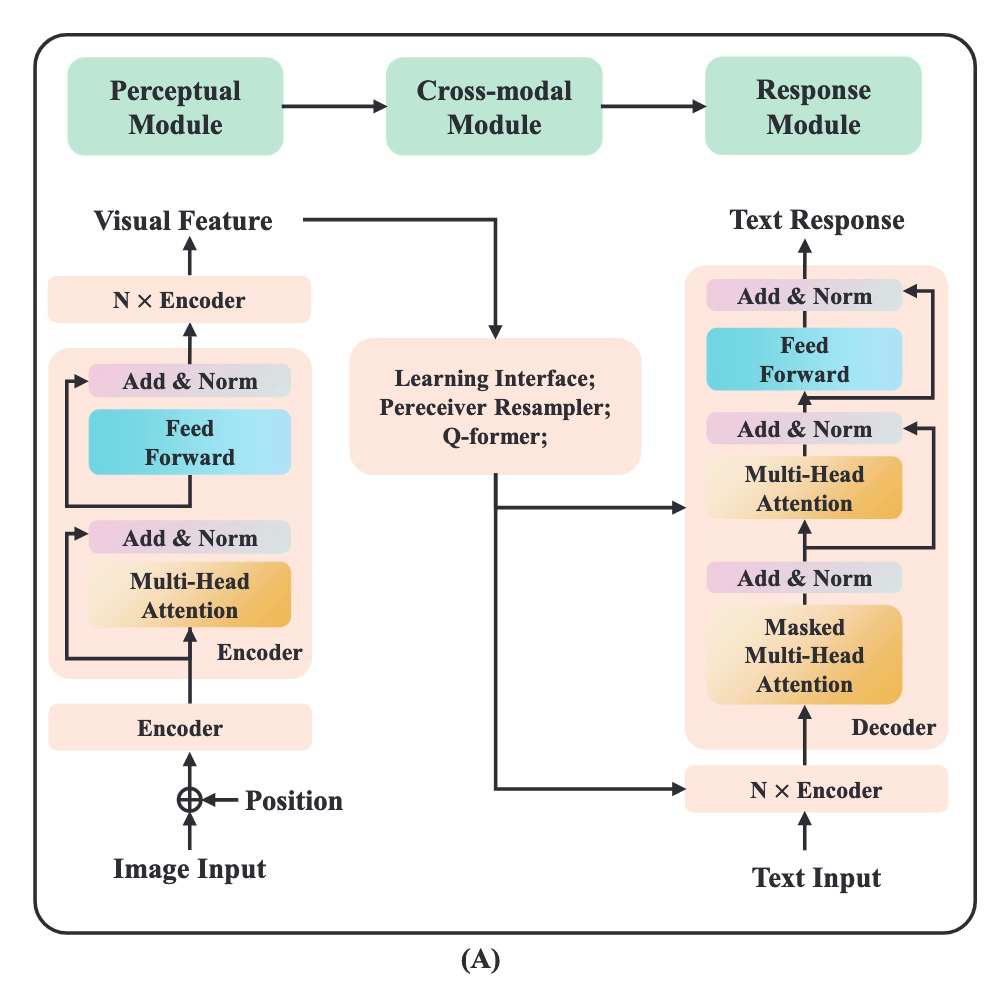
大型视觉语言模型中幻觉现象的综述

2024-10-25 17:41:15

## LVLM 的背景

LVLMs可以分为三个模块：感知模块、跨模态模块和响应模块，通过这三个模块，视觉信息被提取并映射到文本空间，进一步地，视觉信息和文本信息结合以生成最终的响应。



1. 感知模块：通过使用视觉 Transformer 或其变体将图像转换为高维向量。
2. 跨模态模块：弥合视觉和语言之间的模态差异。
3. 响应模块：使用强大的能力来处理和分析视觉和文本的输入，以生成最终答案。

