SOVEREIGN: What is the FLOPs efficiency (tokens per FLOP) of SMoE models versus dense models when evaluated on MMMU subse

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: What is the FLOPs efficiency (tokens per FLOP) of SMoE models versus dense models when evaluated on MMMU subsets with domain shifts (e.g., visual-to-textual mismatch), and how does it scale with expert count?.

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: 1.0/10 → 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
The Bottom-up attention model for VQA v2 dataset is used as the base model and the CNN sub-network is replaced with a cu	×	0.09
There is a very minimal loss in accuracy from 0% sparsity to 50% sparsity on the VQA v2 dataset.	×	0.02
With 75% sparsity, there is a marked 3.62% loss in overall accuracy on the VQA v2 dataset.	×	0.02
The Relational Networks model is used for the CLEVR dataset because it is fully-supervised and trains the CNN in the mai	×	0.06
The CNN used for the CLEVR model has four layers with one residual ResNeXt block each followed by a 1×1 convolutional la	×	0.05
On the CLEVR dataset, the model with 50% sparsity has comparable performance to the model without sparsity in the convol	×	0.08
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.	×	0.03
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.	×	0.03
The baseline Modular CNN with k=12 achieves 94.05% validation accuracy on the CLEVR dataset.	×	0.04
The Modular CNN with k=6 (50% sparsity) achieves 92.23% validation accuracy on the CLEVR dataset.	×	0.04
The gating mechanism assigns weights to each of the 32 paths in the ResNeXt-101 (32 × 4d) residual block.	×	0.02
Gate values are normalized to sum to unity and are conditioned on the LSTM based feature representation of the question.	×	0.04
To optimize computation, only the top-k (out of 32) paths are executed and the contribution of others is zeroed out.	×	0.02

References

- http://arxiv.org/abs/2312.04693v3
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/1909.09192v1