Adversarial Robustness of Contrastive vs. MLM-Based CodeT5 on CWE-200

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does adversarial robustness (measured via targeted perturbation success rate) differ between contrastively pretrained and MLM-based CodeT5 models on the CWE-200 benchmark. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Adversarial Robustness of Vision Transformers. Research question: How does adversarial robustness (measured via targeted perturbation success rate) differ between contrastively pretrained and MLM-based CodeT5 models on the CWE-200 benchmark?.

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 5.8/10.

3 Results

16 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 5.8/10.

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
ViTs are more robust to high-frequency perturbations than CNNs.	✓	0.19
CNN models take up high-frequency patterns that are almost imperceptible to humans but have distribution correlation wit	×	0.10
Original ViT models possess superior adversarial robustness against high-frequency perturbations compared with CNNs.	✓	0.15
Introducing non-transformer modules (e.g., ResNet blocks and T2T blocks) to ViTs diminishes their original adversarial r	×	0.13
Hybrid ViTs (e.g., ResViTs and T2T-ViTs) exhibit inferior adversarial robustness against high-frequency perturbations co	×	0.14
ViTs pay less attention to high-frequency patterns in images compared to CNNs.	✓	0.15
CNNs learn more low-level features compared with ViTs.	×	0.08
ViT feature maps become noisier when ResNet features are introduced (ViT-B/16-Res).	×	0.03
ViT feature maps become noisier when neighboring tokens are aggregated into one token recursively (T2T-ViT-24).	×	0.03
Clean Accuracy (CA) is evaluated on the entire ImageNet-1k test set.	×	0.02
Robust Accuracy (RA) is evaluated on adversarial examples generated with 1,000 test samples.	×	0.07
Under High-pass filtered PGD attack with epsilon 0.001, standard ViT models achieve higher Robust Accuracy than CNN mode	×	0.04
Under High-pass filtered PGD attack with epsilon 0.1, hybrid models like T2T-ViT-24 show significantly lower Robust Accu	×	0.04
Swin-L/4 achieves a Clean Accuracy of 84.2% on ImageNet-1k.	×	0.01
ViT-SAM-B/16 achieves a Robust Accuracy of 63.4% against adversarial perturbations with epsilon 0.001.	×	0.05
Deit-T/16 achieves a Robust Accuracy of 0.3% against adversarial perturbations with epsilon 0.01.	×	0.04

References

- http://arxiv.org/abs/2305.00866v2
- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/2103.15670v3