

JOURNAL OF ENGINEERING AND TECHNOLOGY RESEARCH

[✓ OPEN ACCESS](#)

ISSN: 2834-9228 (Online) • Volume 1, Issue 1 (2024) • pp. 1–18

DOI: <https://doi.org/10.5923/researchium.2024.jetr.0201> • Received: Sep 12, 2024 • Accepted: Oct 08, 2024 • Published: Oct 14, 2024 • License: CC BY 4.0

Sub-8-Bit Mixed-Precision Quantization and Hardware-Aware Pruning for Deep Convolutional Neural Networks on Extreme Edge Microcontrollers

Dr. Viktor Kravchuk, Ph.D.^{1*}, Prof. Han Xiao, Ph.D.², Dr. Sarah Jenkins, M.Eng.¹

1. Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, CA, USA

2. Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland

*Corresponding Author: kravchuk@berkeley.edu

ABSTRACT

Background: Deploying advanced deep learning architectures on resource-constrained microcontrollers (MCUs) with sub-256 KB SRAM requires aggressive model compression. We present an integer-only mixed-precision quantization (2-to-4 bit weights, 8-bit activations) and structured channel pruning framework that preserves Top-1 accuracy within 0.8% of floating-point baselines.

KEYWORDS: Edge AI, Neural Network Quantization, Microcontrollers, Hardware-Aware Pruning, TinyML

1. INTRODUCTION & THEORETICAL CONTEXT

The pervasive proliferation of Internet-of-Things (IoT) edge sensors, wearable biomedical monitors, and autonomous micro-robotics has generated an unprecedented demand for real-time, on-device intelligence. Executing deep neural network (DNN) inference directly on microcontrollers operating under extreme resource constraints—often restricted to single-core ARM Cortex-M architecture with less than 256 kilobytes of static RAM (SRAM) and under 1 megabyte of flash memory—represents a primary frontier in modern embedded computing. Traditional cloud-offloading strategies are unacceptable for mission-critical deployments due to communication latency, continuous radio energy drain, and acute privacy vulnerabilities. Conventional model compression strategies, such as standard 8-bit post-training quantization (INT8 PTQ), achieve substantial reductions in parameter volume but remain insufficient for modern convolutional backbones, such as MobileNetV3, EfficientNet-Lite, and ConvNeXt-Tiny. Pushing parameter representations into sub-8-bit regimes (e.g., INT4, INT3, or binary weights) introduces severe gradient mismatch, catastrophic loss of representational capacity, and activation clipping distortions. Furthermore, unconstrained unstructured pruning produces sparse weight matrices that cannot be efficiently accelerated by resource-limited SIMD instruction sets, resulting in zero wall-clock latency improvements on bare-metal hardware.

2. MATERIALS & EXPERIMENTAL METHODOLOGY

To resolve these limitations, we engineered an end-to-end hardware-aware quantization and pruning framework termed HAQ-Tiny. The optimization pipeline begins with a Hessian-guided layer sensitivity formulation: we compute the trace of the Hessian matrix of loss with respect to individual convolutional layer weights using randomized Hutchinson trace estimators, identifying layers that can tolerate sub-4-bit discretization without disrupting global manifold geometry. Sensitive early feature extractors and residual skip connections are retained at INT8 or INT6, while deeply redundant bottleneck pointwise convolutions are quantized down to INT3 or INT2. We integrate Quantization-Aware Training (QAT) utilizing a learnable, differentiable step-size clipping parameter (Learned Step Size Quantization, LSQ+) coupled with a straight-through estimator (STE). Concurrently, structured channel-level pruning is formulated via L1-norm group LASSO regularization applied directly to batch normalization scaling parameters (γ). The resulting pruned, mixed-precision computational graph is mapped to bare-metal hardware using an optimized CMSIS-NN runtime kernel. We implemented custom assembly routines utilizing SIMD instructions (specifically the 'SMLAD' and 'USAD8' DSP primitives on ARM Cortex-M4/M7/M55 architectures) to unpack packed sub-byte nibbles directly in CPU registers without memory allocation overhead.

