Certified Defense and Adversarial Training Trade-offs in Large-Scale Vision Transformers

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the comparative impact of adversarial training versus certified defense methods on the clean accuracy and robustness trade-off in large-scale Vision Transformers. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CertViT: Certified Robustness of Pre-Trained Vision Transformers. Research question: What is the comparative impact of adversarial training versus certified defense methods on the clean accuracy and robustness trade-off in large-scale Vision Transformers?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

11 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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 training sequence is split into J mini-batches of size T so that K = JT.	×	0.05
The Douglas-Rachford algorithm uses positive parameters β and (λn)n$\in$N.	×	0.03
Proposition 3 [34] Assume that Problem 2 has a solution and that there exists W $\in$C such W is a point in the interior of	×	0.02
The projection step relaxβ reduces the magnitude of each parameter or element of the input matrix Wn by β (i.e (sign(Wn)	×	0.06
CertViT networks have better certified accuracy than state-of-the-art Lipschitz trained networks.	✓	0.34
CertViT is applied on several variants of pre-trained vision transformers and shows adversarial robustness using standar	✓	0.29
CertViT predicts the input image of panda correctly while other methods predict it as badger.	×	0.04
Formal guarantees are of much importance in mission- and safety-critical applications.	×	0.09
Global Lipschitz constant is computationally cheap and scalable but often loose and hence tends to over-regularize the t	×	0.03
Local Lipschitz estimates are tight since it uses information in the local neighborhood of the input data but are hard t	×	0.02

References

- http://arxiv.org/abs/2306.12610v1
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2302.10287v1