Graph Diffusion Models and Sparse GNNs: Robustness to Adversarial Structural Perturbations

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do graph diffusion models and sparse GNN architectures differ in robustness scores against adversarial structural perturbations on large-scale graph benchmarks. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On Adversarial Robustness: A Neural Architecture Search perspective. Research question: How do graph diffusion models and sparse GNN architectures differ in robustness scores against adversarial structural perturbations on large-scale graph benchmarks?.

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 2.4/10.

3 Results

12 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 2.4/10.

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 study presents results only for the CIFAR-10 and ImageNet datasets because pre-trained models are available only for	×	0.07
An ensemble of architectures is known to be more adversarially robust than a single architecture.	×	0.06
The search cost associated with building architectures by randomly sampling cells from the DARTS search space is zero.	×	0.11
For the CIFAR-10 dataset, DARTS architectures with 20 cells are trained for 600 epochs.	×	0.08
In the ensemble method described, the number of training epochs for each network is determined based on the number of ce	×	0.05
The ensemble as a whole is effectively trained for 600 epochs to ensure a fair comparison with existing approaches.	×	0.02
Adversarial training often decreases accuracy on clean (un-perturbed) samples.	×	0.08
EfficientNet-B0 has 5.29M parameters, 91.36% Clean accuracy, 8.11% PGD accuracy, 25.61% AutoPGD accuracy, and 14.90% PP-	×	0.02
EfficientNet-B7 has 66.35M parameters, 91.57% Clean accuracy, 11.20% PGD accuracy, 27.82% AutoPGD accuracy, and 1.59% PP	×	0.02
ResNet-18 has 11.69M parameters, 89.08% Clean accuracy, 2.41% PGD accuracy, 21.65% AutoPGD accuracy, and 4.69% PP-HRS ac	×	0.03
The DARTS architecture uses 5 Max Pool, 0 Avg Pool, 3 Skip connection, 0 Separable Conv, and 0 Dilated Conv operations.	×	0.02
The PC-DARTS architecture uses 1 Max Pool, 0 Avg Pool, 0 Skip connection, 7 Separable Conv, and 0 Dilated Conv operation	×	0.02
When transferring from ResNet-18 to ResNet-18, the error rate is 9.59%.	×	0.05
PC-DARTS achieves an error rate of 7.96% when the source is ResNet-18 and 8.02% when the source is DenseNet-169.	×	0.02
Attackers can target autonomous vehicles by using stickers or paint to create an adversarial stop sign that the vehicle	×	0.03

References

- http://arxiv.org/abs/2007.08428v4
- http://arxiv.org/abs/2104.09630v2
- http://arxiv.org/abs/2104.09369v1