## 幻觉的成因

1. 模态差异：不同模态的特征和表达方式不同
2. 数据集中的有害信息：数据集本来就带有幻觉的数据
3. LLM 的幻觉：LLMS本身很容易产生幻觉

## 幻觉的纠正

### 数据集去幻觉

数据集去幻觉是通过改进或清理训练数据集，以减少或避免模型在推理过程中生成幻觉内容的一种有效方法。

#### 1.数据重写

数据重写是指对训练数据中的错误信息或潜在误导性样本进行修改，以确保模型学到的信息更加准确。

#### 2.去除过度自信

为了解决这一问题，研究者提出了一些方法来降低模型在面对不确定输入时的自信水平。通过调整损失函数，使模型在训练过程中对不确定的样本保持一定程度的怀疑，从而避免模型生成看似合理但实际上错误的响应。

#### 3.打破共现现象

共现现象是指在训练数据中，某些视觉元素和文本描述频繁一起出现，导致模型在推理时不加区分地将这些元素联系在一起。

### 模态差距补偿

由于视觉信息和文本信息之间的特性差异，模型在融合这些多模态数据时，可能会产生错误的推断。通过有效的跨模态模块设计，可以弥合视觉和语言之间的差距，减少因模态不匹配导致的幻觉现象。

#### 1.可学习接口的应用

可学习接口是一种基于投影矩阵的方法，它将视觉信息映射到文本空间。这种方法通过学习视觉和语言之间的对应关系，使模型能够更好地理解视觉信息并将其与文本结合。通过这种方式，模型能够在视觉和语言之间建立更加紧密的联系，从而减少幻觉的产生。

#### 2.Q-former 的引入

Q-former是一种通过交互方式将视觉信息与文本信息连接起来的技术。它通过设计一种特殊的查询机制，使视觉信息在跨模态过程中与文本信息进行交互。这种方法能够更好地处理模态之间的差异，尤其是在复杂的视觉场景中，帮助模型准确地生成文本描述。

#### 3.pereceiver resampler的使用

pereceiver resampler是一种使用交叉注意力机制的技术，旨在将视觉特征编码到文本中。通过这种方法，视觉特征被转换为与文本相匹配的表征，从而减少模态差距导致的幻觉现象。

### 输出纠正

除了通过改进数据集和跨模态机制来减少幻觉，研究人员还致力于直接纠正模型输出的幻觉内容。

#### 1.后处理机制

后处理机制是一种在模型生成响应之后对其进行检查和修正的技术。例如，通过引入额外的验证模块，检查生成的文本是否与视觉信息一致。如果发现生成内容存在逻辑或事实错误，可以通过该模块进行修正。这样的后处理机制可以有效减少模型生成幻觉的可能性。

#### 2.解码策略的优化

解码策略对生成式模型的输出有很大的影响。研究发现，某些解码策略（如贪婪搜索或随机采样）容易引发幻觉现象。为了解决这一问题，研究人员提出了一些新的解码策略，例如基于约束的采样方法或对抗性解码策略。这些方法通过更好地平衡生成的多样性和准确性，减少了模型生成幻觉的概率。

## 幻觉的评估基准

#### 1.判断性基准

判断性基准主要通过分析模型生成的响应是否符合事实或是否与视觉输入一致来评估幻觉。例如，通过引入人工评估或使用预定义的规则来判断模型生成的内容是否准确。这类基准侧重于定量地评估幻觉现象，帮助研究人员了解幻觉在不同场景中的发生频率。

#### 2.生成性基准

生成性基准则侧重于评估模型在生成过程中所表现出的创造性和合理性。尽管模型可能生成出具有一定创新性的内容，但这些内容不一定符合事实或逻辑。因此，生成性基准的评估主要关注模型是否能够在保持生成多样性的同时，减少幻觉的产生。

## 未来研究方向

1. 多模态融合的优化
2. 数据集构建的改进
3. 模型透明性的提高
4. 动态学习机制的引入