1}\Lambda U \tag{3}$$

因此，输出的协方差矩阵 C_X 为：

$$\begin{aligned} C_X &= E\left[(X-\mu_X)(X-\mu_X)^\top\right] \\ &= E\left[(I-G)^{-1}\Lambda(U-\mu_U)(U-\mu_U)^\top\left((I-G)^{-1}\Lambda\right)^\top\right] \\ &= (I-G)^{-1}\Lambda E\left[(U-\mu_U)(U-\mu_U)^\top\right]\left((I-G)^{-1}\Lambda\right)^\top \\ &= \left((I-G)^{-1}\Lambda\right)C_U\left((I-G)^{-1}\Lambda\right)^\top, \end{aligned} \tag{4}$$

其中，$\mu_X=(I-G)^{-1}\Lambda\mu_U$，且 C_U 是外生变量 U 的协方差矩阵。Transformer 中的注意力层从输入序列的嵌入 Y 中估计出一个注意力矩阵 A 和一个值矩阵 V。输出嵌入通过 $Z=AV$ 计算得出，其中第 j 个投影为 $Z(\cdot,j)=AV(\cdot,j)$。Z_j 的协方差 C_{Z_j} 公式如下：

$$\begin{aligned} C_{Z_j} &= E\left[\left(Z_j-\mu_{Z_j}\right)\left(Z_j-\mu_{Z_j}\right)^\top\right] \\ &= AE\left[(V_j-\mu_j)\left(V_j-\mu_{V_j}\right)^\top\right]A^\top = AC_{V_j}A^\top, \quad (5) \end{aligned} \tag{5}$$

其中，C_{V_j} 是输入嵌入的协方差矩阵。根据中心极限定理，可得 $C_V\to I$，进而得出 $C_Z=AA^\top$。将注意力矩阵 A 和值向量 V 解释为结构因果模型（SCM）的组成部分，使得其内生变量的协方差矩阵等于注意力层输出的协方差矩阵，即 $C_X=C_Z$：其中，注意力矩阵 A 度量了对于输入嵌入的任何投影 $V(\cdot,j)$，每个变量对另一个变量的总效应。类似地，对于任何外生值，结构因果模型中的矩阵 $(I-G)^{-1}\Lambda$ 度量了一个变量通过因果图中所有有向路径对另一个变量产生的效应变量 U。因此，矩阵 A 类似于 $(I-G)^{-1}\Lambda$，而 $V(\cdot,j)$ 类似于 U 的赋值，后者充当上下文。

基于注意力的推理关联表示：基于上述观察，进一步将这种类似关系扩展到大型视觉语言模型（LVLM）中的 Transformer 架构，其目的是通过注意力机制来建模 LVLM 中每个标记之间的推理关系。在 LVLM 中，Transformer 解码器由连续的自注意力层组成。每一层在将输出嵌入传递到下一层之前，都会对其进行独立的非线性变换。各层包含多个自注意力头，这些注意力头使用特定的注意力矩阵并行处理输入，输出则在不同头之间进行线性组合。重要的是，嵌入是单独处理的，仅在矩阵乘法过程中才会受到其他嵌入的影响。因此，多层多头部架构可以被概念化为一种深度图形模型，其中前一个注意力层估计结构因果模型（SCM）的外生节点。因果图从最后的（最深的）注意力层中恢复，而较早的层则被视为上下文估计，编码在外生节点的值中。在建立了 LVLM 与 SCM 之间的关系后，从（6）中我们可以基于来自 LVLM 中 Transformer 最深层的注意力 A 来表示成对推理相关性 $\rho_{i,j}$，其可表述为：

$$\rho_{i,j}=C_Z(i,j)/\sqrt{C_Z(i,i)C_Z(j,j)} \tag{6}$$

这种关联 $\rho_{i,j}$ 能够反映出大语言视觉模型（LVLM）推理过程中标记背后的潜在关系。

推理不确定性建模

局部推理不确定性建模：为有效减轻因过度依赖语言先验和数据复杂性而产生的推理不确定性的影响，对推理不确定性进行建模的需求显而易见。根据主观逻辑（SL）[41]，首先利用因果推理关系（在（7）中提及）来获得推理证据 e_{ij} 的预测：其中 g 表示 ReLU 函数。证据 e_{ij} 代表对标记背后所学推理关系的置信度度量。基于此，可以量化局部推理不确定性 $u_{ij}^l=1-e_{ij}$，它与相应的证据成反比，反映生成标记的不可靠程度。全局级推理不确定性建模：根据上一节中提到的推理证据 e_{ij}，得到 α_i，并对全局级推理不确定性 u^g 建模如下：

