Fine-Tuning on Adversarial GLUE for Gradient and Attention Attribution Stability

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does fine-tuning on Adversarial GLUE datasets improve the stability of gradient-based attribution methods compared to attention-based methods under perturbed inputs. Adversarial perturbations are noise-like patterns that can subtly change the data, while failing an otherwise accurate classifier. In this paper, we propose to use such perturbations within a novel contrastive learning setup to build negative samples, which are then used to. 11 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: Contrastive Video Representation Learning via Adversarial Perturbations. Research question: Does fine-tuning on Adversarial GLUE datasets improve the stability of gradient-based attribution methods compared to attention-based methods under perturbed inputs?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 11 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 optimization produces useful representations in about 50 iterations and takes about 5 milli-seconds per frame on a s	×	0.03
The slack regularization constant C is set to 1.	×	0.02
HMDB-51 consists of 6766 Internet videos over 51 classes; each video is about 20 – 1000 frames.	×	0.03
The standard evaluation protocol for HMDB-51 reports average classification accuracy on three-folds.	×	0.01
For HMDB-51, features from the pool5 layer of each stream are sequences of 2048D vectors.	×	0.02
NTU-RGBD has 56,880 video sequences across 60 classes, 40 subjects, and 80 views.	×	0.03
NTU-RGBD videos have on average 70 frames and consist of people performing various actions; each frame is annotated for	×	0.01
For NTU-RGBD, two evaluation protocols are used, namely cross-view and cross-subject evaluation.	×	0.02
For NTU-RGBD, 256D features from the bottleneck layer are used as input to the scheme.	×	0.03
YUP++ dataset has 20 scene classes with 60 videos in each class.	×	0.03
In YUP++, half of the sequences in each class are collected by a static camera and the rest are recorded by a moving cam	×	0.01

References

- http://arxiv.org/abs/1905.11736v5
- http://arxiv.org/abs/1807.09380v3

- http://arxiv.org/abs/2007.04137v3