GCN-Enhanced Multimodal Models vs. Transformer Baselines in Throughput and Adversarial Robustness on Hateful Memes

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How do GCN-enhanced multimodal models compare to transformer-only baselines in terms of throughput and robustness against adversarial text-image perturbations on the Hateful Memes dataset. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: How do GCN-enhanced multimodal models compare to transformer-only baselines in terms of throughput and robustness against adversarial text-image perturbations on the Hateful Memes dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

6 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 Ground Truth Attack by Carlini et al. (2017) uses Reluplex to find the theoretical minimally-distorted adversarial e	×	0.07
Reluplex encodes model parameters and data as subjects of a linear-like programming system to check for eligible adversa	×	0.03
Eykholt et al. (2017) crafted physical adversarial objects by placing stickers on road signs to mislead road sign recogn	×	0.04
Eykholt et al. (2017) used a two-step approach: first an l1-norm based attack to find perturbation regions, followed by	×	0.03
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.	×	0.03
The Ground Truth Attack method involves using an SMT solver, making it slow and not scalable to large networks.	×	0.04
Su et al. (2019) demonstrated that on the CIFAR10 dataset, 63.5% of testing samples for a VGG16 classifier can be attack	×	0.00
VGG16 has 85.5% accuracy on the CIFAR10 test data.	×	0.03
Constraining the l0 norm of a perturbation limits the number of pixels allowed to be changed.	×	0.00

References

- http://arxiv.org/abs/1801.04693v1
- http://arxiv.org/abs/2003.07729v1

- http://arxiv.org/abs/1909.08072v2