# Integrated Decision Gradients and Attention Rollout Correlation Under Adversarial Attacks in GLUE

Assignee Research

May 31, 2026

**Abstract**

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the correlation between Integrated Decision Gradients and Attention Rollout attribution consistency vary across different adversarial attack types in the Adversarial GLUE benchmark. Deep neural networks (DNN) have achieved unprecedented success in numerous machine learning tasks in various domains. However, the existence of adversarial examples has raised concerns about applying deep learning to safety-critical applications. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Research question: How does the correlation between Integrated Decision Gradients and Attention Rollout attribution consistency vary across different adversarial attack types in the Adversarial GLUE benchmark?.

## 2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

## 3 Results

8 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.8/10.

## 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 work of (Carlini et al., 2017) tries to find the 'provable strongest attack' by finding the theoretical minimally-di | × | 0.05 |
| The ground-truth attack is based on Reluplex (Katz et al., 2017), an algorithm for verifying the properties of neural ne | × | 0.06 |
| The ground-truth attack encodes the model parameters F and data (x, y) as the subjects of a linear-like programming syst | × | 0.07 |
| The work of (Eykholt et al., 2017) crafts physical adversarial objects by contaminating road signs to mislead road sign | × | 0.02 |
| Eykholt's attack uses l1-norm based attacks to find regions to perturb and l2-norm based attacks to generate the color f | × | 0.05 |
| The perturbed stop sign can confuse an autonomous vehicle from any distance and viewpoint. | × | 0.04 |
| The work of (Athalye et al., 2017) reported the first work which successfully crafted physical 3D adversarial objects. | × | 0.05 |
| Athalye's 3D adversarial object uses 3D-printing to manufacture an 'adversarial' turtle. | × | 0.02 |
| The work of (Papernot et al., 2017) was the first to introduce an effective algorithm to attack DNN classifiers under bl | × | 0.05 |
| The ground-truth attack is the first work to seriously calculate the exact robustness (minimal perturbation) of classifi | × | 0.03 |
| The ground-truth attack involves using a SMT solver, which makes it slow and not scalable to large networks. | × | 0.03 |
| Recent works (Tjeng et al., 2017; Xiao et al., 2018c) have improved the efficiency of the ground-truth attack. | × | 0.03 |
| In (Su et al., 2019), it shows that on CIFAR10 dataset, for a well-trained CNN classifier (e.g. VGG16), most of the test | × | 0.03 |
| The one-pixel attack demonstrates the poor robustness of deep learning models. | × | 0.07 |

## References


- http://arxiv.org/abs/1801.04693v1
- http://arxiv.org/abs/2407.13111v1
- http://arxiv.org/abs/1909.08072v2