Adversarial Pretraining and Robustness-Accuracy Trade-offs in Multimodal Transformers

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does adversarial pretraining affect the robustness-accuracy trade-off in multimodal transformers compared to standard data augmentation on vision-language benchmarks. 13 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Pretraining of Self-Supervised Deep Networks: Past, Present and Future. Research question: How does adversarial pretraining affect the robustness-accuracy trade-off in multimodal transformers compared to standard data augmentation on vision-language benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

16 papers retrieved. 13 claims extracted; 12 independently verified. Quality review score: 7.3/10.

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
Adversarial pretraining of self-supervised deep networks includes both convolutional neural networks and vision transfor	✓	0.35
Adversarial training with access to labeled examples differs from adversarial pretraining, which only has access to unla	✓	0.25
Existing approaches to incorporating adversaries into pretraining models are largely categorized into two groups: memory	✓	0.35
Memory-free instance-wise attacks impose worst-case perturbations on individual examples.	✓	0.25
Memory-based adversaries are shared across examples over iterations.	✓	0.25
Contrastive Learning (CL) and Masked Image Modeling (MIM) are two popular self-supervised pretraining methods in literat	✓	0.33
The paper reviews representative adversarial pretraining models based on Contrastive Learning (CL).	✓	0.25
The paper reviews representative adversarial pretraining models based on Masked Image Modeling (MIM).	✓	0.25
The paper reviews issues regarding computing overheads in adversarial pretraining.	×	0.14
The paper reviews issues regarding input-level and feature-level adversaries.	✓	0.17
The paper discusses emerging trends regarding the relations between adversarial and cooperative pretraining.	✓	0.17
The paper discusses emerging trends regarding unifying adversarial CL and MIM pretraining.	✓	0.20
The paper discusses the trade-off between accuracy and robustness in adversarial pretraining.	✓	0.17

References

- http://arxiv.org/abs/2505.14042v3
- http://arxiv.org/abs/2405.18770v6

- http://arxiv.org/abs/2210.13463v1