MQuant and Alternative Inference Optimizations on LLaVA Throughput and Accuracy

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does MQuant's post-training quantization compare to other inference optimization techniques (e.g., pruning, distillation) in terms of throughput and accuracy on the LLaVA benchmark. We consider the problem of model compression for deep neural networks (DNNs) in the challenging one-shot/post-training setting, in which we are given an accurate trained model, and must compress it without any retraining, based only on a small amount of calibration input data.. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning. Research question: How does MQuant's post-training quantization compare to other inference optimization techniques (e.g., pruning, distillation) in terms of throughput and accuracy on the LLaVA benchmark?.

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 2.7/10.

3 Results

14 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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
Calibration datasets consist of 1024 random training samples.	×	0.04
For ImageNet, roughly 0.1% of the training data is used with standard flipping and cropping augmentations to increase th	×	0.05
For ResNet models, batchnorm statistics are reset using 100 batches of 128 samples from the calibration set.	×	0.04
For YOLO model compression, mean and variance correction is applied on a single batch of 128 samples after normalization	×	0.06
For BERT model compression, mean and variance correction is applied on a single batch of 512 samples after normalization	×	0.06
The method uses layer-wise calibration losses without augmentations for all models with exactly one layer compressed to	×	0.03
ResNet50 achieves 76.13 accuracy in the dense model setting.	×	0.04
ResNet50 achieves 74.75 accuracy with 4:8 pruning in ExactOBS method.	×	0.04
BERT3 model achieves 84.66 accuracy in the dense model setting.	×	0.04
BERT3 model achieves 82.75 accuracy with AdaPrune method.	×	0.02
ResNet18 achieves 69.76 accuracy in the dense model setting.	×	0.03
ResNet50 achieves 76.13 accuracy in the dense model setting.	×	0.04
AdaQuant method achieves 75.84, 75.14, 71.58 accuracies for 4bit, 3bit, 2bit quantization levels respectively.	×	0.02
OBQ method achieves 69.56, 68.69, 64.04 accuracies for ResNet18 at 4bit, 3bit, 2bit quantization levels respectively.	×	0.02

References

- `http://arxiv.org/abs/2401.10484v1`

- http://arxiv.org/abs/2208.11580v2
- http://arxiv.org/abs/2406.16299v1