SOVEREIGN: What is the throughput (queries per second) trade-off for dense retrievers (e.g., Contriever) on MuSiQue 2-hop

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, we used non-saturat

1 Introduction

Analysis of: ImageNet classification with deep convolutional neural networks. Research goal: What is the throughput (queries per second) trade-off for dense retrievers (e.g., Contriever) on MuSiQue 2-hop vs 3-hop queries under fixed GPU memory, compared to sparse retrievers on CPU?.

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. 6 claims extracted, 6 verified. Tribunal: 6.0/10 → REVISE (revision_round=1). Policy: SOFT_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
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet	✓	0.47
On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively.	✓	0.30
The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of	✓	0.52
To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution opera	✓	0.33
To reduce overfitting in the fully connected layers we employed a recently developed regularization method called 'dropo	✓	0.39
We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 1	✓	0.43

References

- https://doi.org/10.1145/3065386
- https://doi.org/10.4230/lipics.cosit.2022.18
- https://doi.org/10.1186/s40537-021-00444-8