Adversarial Training Effects on GAS and Mul-GAD Anomaly Detection Robustness

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of adversarial training on the anomaly detection accuracy of GAS versus Mul-GAD when evaluated on perturbed graph structures. Deep neural networks (DNN) have achieved unprecedented success in numerous machine learning tasks in various domains. However, the existence of adversarial examples has raised concerns about applying deep learning to safety-critical applications. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Research question: What is the impact of adversarial training on the anomaly detection accuracy of GAS versus Mul-GAD when evaluated on perturbed graph structures?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

10 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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 work of Carlini et al. (2017) uses Reluplex to find the theoretical minimally-distorted adversarial examples, referr	×	0.08
Reluplex encodes model parameters and data as a linear-like programming system to check for the existence of an eligible	×	0.03
Eykholt et al. (2017) crafted physical adversarial objects by placing stickers on road signs to mislead road sign recogn	×	0.01
Eykholt et al. (2017) used a two-step approach: first using an l1-norm based attack to find perturbation regions, then u	×	0.05
The perturbed stop sign created by Eykholt et al. (2017) can confuse an autonomous vehicle's road sign recognizer from a	×	0.00
Athalye et al. (2017) reported the first work to successfully craft physical 3D adversarial objects using 3D-printing.	×	0.03
Athalye et al. (2017) optimized a 3D object's texture so that rendering images remain adversarial under varying camera d	×	0.03
Papernot et al. (2017) introduced the first effective algorithm to attack DNN classifiers using a substitute model.	×	0.05
The ground-truth attack is the first work to calculate the exact robustness (minimal perturbation) of classifiers.	×	0.05
The ground-truth attack method involves using an SMT solver, making it slow and not scalable to large networks.	×	0.02
Su et al. (2019) demonstrated that on the CIFAR10 dataset, 63.5% of testing samples can be attacked by changing only one	×	0.01
The VGG16 classifier has 85.5% accuracy on the CIFAR10 test data.	×	0.02
The One-pixel Attack constrains the perturbation's l0 norm to limit the number of pixels allowed to be changed.	×	0.02

References

- http://arxiv.org/abs/2104.09369v1
- http://arxiv.org/abs/1909.08072v2
- http://arxiv.org/abs/2305.02496v1