RankVQA vs. Contrastive Learning in VQA v2 and GQA Benchmark Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the Rank VQA model's ranking-inspired hybrid training strategy compare to contrastive learning methods in terms of VQA accuracy on the VQA v2 and GQA benchmarks. Visual Question Answering (VQA) is a challenging task that requires systems to provide accurate answers to questions based on image content. Current VQA models struggle with complex questions due to limitations in capturing and integrating multimodal information effectively. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Enhancing Visual Question Answering through Ranking-Based Hybrid Training and Multimodal Fusion. Research question: How does the Rank VQA model's ranking-inspired hybrid training strategy compare to contrastive learning methods in terms of VQA accuracy on the VQA v2 and GQA benchmarks?.

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 3.9/10.

3 Results

4 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 3.9/10.

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 RankVQA model was evaluated using the VQA v2.0 and COCO-QA datasets.	×	0.14
The VQA v2.0 dataset contains over 200,000 images and 600,000 questions.	×	0.05
The COCO-QA dataset comprises 123,287 images and over 117,000 questions.	×	0.04
The experimental environment utilized an NVIDIA Tesla V100 GPU with 32GB memory.	×	0.01
The experimental environment utilized an Intel Xeon E5-2698 v4 CPU.	×	0.01
The system memory configuration was 256GB DDR4.	×	0.00
The storage configuration was a 2TB SSD.	×	0.00
The operating system used was Ubuntu 20.04 LTS.	×	0.00
The deep learning framework used was PyTorch 1.10.0.	×	0.01
The CUDA version used was 11.2.	×	0.00
The cuDNN version used was 8.1.	×	0.00
The Python version used was 3.8.10.	×	0.00
During data preprocessing, all images were resized to 224x224 pixels.	×	0.01
During data preprocessing, image pixel values were normalized to the range of 0 to 1.	×	0.04
The RankVQA model architecture includes a visual feature extraction module, a text feature extraction module, a multimod	✓	0.18
The RankVQA network employs the Faster R-CNN model to extract visual features from images.	×	0.11
The RankVQA model utilizes a pre-trained BERT model to extract text features.	×	0.14
The RankVQA model uses a multi-head self-attention mechanism for multimodal fusion.	×	0.15

References

- http://arxiv.org/abs/1911.06352v1
- http://arxiv.org/abs/2408.07303v2

- http://arxiv.org/abs/2305.17369v2