Adversarial Training in the Frequency Domain for Robust Multimodal Image Captioning Against Structured Perturbations

Assignee Research

June 13, 2026

Abstract

Multimodal machine learning models that combine visual and textual data are increasingly being deployed in critical applications, raising significant safety and security concerns due to their vulnerability to adversarial attacks. This paper presents an effective strategy to enhance the robustness of multimodal image captioning models against such attacks. By leveraging the Fast Gradient Sign Method (FGSM) to generate adversarial examples and incorporating adversarial training techniques, we demonstrate improved model robustness on two benchmark datasets: Flickr8k and COCO. Our findings indicat

1 Introduction

This paper examines: AI Safety in Practice: Enhancing Adversarial Robustness in Multimodal Image Captioning. Research question: To what extent does adversarial training in the frequency domain improve the robustness of multimodal models against structured perturbations in image captioning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

13 papers retrieved. 12 claims extracted; 11 independently verified. Quality review score: 7.6/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
In prior work [15], authors found that attacking only one input modality in multimodal deep learning models causes signi	×	0.12
The Fast Gradient Sign Method (FGSM) applies perturbations to input data in the direction of the gradient of the loss fu	✓	0.25
Projected Gradient Descent (PGD) generates adversarial examples by iteratively taking small steps toward the direction o	✓	0.29
JSMA is an iterative adversarial attack technique that computes a saliency map based on the Jacobian matrix to identify	✓	0.31
The C&W attack is an optimization-based adversarial attack that finds quasi-imperceptible perturbations to cause misclas	✓	0.24
Prior work [7] presented adversarial attacks against CLIP involving typographical manipulations (misspellings, font chan	✓	0.16
The authors in [13] proposed a fast adversarial training algorithm that recycles gradient information computed during tr	✓	0.19
The study demonstrates improved model robustness on the Flickr8k and COCO benchmark datasets using adversarial training	✓	0.18
Selectively training only the text decoder of the multimodal architecture yields performance comparable to full adversar	✓	0.26
Selectively training only the text decoder offers increased computational efficiency compared to full adversarial traini	✓	0.20
The architecture used combines GPT-2 as a decoder and Vision Transformer (ViT) as an encoder.	✓	0.18
In the described architecture, ViT patches are flattened and projected linearly into a space with dimensions of 768.	✓	0.19

References

- `http://arxiv.org/abs/2405.18770v6`

- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2407.21174v1