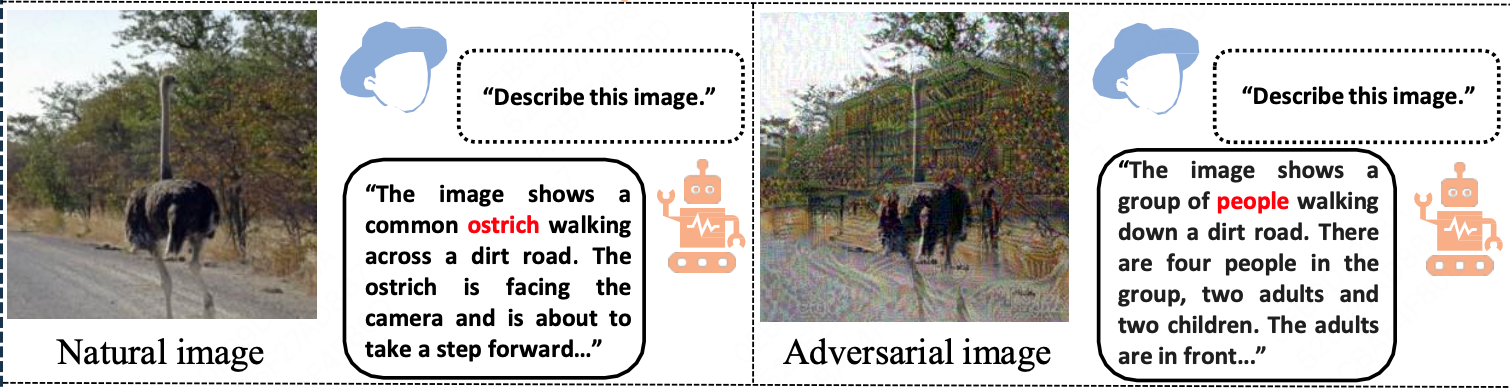
**（自己想的题目）基于图文对抗和思维链扰乱的语义多模态对抗攻击研究**

**多模态对抗攻击的研究必要性**

1. 随着人工智能技术的不断进步，多模态大模型在多种任务中的应用越来越广泛，展现出强大的能力和潜力。
2. 多模态对抗攻击是为了测试和展示多模态大模型在实际应用中的潜在安全风险和脆弱性。通过构造和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。

**背景**

定向攻击：通过图像上的对抗扰动引导模型输出攻击者想要的目标文本



非定向攻击：1. 通过图像上的对抗扰动引导模型尽量不输出正确的文本 2. 在一定大小的图像扰动下尽量使得图像经过编码器之后的嵌入偏离原始嵌入

**研究点**

1. **当前的多模态攻击停留在引导模型避免输出正确的文本描述的阶段，在图像语义上的非定向攻击缺乏探索**。以图像描述任务为例，非定向攻击尝试引导模型避免输出正确的文本描述，但由于图像的正确描述可能有无穷多种，尽管攻击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，旨在引导模型不输出图像中的关键要素，如忽视关键实体、背景颜色等等。
2. **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，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，从而让攻击更有效率。
3. **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。**思维链是一种逐步推理的方法，它通过展示中间推理步骤来帮助模型得出最终的答案。这种思维方式对模型的推理具有很强的指导作用，使得模型能够在每一步的思考过程中进行“自我纠正”，从而可能输出正确的文本内容，导致攻击效果不佳。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。
4. **备选：当前的多模态攻击往往针对特定任务进行设计，而没有考虑对抗图像在其他任务下也都能成功对模型造成误导或干扰，即缺乏任务上的攻击迁移性探索**。由于各种任务中对抗攻击的损失函数和优化目标不尽相同，导致为特定任务设计的对抗图像在其他任务上的攻击成功率往往不高。除此以外，当前的多模态大模型通常基于特定任务的攻击进行优化和防御，然而缺乏跨任务的对抗攻击研究可能会导致多模态大模型的潜在应用风险。因此，任务上的攻击迁移性既具备难点，也具备研究必要性。