$$\alpha_i=e_i+1, u^g=\frac{N}{S} \tag{7}$$

其中 $S=\sum_{i=1}^N\alpha_i$ 可被视为狄利克雷分布的强度。其特征在于　可定义为：

$$D(p|\alpha)=\begin{cases}\frac{1}{B(\alpha)}\prod_{j=1}^N p_j^{\alpha_j-1} & forp\in\mathcal{S}_N,\\ 0 & otherwise,\end{cases} \tag{8}$$

其中 $B(\alpha)$ 表示 N 维贝塔函数，而 S_N 是 N 维单位单纯形。全局级不确定性 u^g 可被视为单句不可靠性的度量。

基于多级不确定性的调整

基于局部不确定性的调整：与现有的大型视觉语言模型（LVLMs）类似，首先使用一个词汇头（记为 f_h）对隐藏状态向量 h_n 进行投影。这会生成用于预测后续 tokens 的常规对数概率（或概率）$p(t_n|x_{<n})$，其公式为：

$$p\left(t_n|x_{<n}\right) = SoftMax\left[f_h\left(h_n\right)\right]_{t_n}, t_n \in T \tag{9}$$

其中，h_t 表示来自 LVLMs 最后一层的输出隐藏状态。符号 $t_{<n}$ 简化了前面的序列 $t_{i\,i=0}^{n-1}$，T 表示完整的词汇集。随后，引入建模的局部推理不确定性 u^l 来优化预测概率，其表达式为：

$$p(t_n|t_{<n}) = \alpha p(t_n|t_{<n}) + (1-\alpha)\left(1 - \beta\frac{u^g}{\|u^g\|}\right) \tag{10}$$

其中，　和　为调和因子，$\frac{1}{\|u^g\|}$ 表示全局级不确定性的归一化。这种方法能够将不确定性细致地融入预测过程，提高 token 生成的稳健性和准确性。基于全局不确定性的调整：如前所述，显然较低的 u^g 与幻觉可能性增加显著相关，而接近零的值则表明所描述内容的真实性更高。因此，u^9 充当初步过滤器，利用严格定义的幻觉阈值　来识别句子中潜在的不准确之处。

作者引入了一种自检机制来增强这种初始过滤的稳健性。当全局层面的不确定性超过阈值　时，该机制便会启动，以评估响应中对象层面和属性层面的语义与模型生成的语义是否一致。这包括提示模型去验证每个所讨论的对象在大视觉语言模型的响应中是否真实存在。对于响应中提到的每个对象，会采用一组双重问题序列：第一个问题针对该对象的属性，紧接着的另一个问题则探究哪个对象拥有这些属性。模型严格评估响应与所描述属性的一致性。如果检测到不一致，这些对象会被视为幻觉内容，随后从输出中排除，以最大限度地减少错误信息的传播。

