SOVEREIGN: To what extent does SMoES improve expert specialization for cross-modal alignment tasks (e.g., visual question

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: To what extent does SMoES improve expert specialization for cross-modal alignment tasks (e.g., visual question answering, captioning) over modality-agnostic routing, as measured by downstream task accuracy and inference throughput on the COCO and VQA v2 benchmarks?.

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

11 papers retrieved. 15 claims extracted, 0 verified. Tribunal: 4.8/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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
Using the Bottom-up attention model for VQA v2 dataset with 0% sparsity to 50% sparsity results in very minimal loss in	×	0.04
With 75% sparsity, there is a marked 3.62% loss in overall accuracy on the VQA v2 dataset.	×	0.03
On the CLEVR dataset, the model using 50% sparsity has comparable performance with the one without sparsity in the convo	×	0.08
The baseline Modular CNN (k=12) achieves 94.05% validation accuracy on the CLEVR dataset.	×	0.05
The Modular CNN with 50% sparsity (k=6) achieves 92.23% validation accuracy on the CLEVR dataset.	×	0.05
The baseline ResNeXt-32 (101 x 32d) achieves 54.51% accuracy on the VQA v2 dataset.	×	0.03
The Modular ResNeXt-32 (101 x 32d) with k=32 (0% sparsity) achieves 54.90% accuracy on the VQA v2 dataset.	×	0.04
The Modular ResNeXt-32 (101 x 32d) with k=16 (50% sparsity) achieves 54.47% accuracy on the VQA v2 dataset.	×	0.03
The Modular ResNeXt-32 (101 x 32d) with k=8 (75% sparsity) achieves 51.28% accuracy on the VQA v2 dataset.	×	0.04
The baseline ResNeXt-32 (101 x 32d) has 156.04E+09 FLOPS.	×	0.01
The Modular ResNeXt-32 (101 x 32d) with k=32 (0% sparsity) has 181.39E+09 FLOPS.	×	0.01
The Modular ResNeXt-32 (101 x 32d) with k=16 (50% sparsity) has 77.72E+09 FLOPS.	×	0.01
The Modular ResNeXt-32 (101 x 32d) with k=8 (75% sparsity) has 45.94E+09 FLOPS.	×	0.01
The baseline Modular CNN (k=12) has 5.37E+07 FLOPS.	×	0.04
The Modular CNN with 50% sparsity (k=6) has 3.21E+07 FLOPS.	×	0.04

References

- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/1909.09192v1