Dataset Compression Ratio Effects on CodeT5 Adversarial Robustness under DKD Distillation

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of dataset compression ratio on the adversarial robustness of CodeT5 when distilled with DKD, as measured by accuracy metrics across different adversarial attack strengths. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: BEARD: Benchmarking the Adversarial Robustness for Dataset Distillation. Research question: What is the impact of dataset compression ratio on the adversarial robustness of CodeT5 when distilled with DKD, as measured by accuracy metrics across different adversarial attack strengths?.

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 4.5/10.

3 Results

12 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 4.5/10.

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 adversarial robustness of existing DD methods is evaluated across multiple datasets using three key metrics: Robustn	✓	0.28
The Robustness Ratio (RR) measures attack effectiveness.	×	0.10
The Attack Efficiency Ratio (AE) measures attack efficiency.	×	0.13
The Comprehensive Robustness-Efficiency Index (CREI) provides an overall evaluation of adversarial robustness.	✓	0.15
The adversarial game framework is used to enhance the evaluation of adversarial robustness in dataset distillation.	✓	0.18
The defender function D encompasses DD methods with diverse IPC settings.	×	0.10
The model M is trained on the distilled dataset generated by D with diverse IPC settings d $\in$ D.	×	0.08
The attacker function A has a perturbation budget ϵ $\in$ P.	×	0.02
Previous studies on adversarial robustness in DD have mainly focused on model accuracy, which provides an incomplete vie	×	0.08
The adversarial perturbation is constrained by x -xp $\leq$$\epsilon$.	×	0.04
The attacker function A maps an input and a hypothesis to an adversarially perturbed version of the input.	×	0.00
The defender function D aims to generate a synthetic dataset S such that a model M trained on S performs comparably to o	×	0.06
The CREI on CIFAR-10 for targeted attack is 43.97 for Full-size and 52.62 for DC.	×	0.04
The CREI on CIFAR-10 for untargeted attack is 21.91 for Full-size and 23.89 for DC.	×	0.04
The CREI on CIFAR-100 for untargeted attack is 17.29 for Full-size and 14.44 for DC.	×	0.03
The CREI on TinyImageNet for untargeted attack is 29.83 for Full-size and 29.96 for DC.	×	0.03

References

- http://arxiv.org/abs/2404.04245v1
- http://arxiv.org/abs/2403.13322v3
- http://arxiv.org/abs/2411.09265v1