Gradient Masking Effects on PGD Adversarial Transferability in CodeT5 Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of gradient masking preprocessing techniques on the transferability of PGD-generated adversarial examples against CodeT5 models trained with FGSM adversarial examples. 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: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. Research question: What is the impact of gradient masking preprocessing techniques on the transferability of PGD-generated adversarial examples against CodeT5 models trained with FGSM adversarial examples?.

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 3.1/10.

3 Results

12 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
Adversarial retraining has been shown to be difficult at ImageNet scale.	×	0.03
Training exclusively on ∞ adversarial examples provides only limited robustness to adversarial examples under other dis	×	0.06
Cascade adversarial training involves training a first model, generating adversarial examples on that model using iterat	×	0.07
Cascade adversarial training achieves 16% accuracy with ϵ = 0.015.	×	0.06
Adversarial training as described in Madry et al. (2018) achieves over 70% accuracy at a perturbation budget of ϵ = 0.01	×	0.07
Goodfellow et al. (2014b) argue that adversarial examples exist because neural networks behave in a largely linear manne	×	0.04
Thermometer encoding is designed to break the linearity of neural networks.	×	0.03
On CIFAR-10, performing only thermometer encoding yields 50% accuracy within ϵ = 0.031 under ∞ distortion.	×	0.03
Performing adversarial training with 7 steps of Logit-Space Projected Gradient Ascent (LS-PGA) on thermometer encoded ne	×	0.09
Adversarial examples generated on a standard adversarially trained model transfer to a thermometer encoded model, reduci	×	0.08
The authors construct adversarial examples using iterative optimization-based methods, specifically Projected Gradient D	×	0.09
The authors use the Lagrangian relaxation of Carlini & Wagner (2017c) to generate 2 bounded adversarial examples.	×	0.03
Seven of the ICLR 2018 defenses rely on the effect of gradient masking.	×	0.15
Gradient masking is known to be an incomplete defense to adversarial examples.	×	0.08

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2304.06908v1
- http://arxiv.org/abs/1802.00420v4