SOVEREIGN: Learning Sparse Mixture of Experts for Visual Question Answering

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

There has been a rapid progress in the task of Visual Question Answering with improved model architectures. Unfortunately, these models are usually computationally intensive due to their sheer size which poses a serious challenge for deployment. We aim to tackle this issue for the specific task of Visual Question Answering (VQA). A Convolutional Neural Network (CNN) is an integral part of the visual processing pipeline of a VQA model (assuming the CNN is trained along with entire VQA model). In this project, we propose an efficient and modular neural architecture for the VQA task with focus on

1 Introduction

Analysis of: Learning Sparse Mixture of Experts for Visual Question Answering. Research goal: What is the scaling efficiency trade-off between dynamic expert specialization and fixed routing in MoE-VLMs when measuring throughput vs. accuracy on visual question answering benchmarks under increasing active parameter counts?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

3 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.7/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
There has been a rapid progress in the task of Visual Question Answering with improved model architectures.	✓	0.39
These models are usually computationally intensive due to their sheer size which poses a serious challenge for deploymen	✓	0.31
A Convolutional Neural Network (CNN) is an integral part of the visual processing pipeline of a VQA model (assuming the	✓	0.48
A sparsely activated CNN based VQA model achieves comparable performance to a standard CNN based VQA model architecture.	✓	0.47

References

- `http://arxiv.org/abs/1803.07724v1`
- `http://arxiv.org/abs/1909.09192v1`
- `http://arxiv.org/abs/1912.02145v1`