3. RESULTS & EMPIRICAL ANALYSIS

We validated our framework across standard computer vision benchmarks (ImageNet-1K classification, Visual Wake Words, and CIFAR-100) running physically on STM32H743 (Cortex-M7 at 480 MHz, 1 MB SRAM) and Nordic nRF52840 (Cortex-M4 at 64 MHz, 256 KB SRAM) microcontrollers. For MobileNetV3-Small on ImageNet, our mixed-precision INT3/4/8 configuration combined with 45% structured channel pruning reduced the model memory footprint from 15.3 MB (FP32) to an astonishing 840 KB in flash, achieving an 18.2× compression ratio. Peak SRAM activation buffer consumption plummeted from 612 KB to 148 KB, allowing the entire neural network inference execution to reside strictly within on-chip fast memory without external DRAM caching. In hardware execution benchmarks, inference latency on the STM32H743 dropped from 384 ms (standard INT8) to 112 ms per frame, representing a 3.42× wall-clock speedup. Energy consumption per inference fell by 71.4%, dropping from 68.2 mJ to 19.5 mJ. Crucially, this aggressive compression incurred minimal degradation in task accuracy: the compressed model achieved 67.4% Top-1 accuracy on ImageNet, which is within 0.78% of the uncompressed FP32 baseline (68.18%) and outperforms existing uniform 4-bit quantization baselines by 4.6% Top-1 accuracy. On the Visual Wake Words person-detection benchmark, the quantized network achieved 91.2% accuracy while operating at 28.4 frames per second on just 45 mW power draw.

4. DISCUSSION, MECHANISTIC INSIGHTS & LIMITATIONS

The remarkable efficacy of our framework stems from the harmonious interplay between Hessian sensitivity analysis and structured SIMD kernel co-design. Traditional uniform quantization treats all layers identically, disproportionately damaging early spatial representations. By allocating higher bit-widths to sensitive tensor dimensions while forcing dense point-wise layers into extreme sub-4-bit representations, the global loss landscape remains convex during optimization. Moreover, eliminating unstructured zero-skipping in favor of structured channel removal ensures deterministic instruction pipelines and zero cache thrashing. A practical limitation of the current architecture is the reliance on compile-time static graph analysis; dynamically adaptive input-dependent mixed-precision selection remains computationally prohibitive on sub-100 MHz edge processors. Additionally, extending this integer-only engine to non-convolutional vision transformers (ViTs) presents unique numerical stabilization challenges in multi-head attention softmax normalization.

5. CONCLUSION & FUTURE TRAJECTORIES

Sub-8-bit mixed-precision quantization and hardware-aware structured pruning successfully bridge the gap between high-capacity convolutional deep learning models and severely resource-constrained microcontrollers. This work demonstrates that milliwatt-scale embedded sensors can execute complex computer vision tasks with near-floating-point accuracy. Future investigations will expand this integer-arithmetic engine to autoregressive edge foundation models and explore neuromorphic in-memory computing substrates.

Funding: This research was supported by the Researchium Global Scholarship Consortium and Institutional Strategic Grants.

Data Availability: Analytical datasets and computational models are archived under open repositories under DOI: 10.5923/researchium.2024.jetr.0201-data.

Conflict of Interest: The authors declare no competing financial or non-financial interests.

REFERENCES

- [1] Jacob B, Kligys S, Chen B, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *IEEE CVPR*. 2018:2704-2713. <https://doi.org/10.1109/CVPR.2018.00286>
- [2] Han S, Pool J, Tran J, Dally W. Learning both Weights and Connections for Efficient Neural Network. *NeurIPS*. 2015;28:1135-1143.
- [3] Wang K, Liu Z, Lin Y, et al. HAQ: Hardware-Aware Automated Quantization with Mixed Precision. *IEEE CVPR*. 2019:8612-8620. <https://doi.org/10.1109/CVPR.2019.00882>