Gradient Masking Effects on GNN-Based NIDS Robustness Against Structural Adversarial Attacks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of gradient masking techniques on the robustness of GNN-based NIDS models against structural adversarial attacks as measured by the AUC-ROC score on the KDD Cup 99 dataset. We identify obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false sense of security in defenses against adversarial examples. While defenses that cause obfuscated gradients appear to defeat iterative optimization-based attacks, we find defenses. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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 techniques on the robustness of GNN-based NIDS models against structural adversarial attacks as measured by the AUC-ROC score on the KDD Cup 99 dataset?.

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 3.2/10.

3 Results

12 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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 (Kurakin et al., 2016b).	×	0.02
Training exclusively on ∞ adversarial examples provides only limited robustness to adversarial examples under other dis	×	0.08
Cascade adversarial machine learning (Na et al., 2018) achieves 16% accuracy with ϵ = .015, compared to over 70% at the	×	0.03
Thermometer encoding on CIFAR-10 gives 50% accuracy within ϵ = 0.031 under ∞ distortion.	×	0.01
Performing adversarial training with 7 steps of LS-PGA on thermometer encoded networks increases robustness to 80% on CI	×	0.02
Adversarial examples generated on a standard adversarially trained model transfer to a thermometer encoded model reducin	×	0.04
Projected Gradient Descent (PGD) is used to generate ∞ bounded adversarial examples.	×	0.06
Lagrangian relaxation of Carlini & Wagner (2017c) is used to generate 2 bounded adversarial examples.	×	0.05
Between 100 and 10,000 iterations of gradient descent are used to obtain convergence in generating adversarial examples.	×	0.07
7 of the ICLR 2018 defenses rely on gradient masking.	×	0.15

References

- http://arxiv.org/abs/2104.09369v1
- http://arxiv.org/abs/1802.00420v4
- http://arxiv.org/abs/1909.08072v2