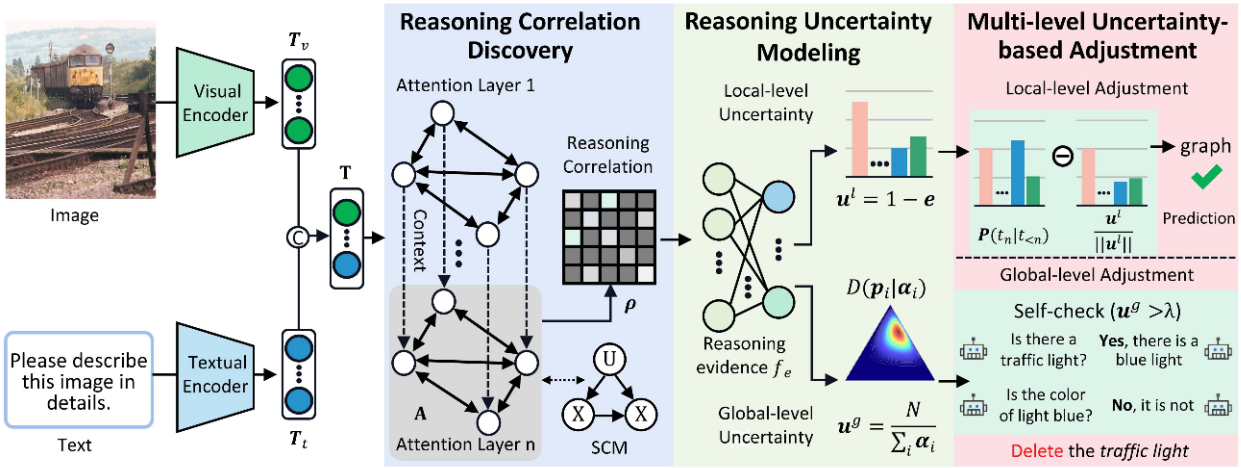


图 2: The overall framework of our proposed RUR method. It consists of three key components: 1) Reasoning Correlation Discovery (RCD); 2) Reasoning Uncertainty Modeling (RUM); and 3) Multi-level Uncertainty-based Adjustment (MUA).

第一步（找关联）：利用模型的注意力机制，找出每个词（标记）之间的因果关系（比如“火车”和“轨道”强相关，“客厅”和“火箭”弱相关）；

把 Transformer 早期注意力层的输出，当作 SCM 的“背景因素”（外生变量 U），负责提供上下文

信息；把 Transformer 最深层的注意力层输出，当作 SCM 的“核心变量”（内生变量 X），核心关注这些变量间的因果影响。计算最深层输出嵌入 Z 的协方差矩阵 C_Z，用公式计算标记 i 和标记 j 的相关性。如，若“火车”和“轨道”的 $C_Z(i,j)$ 值大、方差小， 值会接近 1，说明两者因果关联强；若“客厅”和“火箭”的 $C_Z(i,j)$ 值小， 值接近 0，说明关联弱。
第二步（判可靠）：根据这些关联的强弱，给每个词打“不可靠分”（局部不确定性），给每句话打“不可靠分”（全局不确定性）——关联越弱、越没依据，分数越高；
第三步（做修正）：对“不可靠分”高的词，降低其被生成的概率；对“不可靠分”高的句子，通过反问验证的方式，判断是否是幻觉并剔除。
局部不确定性 u_{ij}^{l} 高的标记，原始概率会被 α 加权压低，同时结合归一化后的全局不确定性 u^{g} 进一步调整。标记越不可靠，修正后的生成概率越低，最终被模型“放弃生成”。
句子不确定性：第一步筛选：设定阈值 $\lambda = 0.5$，若句子的全局不确定性 u^{g} ，直接纳入“可疑句子列表”；第二步自检：对可疑句子中的每个对象，自动生成两组验证问题：第一组问属性：“该对象的颜色 / 位置 / 类别是什么?”；第二组问归属：“具备该属性的对象是什么?”；第三步判定：若模型对两组问题的回答与原句子描述矛盾（如原句说“客厅有红色椅子”，但验证时回答“没有红色家具”），则判定为幻觉句子，直接剔除该句子或删除其中的幻觉对象描述。

(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?】
从论文的描述来看，该方法在计算效率方面可能存在问题，因为引入因果推理关系提取和不确定性建模可能会增加模型的计算复杂度，从而影响其在实际应用中的实时性。此外，虽然本文通过主观逻辑建模了推理不确定性，但这种建模方法可能存在一定的主观性，其是否能够完全准确地反映模型的推理不确定性还有待进一步研究。

(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，从研究问题的视角来看，本文将推理不确定性作为缓解 LVLMs 中幻觉问题的关键因素，这种从不确定性角度出发的思维方式为研究其他类型的生成模型中的问题提供了新的思路。其次，在技术方法上，将结构因果模型与 Transformer 架构相结合来提取因果推理关系的方法具有很强的创新性，为建模复杂的推理过程提供了一种有效的技术手段。

(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……】
本文涉及的结构因果模型和主观逻辑等技术在处理不确定性方面具有独特的优势，这些技术之间的结合为解决复杂的推理问题提供了新的可能性。

5 评阅人

姓名:
时间: