Adversarial Pre-Training for Zero-Shot Robustness in Large Language and Multimodal Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: Can the adversarial pre-training strategy in APT be extended to other foundational models (e.g., LLMs or multimodal models) to improve their zero-shot performance on out-of-distribution language or. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving Adversarial Transferability of Vision-Language Pre-training Models through Collaborative Multimodal Interaction. Research question: Can the adversarial pre-training strategy in APT be extended to other foundational models (e.g., LLMs or multimodal models) to improve their zero-shot performance on out-of-distribution language or vision-language benchmarks like GLUE, SuperGLUE, or VQA v2?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

10 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
VLP models play a crucial role in offering a universal solution for multiple tasks, including image-text retrieval (ITR)	×	0.09
Recent studies have elucidated the vulnerability and sensitivity of VLP models to adversarial examples.	×	0.06
Single-modal attacks such as PGD and BERT-Attack exhibit good adversarial performance in the visual and text domains.	×	0.06
Applying single-modal attacks directly to VLP models still poses challenges because VLP models integrate multimodal info	×	0.08
Sep-Attack directly combines both BERT-Attack and PGD.	×	0.03
Co-Attack considers image-text collaborative information and is specifically designed for customized attack forms for di	×	0.10
The proposed Collaborative Multimodal Interaction Attack demonstrates effectiveness in experimental results.	✓	0.18
VLP models can be classified into two categories: Fused VLP models and Aligned VLP models.	×	0.06
Fused VLP models (e.g., ALBEF, TCL) employ a structural approach that initially utilizes a single encoder to extract fea	×	0.06
Aligned VLP models (e.g., CLIP) use a single encoder to independently learn feature representations.	×	0.05

References

- http://arxiv.org/abs/2306.05540v1

- http://arxiv.org/abs/2404.14700v4
- http://arxiv.org/abs/2403.10883v2