What is the difference in robustness scores between ViLT models trained with SMOTE versus StyleGAN-augmented d

Assignee Research

June 10, 2026

Abstract

Benefiting from large-scale pre-training, we have witnessed significant performance boost on the popular Visual Question Answering (VQA) task. Despite rapid progress, it remains unclear whether these state-of-the-art (SOTA) models are robust when encountering examples in the wild. To study this, we introduce Adversarial VQA, a new large-scale VQA benchmark, collected iteratively via an adversarial human-and-model-in-the-loop procedure. Through this new benchmark, we discover several interesting findings. (i) Surprisingly, we find that during dataset collection, non-expert annotators can easily

1 Introduction

This paper examines: Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models. Research question: What is the difference in robustness scores between ViLT models trained with SMOTE versus StyleGAN-augmented datasets when evaluated on adversarial VQA benchmarks?.

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 3.8/10.

3 Results

14 papers retrieved. 17 claims extracted; 1 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
Model accuracy on the VQA v2 dataset improved from 50% to 76% since its inception.	×	0.07
Current robust VQA benchmarks are often designed with heuristic rules.	×	0.06
Current robust VQA benchmarks often focus on a single type of robustness.	×	0.08
Current robust VQA benchmarks are based on VQA v2 images or questions which state-of-the-art models are trained on.	×	0.12
Images or questions in previous robust VQA benchmarks are often synthesized rather than provided by humans.	×	0.05
The Adversarial VQA (AVQA) dataset is collected using Human-And-Model-in-the-Loop Enabled Training (HAMLET).	✓	0.15
The AVQA dataset includes web images from Conceptual Captions, user-generated images from Fakeddit, and movie images fro	×	0.04
In the AVQA data collection process, human annotators create examples that current best models cannot answer correctly.	×	0.06
The BUTD model trained on VQA v2 + VGQA achieved an accuracy of 67.60 on the VQA v2 test-dev set.	×	0.05
The BUTD model trained on all data (ALL) achieved an accuracy of 67.52 on the VQA v2 test-dev set.	×	0.04
The UNITER-B model trained on VQA v2 + VGQA achieved an accuracy of 72.70 on the VQA v2 test-dev set.	×	0.06
The UNITER-B model trained on all data (ALL) achieved an accuracy of 72.66 on the VQA v2 test-dev set.	×	0.06
The UNITER-B model trained on all data achieved an accuracy of 24.10 on the AVQA test set.	×	0.04
The ClipBERT model trained on all data achieved an accuracy of 24.35 on the AVQA test set.	×	0.03
The VILLA-B model trained on all data achieved an accuracy of 26.08 on the AVQA test set.	×	0.03
The Sememe+PSO attack method resulted in an error rate of 88.6% with an original accuracy of 84.9 and an adversarial acc	×	0.03
The AVQA collection process involves workers attempting to attack the VQA model for at most 5 times.	×	0.04

References

- http://arxiv.org/abs/2106.00245v2
- http://arxiv.org/abs/2407.04255v1
- http://arxiv.org/abs/1912.02145v1