暂定题目：基于思维链扰乱和图文对抗的多模态语义对抗攻击方法研究

- 多模态对抗攻击的必要性

随着人工智能技术的进步，多模态大模型在处理文本、图像、音频等多种数据时展现出强大能力，广泛应用于自然语言处理、医疗诊断、自动驾驶等领域。它们提升了机器翻译、情感分析和内容生成的效果，使人机交互更加自然。

多模态对抗攻击是指在多模态任务中，通过对输入数据（如图像和文本）进行精心设计的扰动，来欺骗或操纵多模态模型，使其做出错误的预测或行为。这种攻击利用了模型在处理多模态数据时的弱点，尤其是在模型的预训练阶段和微调阶段的潜在差异。

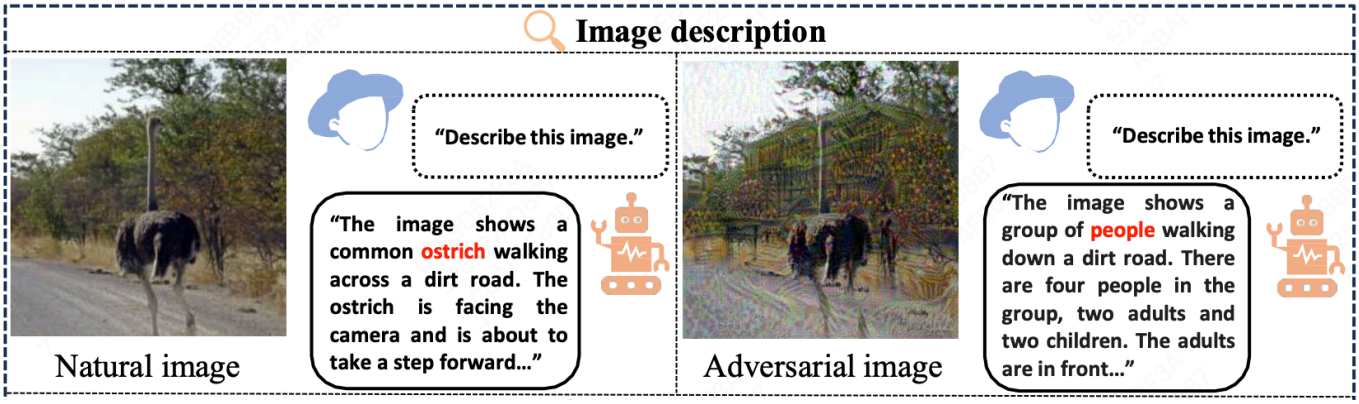

图1 图像描述任务中的多模态攻击展示

研究多模态对抗攻击是为了测试和揭示多模态大模型在实际应用中的潜在安全风险和脆弱性。通过研究多模态对抗攻击，可以更全面地评估这些模型在实际对抗环境中的鲁棒性和安全性并且可以为多模态模型对抗训练提供数据、提升多模态模型的鲁棒性。

- 研究点

1️⃣**当前的多模态非定向攻击停留在降低模型输出正确文本概率的阶段，在图像语义上的非定向攻击缺乏探索。**目前非定向攻击主要分为两类：一类是通过在图像上施加对抗扰动以降低模型生成给定正确文本的概率；另一类是在限定扰动范围的基础上，尽量使图像编码后的嵌入向量偏离初始状态。然而，这些方法并未充分考虑图像的语义层面。因为图像的正确描述可能有无穷多种，尽管对抗扰动能够成功避免了一种正确描述，模型仍有可能生成另一种符合图像的正确描述，这在一定程度上削弱了攻击的效果。例如，[图像="一只小猫趴在桌子上", prompt_1="你能描述一下这张图像吗"]，图像的正确描述可以是"一只小猫趴在桌子上"，也可以是"桌子上趴了一只小猫"，降低了其中一种正确描述但模型仍然可能生成另一种正确描述。对于这个问题，我们提出一种语义上的非定向攻击，旨在引导模型忽视图像中的关键要素，如关键实体、实体之间的相对位置、背景等等。

2️⃣**当前的多模态攻击忽视了链式思维（Chains of Thought）对模型推理的引导作用，这可能导致攻击效果下降。**当前的多模态攻击尚未包括对大模型思维过程的针对性攻击。然而，思维链作为一种逐步推理的方法，通过展示中间步骤帮助模型得出最终答案。这种方式强化了模型的每一步推理过程，使其能够进行"自我纠正"，从而更可能生成正确的文本内容，这很可能会导致多模态对抗攻击效果可能不理想。对于这个问题，我们可以将"引导模型直接思考"、"不要按照思维链的思维方式进行思考"等的语义内容嵌入图像扰动中，降低多模态大模型链式思维对攻击的抵抗作用。

3️⃣**当前的多模态攻击在计算图像扰动的过程中，通常忽视文本模态在生成内容时的引导作用。**以图像描述任务为例，绝大多数攻击方法专注于通过最大化攻击者期望输出的文本概率以计算图像扰动。然而这些方法没有充分考虑文本模态在内容生成中的协同作用，这很可能会导致图像扰动的效果可能不够理想。例如，在图像描述任务中，[图像="校运会上小明和小红在接力跑步", prompt_1="你能描述一下这张图像吗", prompt_2="这是一张校运会上拍的照片，你能描述一下这张图像吗"]，prompt_2在指导内容生成上比prompt_1更具优势。对于这个问题，我们就可以利用文本这种指导作用，在生成图像扰动的时候，正向优化文本引导模型输出正确文本，反向优化图像引导模型不输出正确文本，让两者在生成对抗图像的过程中具有一种对抗关系，从而让攻击更加有效。