Adversarial Training Batch Size Scaling and Robustness Gaps in CodeT5 on MBXP Python

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of scaling adversarial training batch sizes on the FGSM vs. PGD robustness gap in CodeT5 for the MBXP Python subset, measured by accuracy differentials under varying perturbation. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. 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: What is the impact of scaling adversarial training batch sizes on the FGSM vs. PGD robustness gap in CodeT5 for the MBXP Python subset, measured by accuracy differentials under varying perturbation budgets?.

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 3.1/10.

3 Results

13 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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.	×	0.06
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.	×	0.06
Reference [6] does not propose any improved detection mechanisms for the attacks it introduces.	×	0.05
A novel adversarial attack targeting image captioning models using attention-based optimization techniques was proposed	×	0.08
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	×	0.02
With the proposed defense on MNIST, the test accuracy at noise level 0.00 is 0.9569 with a defense time of 0.003 seconds	×	0.02
With the proposed defense on MNIST, the test accuracy at noise level 1.00 is 0.3571 with a defense time of 0.0003 second	×	0.02
With the proposed defense on MNIST Fashion, the test accuracy at noise level 0.00 is 0.7479.	×	0.02
In the second defense evaluation table, the test accuracy on MNIST at noise level 0.05 is 0.9342.	×	0.03
In the second defense evaluation table, the time for defending an attack is consistently 0.0012 seconds across all teste	×	0.03

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2011.05157v2
- http://arxiv.org/abs/2410.21676v4