Quantization-Aware Training Effects on Latency-Throughput Trade-offs in FPGA-Deployed Transformers

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does quantization-aware training impact the latency-throughput trade-off of hls4ml-deployed transformer models on FPGA accelerators compared to post-training quantization. 9 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Fast machine learning for building management systems. Research question: How does quantization-aware training impact the latency-throughput trade-off of hls4ml-deployed transformer models on FPGA accelerators compared to post-training quantization?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

10 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 7.7/10.

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
Building management systems (BMSs) are increasingly integrating advanced machine learning (ML) and artificial intelligen	✓	0.39
Existing BMSs can only provide local adaptability by creating and managing information for a built asset.	✓	0.25
Existing BMSs lack the capability to learn and adapt based on performance objectives.	✓	0.19
The study reviews optimization methods for ML algorithms, focusing on Long Short-Term Memory (LSTM) networks for energy	✓	0.33
The study explores solutions that leverage hardware accelerators for low-latency and high-throughput processing.	✓	0.21
The High-Level Synthesis for Machine Learning (HLS4ML) framework facilitates deployment of fast machine learning models	✓	0.34
The use of the HLS4ML framework achieves substantial gains in hardware efficiency and inference speed in resource-constr	✓	0.21
HLS4ML-optimized models maintain accuracy while offering computational efficiency through techniques like pruning and qu	✓	0.27
HLS4ML-optimized models support real-time BMS applications.	✓	0.15

References

- https://doi.org/10.21275/sr25409073105
- https://doi.org/10.1145/3613963
- https://doi.org/10.1007/s10462-025-11226-6