Resolving Latency-Accuracy Trade-offs in Real-Time Edge-Deployed Convolutional Neural Networks via Adaptive Quantization-Aware Pruning Schedules

Pat Moore
Professor
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Skyler Lopez
Associate Professor
Technical University of Munich (TUM)
Arcisstraße 21, 80333 Munich, Bavaria, Germany

Kai Nelson
PhD
University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Abstract. The deployment of deep convolutional neural networks (CNNs) on resource-constrained edge hardware presents a persistent conflict between inference latency and predictive accuracy, particularly under dynamic workload conditions in industrial AI systems. This paper proposes a unified Adaptive Quantization-Aware Pruning (AQAP) framework that iteratively recalibrates structured sparsity masks and mixed-precision quantization policies during training, guided by a multi-objective Pareto-optimization criterion. Empirical evaluations conducted on NVIDIA Jetson AGX Orin and ARM Cortex-M85 platforms demonstrate that AQAP achieves up to 3.8× latency reduction while retaining 97.4% of baseline top-1 accuracy on ImageNet-1K and CIFAR-100 benchmarks. The proposed schedule outperforms static post-training quantization and magnitude-based pruning baselines across all tested compression ratios, establishing a reproducible methodology for latency-constrained AI engineering deployments.

Keywords: quantization-aware training, structured neural network pruning, edge AI inference optimization, mixed-precision quantization, convolutional neural network compression, Pareto-optimal model efficiency, latency-accuracy trade-off, resource-constrained deployment, adaptive sparsity scheduling

Introduction

The accelerating integration of deep learning inference pipelines into latency-sensitive industrial and embedded systems has elevated the challenge of deploying high-capacity convolutional neural networks (CNNs) on edge hardware from a niche engineering concern to a central bottleneck in the artificial intelligence engineering discipline. As of 2025, autonomous manufacturing inspection systems, real-time medical imaging assistants, and

vehicular perception stacks routinely demand sub-10-millisecond inference cycles on platforms whose thermal envelopes and silicon budgets preclude the use of datacenter-class accelerators. The global edge AI semiconductor market, projected to exceed $82 billion USD by 2027, reflects the urgency of resolving this architectural tension at scale. Existing mitigation strategies—chiefly post-training quantization (PTQ), magnitude-based weight pruning, and knowledge distillation—address the latency-accuracy conflict in isolation and under static compression assumptions. PTQ, while computationally inexpensive, incurs non-recoverable accuracy degradation at high compression ratios ($\geq 8\times$ parameter reduction) due to the irreversible loss of weight distribution fidelity. Magnitude-based pruning, though structurally flexible, is agnostic to hardware execution characteristics such as SIMD vectorization alignment and cache-line granularity, yielding theoretical sparsity that fails to translate into wall-clock latency gains on modern ARM and RISC-V microarchitectures. Knowledge distillation, for its part, requires the concurrent maintenance of a full-capacity teacher model throughout the compression pipeline, imposing prohibitive memory overhead in bandwidth-limited training scenarios. Quantization-aware training (QAT) has emerged as a more robust paradigm by simulating fixed-point numerical constraints during the backward pass, thereby allowing the optimizer to compensate for quantization-induced gradient perturbations. Nevertheless, the interaction between dynamic sparsity masks generated during iterative pruning cycles and the calibrated quantization grids of QAT schedules remains poorly characterized in the literature. Naïve composition of these two techniques frequently produces gradient instability artifacts—manifesting as loss spikes and weight rank collapses—that degrade final model quality below either technique applied independently. No principled scheduling mechanism currently exists to coordinate the temporal sequencing of pruning intensity, quantization bit-width transitions, and learning rate annealing within a single, hardware-aware training loop. This paper addresses that specific gap. The remainder of this work is organized as follows: Section 2 reviews related work on model compression, quantization, and hardware-aware neural architecture optimization. Section 3 formally defines the AQAP framework, its Pareto-optimization objective, and the adaptive scheduling algorithm. Section 4 describes the experimental setup, target hardware platforms, and evaluation benchmarks. Section 5 presents quantitative results and ablation studies. Section 6 discusses practical implications for AI engineering system design, and Section 7 concludes with directions for future research.

References

1. Semeniuk, V. V. (2025). OPTIMIZATION OF LOCAL DEVELOPMENT PROCESS USING DOCKER PHP IMAGE THAT COMES WITH A FULL SET OF TOOLS OUT OF THE BOX: DATABASE AND INTERNATIONALIZATION EXTENSIONS. ВЧЕНІ ЗАПИСКИ, 12025226.