Contrastive Learning and Adversarial Perturbations Enhance CodeT5 Robustness to Identifier Renaming

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does integrating contrastive learning with adversarial perturbations during pre-training improve CodeT5's resilience to identifier renaming variations in functional correctness metrics. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robust Pre-Training by Adversarial Contrastive Learning. Research question: Does integrating contrastive learning with adversarial perturbations during pre-training improve CodeT5's resilience to identifier renaming variations in functional correctness metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 proposed ACL (DS) yields a significant improvement on [TA, RA] by [2.14%, 2.99%] on CIFAR-10, and [2.14%, 3.58%] on	×	0.07
ACL (DS) establishes the new benchmark TA/RA numbers on CIFAR-10 and CIFAR-100.	×	0.09
The proposed ACL (DS) leads to an improvement of 2.42% when averaged over all unforeseen attacks.	×	0.03
ACL (DS) outperforms on most unforeseen attack types, showing more general robustness gains.	×	0.05
For the fully-supervised tuning, we employ the loss in TRADE [29] for adversarial fine-tuning, with the regularization w	×	0.02
We use SGD with 0.9 momentum and batch size 128.	×	0.07
By default, we fine-tune 25 epochs, with initial learning rate set as 0.1 and then decaying by 10 times at epoch 15 and	×	0.03
For contrastive pre-training, we identically follow SimCLR [2] for all the optimizer settings, augmentation and projecti	×	0.08
We choose 512 for batch size and train for 1000 epochs.	×	0.09
To generate adversarial perturbations, we use the ∞PGD attack [15], following all hyperparameters used by [1], except t	×	0.09

References

- `http://arxiv.org/abs/2212.12411v1`

- http://arxiv.org/abs/2302.09195v5
- http://arxiv.org/abs/2010.13337v1