Kolmogorov-Arnold Networks vs. MLPs in Adversarial Robustness on CIFAR-10

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the robustness of KANs against adversarial attacks compare to MLPs when measured using the FGSM attack success rate on the CIFAR-10 dataset. 16 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Adversarial Robustness: A Comparison Of FGSM, Carlini-Wagner Attacks, And The Role of Distillation as Defense Mechanism. Research question: How does the robustness of KANs against adversarial attacks compare to MLPs when measured using the FGSM attack success rate on the CIFAR-10 dataset?.

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 4.2/10.

3 Results

10 papers retrieved. 16 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
The study evaluated the impact of FGSM and CW attacks on three pre-trained CNN architectures: resnext50_32xd, Densenet20	✓	0.16
The models were initially trained on the ImageNet dataset.	×	0.09
The study explored the efficacy of Defensive Distillation in mitigating the effects of adversarial attacks.	✓	0.15
The Tiny ImageNet dataset, which has 200 classes, was used for the research.	×	0.07
The study used both top 1 and top 5 accuracy scores for evaluation.	×	0.03
Top 1 accuracy indicates the percentage of successfully categorized images when the top prediction matches the actual la	×	0.02
Top 5 accuracy evaluates if the true label occurs in any of the top 5 predicted classes by the model.	×	0.02
The classification performance of Resnext50_32x4d, DenseNet201, and VGG-19 models before attack is shown in Table 1.	×	0.07
The top-1 error rates for Resnext50_32x4d, DenseNet201, and VGG-19 are 10.16%, 13.92%, and 19.88% respectively.	×	0.06
The top-5 error rates for Resnext50_32x4d, DenseNet201, and VGG-19 are 1.20%, 2.22%, and 4.38% respectively.	×	0.05
The study used PyTorch's torchvision package to evaluate the models.	×	0.04
The Tiny ImageNet dataset was used to evaluate the models and define baselines for essential performance.	×	0.06
The study measured the models' intrinsic ability to perform picture classification tasks by computing key classification	×	0.11
The study utilized two well-known adversarial attack techniques: the CW attack and the FGSM.	×	0.13
The study systematically altered the epsilon values, ranging from 1% to 10%, for both attack approaches.	×	0.03
The study investigated classification accuracy metrics in detail and documented classification mistakes at various epsil	×	0.06

References

- http://arxiv.org/abs/2404.04245v1
- http://arxiv.org/abs/2307.02055v1
- http://arxiv.org/abs/2102.08868v2