Quantization Strategies and Inference Throughput in Deep Convolutional Networks on ImageNet

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the inference throughput of deep convolutional neural networks on ImageNet vary when applying different quantization strategies while maintaining top-5 error rates below 17.0%. Deep convolutional neural networks (CNNs) are powerful tools for a wide range of vision tasks, but the enormous amount of memory and compute resources required by CNNs pose a challenge in deploying them on constrained devices. Existing compression techniques, while excelling at. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Focused Quantization for Sparse CNNs. Research question: How does the inference throughput of deep convolutional neural networks on ImageNet vary when applying different quantization strategies while maintaining top-5 error rates below 17.0%?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

14 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Focused compression (FC) consists of pruning, Focused Quantization (FQ), and Huffman encoding.	×	0.11
FC was applied to MobileNets and ResNets on the ImageNet dataset.	×	0.05
FC produced models with high compression ratios while permitting a multiplication-free hardware implementation of convol	×	0.08
FC had minimal impact on task accuracy for the tested models.	×	0.04
Models were initially sparsified using Dynamic Network Surgery.	×	0.03
During fine-tuning, Incremental Network Quantization (INQ) was employed to gradually increase the proportion of quantize	×	0.05
Models were fine-tuned for 3 epochs at a learning rate of 0.001 at each quantization step, except for the final 100% ste	×	0.04
The learning rate was decayed every 3 epochs during fine-tuning.	×	0.04
Huffman encoding was applied to model weights to further reduce model sizes.	×	0.09
In the experiments, μ- and μ+ were quantized to the nearest powers-of-two values.	×	0.03
In the experiments, σ- and σ+ were constrained to be equal.	×	0.02
Shift quantization constrains weight values to powers-of-two or zero values.	×	0.05
A representable value in a (k + 2)-bit shift quantization is defined by the formula v = s $\cdot$ 2^(e-b).	×	0.06
In shift quantization, s denotes the sign or zero ({-1, 0, 1}), e is an integer bounded by [0, 2^k-1], and b is a layer-	×	0.02
Shift quantization on sparse layers results in a distribution that is a poor approximation of the original layer weight	×	0.06
Recentralized quantization (Q[θ]) is designed to concentrate quantization effort on high probability regions in the weig	×	0.06
Recentralized quantization is applied in a layer-wise fashion.	×	0.02

References

- http://arxiv.org/abs/2205.04596v2
- http://arxiv.org/abs/2305.05274v2
- http://arxiv.org/abs/1903.03046v3