View Review

Paper ID
1076

Paper Title
Embracing the Uncertainty: Evaluating Adversarial Robustness of Vision-Language Models from an Untargeted Attack Perspective

REVIEW QUESTIONS

1. How confident are you in your evaluation of this paper?
Confident

2. Importance/relevance to ICME
Of limited interest

3. Justification for importance/relevance
Assessing the robustness of VLMs against adversarial attacks is essential for their practical deployment. The paper addresses the limitations of previous methods that relied solely on optimizing the model's output, providing a more comprehensive approach by considering both output and internal feature uncertainties.

The proposed MIE method introduces a novel approach to generating adversarial examples by leveraging internal features of the model. This advancement in attack techniques helps in identifying the weaknesses of VLMs and can inform the development of more robust models.

4. Novelty/originality
Minor Originality

5. Justification for novelty/originality
The paper introduces a novel untargeted attack method based on Maximizing Information Entropy (MIE). Unlike previous methods that primarily focused on optimizing the model's output, MIE considers both the output distribution and the internal features of the model. This dual approach enhances the effectiveness of the attack by leveraging both external and internal characteristics of the model.

6. Technical correctness
Probably Correct

7. Justification for technical correctness
MIE does not require predefined textual targets, which allows for a broader perturbation space and makes the attack more versatile. This aspect of the method is novel as it removes the dependency on specific target texts, making the attack more adaptable to different scenarios.

8. Experimental validation and reproducibility

Limited but Convincing

9. Justification for experimental validation and reproducibility

It also provide extensive empirical validation across multiple models and datasets, demonstrating the effectiveness of the MIE method. The results show that MIE outperforms previous methods, establishing a new benchmark for assessing the adversarial robustness of VLMs.

10. Clarity of presentation

Clear Enough

11. Justification for clarity of presentation

The paper is well-organized with clear sections that follow a logical flow. It begins with an introduction that outlines the problem and motivation, followed by related work, methodology, experiments, results, and conclusion. This structure helps readers understand the context and progression of the research.

12. Reference to prior work

References Adequate

13. Justification for references

The references cover a wide range of topics related to vision-language models, adversarial attacks, and robustness evaluation. They include foundational works on VLMs, adversarial examples, and methods for enhancing model robustness, ensuring that the paper builds on existing literature.

14. Overall evaluation of the paper

Weak Reject

15. Justification for overall evaluation (required)

The method seems not new, jut changing partical of the original attack method. The paper focuses primarily on attacking the models rather than exploring defense mechanisms. While this is a valid approach for highlighting vulnerabilities, a more balanced evaluation would include a discussion of potential defenses and their effectiveness. While the paper evaluates multiple models and datasets, the diversity of these models and datasets could be expanded to include more recent and diverse architectures. This could provide a more comprehensive assessment of the research findings.

18. Is this an award-quality paper? (Only for Definite Accept papers)

Not a candidate for award