Integrated Decision Gradients and Attention Rollout Robustness Under Adversarial Text Perturbations

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the robustness of Integrated Decision Gradients compare to Attention Rollout in maintaining feature attribution consistency under adversarial text perturbations across standard NLP benchmark. Large-scale pre-trained language models have achieved tremendous success across a wide range of natural language understanding (NLU) tasks, even surpassing human performance. However, recent studies reveal that the robustness of these models can be challenged by carefully. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. Research question: How does the robustness of Integrated Decision Gradients compare to Attention Rollout in maintaining feature attribution consistency under adversarial text perturbations across standard NLP benchmark datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

14 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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 data curation phase serves as a comprehensive benchmark over existing adversarial attack methods.	×	0.13
The data curation phase provides a fair standard for all adversarial attacks and systematic human annotations to evaluat	×	0.07
Table 4 reports model performance on the AdvGLUE test set, including BERT (Large) and RoBERTa (Large) fine-tuned with di	×	0.04
For MNLI, Table 4 reports test accuracy on the matched and mismatched test sets.	×	0.01
For QQP, Table 4 reports accuracy and F1.	×	0.02
For other tasks, Table 4 reports accuracy.	×	0.03
Table 4 reports the macro-average (Avg) of per-task scores for different models.	×	0.04
Table 9 reports model performance on the AdvGLUE test set and GLUE dev set.	×	0.06
The first four typo strategies guarantee the word edit distance between the typo word and its original word to be 1, and	×	0.01
In Strategy (i), a space is inserted into a word only when the word contains less than 6 characters.	×	0.02
In Strategy (v), characters in a word are swapped only when the word has more than 4 characters.	×	0.02
For sentiment analysis tasks, the cosine similarity threshold is set to 0.8 to encourage synonyms to be semantically clo	×	0.03
Table [Table (p4)] lists different adversarial attack methods including Embedding, TextBugger, TextFooler, BERT-ATTACK,	×	0.04
Table [Table (p17)] reports the average performance of various models on different tasks including SST-2, MNLI, RTE, QNL	×	0.07
Table [Table (p20)] reports metrics for word-level attacks including Transferability, Fidelity, Human Consensus, Utility	×	0.03

References

- http://arxiv.org/abs/2111.02840v2
- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2212.09155v1