【通过对比层解码提升大语言模型的事实准确性】

——【DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models】

1 相关资源

pdf: https://arxiv.org/pdf/2309.03883
ppt:
短视频:
数据集:
源码: https://github.com/voidism/DoLa
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: ICLR 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: Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, Pengcheng He
单位:

1. Massachusetts Institute of Technology
2. Microsoft

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?】
随着大模型 LLMs 在自然语言处理 NLP 领域的广泛应用，其在众多任务中展现出强大的性能，但

同时也存在着一个严重的问题，即容易出现“幻觉”，也就是生成了与预训练时观察到的真实世界事实相偏离的内容，这在一些对文本可靠性要求极高的场景中极大地限制了它们的应用。
因此，本文提出了一种简单的解码策略，且无需依赖检索外部知识和额外的微调，就可以减少大模型生成的幻觉，使其生成更符合事实的文本。

(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】
目前，尽管 LLMs 的性能不断提升，但其幻觉问题仍未完全解决。一些研究认为，由于大模型采用的最大似然语言建模目标可能会导致模型对不符合训练数据知识的句子也赋予非零概率，从而产生幻觉。
从模型可解释性的角度来看，研究表明 Transformer 语言模型的前置层倾向于编码“低级”信息，而后期层则包含更多“语义”信息，且事实知识在大模型中通常被发现在特定的 Transformer 层中。
现有的减少幻觉的方法包括强化学习人类反馈、推理时的自我一致性检查、多智能体辩论以及使用人类标签进行推理时干预等，但这些方法要么需要额外的训练或数据，要么效率比较低。

(3) 本文解决的问题
【一句话概括本文解决的核心问题】
本文提出了一个解码方法 DoLa，旨在不依赖外部知识检索和额外微调的情况下，通过对比层解码来减少模型的幻觉现象，从而提高其生成内容的真实性。

(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. 提出了新的解码方法 DoLa，该方法不依赖于检索外部知识或对模型进行额外的微调，而是通过对比模型内部不同层的输出来增强事实知识，减少幻觉。
2. 利用 LLMs 中事实知识在特定层中分布的特点，通过计算后期层与早期层输出概率的差异来优化生成结果。。
3. DoLa 方法在解码过程中仅引入了较小的延迟，具有较高的效率。

(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】
论文提出了通过对比层解码 (DoLa) 的方法来减少大模型的幻觉现象。在解码过程中，DoLa 动态地选择一个未成熟层 (前置层)，计算最终成熟层与前置层之间的输出概率差异。具体来说，DoLa 会根据前置层与最终层之间的 JS 散度来动态选择最合适的前置层，以确保所选层与最终层的输出差异最大；然后 DoLa 通过从最终层的输出概率中减去前置层的输出概率，来增强最终层的输出，抑制前置层的输出，使其更倾向于生成事实正确的词汇；此外，DoLa 还引入了一些辅助策略，如自适应可接受性约束和重复惩罚，以进一步提高生成质量和避免重复生成的问题。这种方法充分利用了大模型内部不同层的知识差异，通过内部对比来提升模型的事实性，而无需依赖外部知识或额外的微调。
动态前置层选择
DoLa 首先要动态选择一个前置层 M，并与最终层 N 进行对比。前置层的选择基于其与最终层之间的 JS 散度，以确保所选层与最终层的输出差异最大，从而更有可能包含额外的事实知识。具体来

说，前置层 M 的选择如下：

$$M = \arg\max_{j \in J} \mathrm{JSD}(q_j(x_t|x_{<t}), q_N(x_t|x_{<t})) \tag{1}$$

其中，J 为候选前置层的集合，q 为输出概率分布。**对比预测**

在选择了前置层 M 后，DoLa 通过从最终层的输出概率中减去前置层 M 的输出概率（在对数域中进行），来增强最终层的输出，同时抑制前置层的输出。具体来说，DoLa 计算下一个词的预测概率如下：

$$p_{\mathrm{DoLa}}(x_t|x_{<t}) = \mathrm{softmax}\left(\log q_N(x_t|x_{<t}) - \log q_M(x_t|x_{<t})\right) \tag{2}$$

自适应可接受性约束 APC

为减少对比过程中的误判，DoLa 引入了自适应可接受性约束。即如果某个词在最终层的预测概率过低，则不太可能是合理预测，因此将其概率设为零以最小化假阳性与假阴性情况。在 DoLa 语境中，假阳性指因不同层间低概率区间不稳定，导致低分不可信词元经过对比后可能获得告分；假阴性指当模型对简单决策非常确信时，高分词元在各层的输出概率变化不大，对比后得分反而降低，这时需要强制模型仍从这些高分次元中选择。

具体来说，DoLa 定义了一个阈值 ，如果最终层中某个词的概率低于 ，则将其概率设为零：

$$p_{\mathrm{DoLa}}(x_t|x_{<t}) = \begin{cases} p_{\mathrm{DoLa}}(x_t|x_{<t}), & \text{if } q_N(x_t|x_{<t}) \geq \alpha \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

重复惩罚

在对比过程中，DoLa 可能会导致模型倾向于重复之前生成的句子，尤其是在生成长序列时。为了减少这种重复现象，DoLa 引入了重复惩罚机制。具体来说，DoLa 在解码过程中对重复的词或短语施加一个惩罚因子 ，以降低其生成概率：

$$p_{\mathrm{DoLa}}(x_t|x_{<t}) = p_{\mathrm{DoLa}}(x_t|x_{<t}) \times \left(\frac{1}{\theta}\right)^{\mathrm{repeat}(x_t)} \tag{4}$$

其中，repeat(xt) 是词 xt 在之前生成的序列中出现的次数。

(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?】

该方法主要关注于提升模型的事实性，而对于模型生成内容的其他维度，如连贯性、多样性等的影响没有深入探索。

其次，没有利用外部知识库和人类标注数据进行微调，可能在最新事实或特定领域的任务中受到限制。

(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】

本文通过观察模型内部不同层的输出，以此提出了这样的一个新的解码方法，即通过对比模型内部不同层的输出来增强事实知识，这种思想值得学习。

此外，DoLa 在实验中对多种任务和模型规模进行了广泛的验证，这种全面的实验设计方法也为其他研究提供了一个很好的参考。

(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……】

我了解到解码策略的优化可以发挥很重要的作用，让我认识到除了模型结构和训练方法外，解码过程的设计也同样关键。

5 评阅人

姓名:
时间: