Clean Accuracy Degradation in Certified Lipschitz-Bounded Versus Adversarially Trained Vision Transformers Beyond 100M Parameters

Assignee Research

June 12, 2026

Abstract

Lipschitz bounded neural networks are certifiably robust and have a good trade-off between clean and certified accuracy. Existing Lipschitz bounding methods train from scratch and are limited to moderately sized networks ($<$ 6M parameters). They require a fair amount of hyper-parameter tuning and are computationally prohibitive for large networks like Vision Transformers (5M to 660M parameters). Obtaining certified robustness of transformers is not feasible due to the non-scalability and inflexibility of the current methods. This work presents CertViT, a two-step proximal-projection method to a

1 Introduction

This paper examines: CertViT: Certified Robustness of Pre-Trained Vision Transformers. Research question: How does the clean accuracy degradation of certified Lipschitz-bounded Vision Transformers compare to adversarially trained ViTs when scaling beyond 100M parameters on ImageNet-1K?.

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 6.8/10.

3 Results

12 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 6.8/10.

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 training sequence is split into J mini-batches of size T so that K = JT.	✓	0.25
The Douglas-Rachford algorithm uses positive parameters β and (λn)n$\in$N.	✓	0.27
The projection step relaxβ reduces the magnitude of each parameter or element of the input matrix Wn by β (i.e (sign(Wn)	✓	0.35
CertViT networks have better certified accuracy than state-of-the-art Lipschitz trained networks.	✓	0.32
CertViT predicts the input image of panda correctly, while other methods predict it as badger.	✓	0.16
Global Lipschitz constant is computationally cheap and scalable but often loose and hence tends to over-regularize the t	✓	0.27
Local Lipschitz estimates are tight since it uses information in the local neighborhood of the input data but are hard t	✓	0.25

References

- http://arxiv.org/abs/2302.10287v1
- http://arxiv.org/abs/2111.15121v2
- http://arxiv.org/abs/2101.11986v3