Growth Bound Matrix for Adversarial Robustness in Recurrent and State Space Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Does the Growth Bound Matrix method maintain accuracy on out-of-distribution datasets while defending against adversarial synonym substitutions in recurrent and state space architectures. 15 claims were extracted from source literature; 1 was 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: Frequency Centric Defense Mechanisms against Adversarial Examples. Research question: Does the Growth Bound Matrix method maintain accuracy on out-of-distribution datasets while defending against adversarial synonym substitutions in recurrent and state space architectures?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 15 claims extracted; 1 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 time taken to classify a single CIFAR-10 image with one feature is 6.46 $\times$ 10-5 seconds.	×	0.06
The time taken to classify a single ImageNet image with one feature is 4.9 $\times$ 10-3 seconds.	×	0.05
The time taken to classify a single CIFAR-10 image with all features is 1.35 $\times$ 10-4 seconds.	×	0.06
The time taken to classify a single ImageNet image with all features is 1.02 $\times$ 10-2 seconds.	×	0.05
The proposed scheme achieves a detection accuracy of 99.5% for FGSM attacks and 99.8% for PGD attacks on the CIFAR-10 da	✓	0.18
The proposed scheme achieves a detection accuracy of 72.6% for DeepFool attacks and 62.6% for CW attacks on the CIFAR-10	×	0.12
SpectralDefense (InputMFS) achieves a detection accuracy of 98.1% for FGSM and 93.6% for PGD on the CIFAR-10 dataset.	×	0.10
Local Intrinsic Dimensionality (LID) achieves a detection accuracy of 78.9% for DeepFool and 78.1% for CW on the CIFAR-1	×	0.08
Mahalanobis Distance (MD) achieves a detection accuracy of 95.6% for FGSM and 96.0% for PGD on the CIFAR-10 dataset.	×	0.09
The combined approach (Entropy + MFS + PFS) achieves an accuracy of 98.8% against FGSM attacks.	×	0.05
The combined approach (Entropy + MFS + PFS) achieves an accuracy of 51.8% against DeepFool attacks.	×	0.05
The magnitude of perturbation (δ) observed in the ImageNet dataset is lesser than in the CIFAR-10 dataset for every atta	×	0.08
The proposed approach fails to detect adversarial examples generated by DeepFool and CW on the ImageNet dataset.	×	0.08
The CIFAR-10 model architecture uses 10 kernels in the first convolutional layer and 20 nodes in the first dense layer.	×	0.05
The ImageNet model architecture uses 100 kernels in the fourth convolutional layer and 30 nodes in the first dense layer	×	0.03

References

- http://arxiv.org/abs/2110.13935v1
- http://arxiv.org/abs/2212.06370v4
- http://arxiv.org/abs/1903.06620v2