Preprocessing Defenses and Adversarial Training for Robust Code Generation in CodeT5

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does combining preprocessing defenses with adversarial training yield higher robustness scores for CodeT5 on code generation tasks under iterative PGD attacks compared to single-step FGSM. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robust Image Classification: Defensive Strategies against FGSM and PGD Adversarial Attacks. Research question: Does combining preprocessing defenses with adversarial training yield higher robustness scores for CodeT5 on code generation tasks under iterative PGD attacks compared to single-step FGSM?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

12 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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
The concepts of adversarial examples and the FGSM attack were introduced in reference [1].	×	0.09
Adversarial training can be computationally expensive and may not generalize well to unseen types of attacks.	×	0.05
PGD was proposed in reference [2] as a robust method for generating adversarial examples and for use in adversarial trai	×	0.11
Adversarial training with PGD significantly enhances the robustness of deep learning models.	×	0.14
Reference [6] introduces sophisticated attacks that successfully bypass ten state-of-the-art detection methods.	×	0.05
Reference [6] does not propose any improved detection mechanisms for the attacks it presents.	×	0.05
A novel adversarial attack targeting image captioning models using attention-based optimization techniques was proposed	×	0.05
On the MNIST dataset, the model accuracy drops from 0.9927 at noise level 0.00 to 0.0122 at noise level 0.30.	×	0.03
On the MNIST Fashion dataset, the model accuracy remains relatively stable between 0.2943 and 0.2956 for noise levels ra	×	0.03
In the second experiment table, MNIST accuracy decreases from 0.9909 at noise level 0.00 to 0.0528 at noise level 1.00.	×	0.01
For the defense mechanism tested in Table (p7), the time required to defend an attack on MNIST is 0.003 seconds at noise	×	0.04
At noise level 0.00, the test accuracy on MNIST Fashion is 0.7479 with a defense time of 0.003 seconds.	×	0.02
In the final table presented, the defense time for both MNIST and MNIST Fashion datasets is consistently 0.0012 seconds	×	0.02
At noise level 0.00 in the final table, the test accuracy on MNIST is 0.9203 and on MNIST Fashion is 0.7049.	×	0.01

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2011.05157v2
- http://arxiv.org/abs/1702.06763v8