NeurArch 2026 — International Symposium on Neural Computing Architectures

Adaptive Quantization-Aware Neural Architectures for Low-Power Edge Inference in Industrial IoT Systems

Elena R. Vasquez [1,*] Jonas M. Breitfeld [1,2] Priya S. Nambiar [2] Tobias F. Engel [3] Li-Wei Chen [1]

[1]Institute for Embedded Intelligence, ETH Zurich, Zurich, Switzerland
[2]Siemens AG, Corporate Technology, Munich, Germany
[3]Department of Electrical Engineering, KU Leuven, Leuven, Belgium

e.vasquez@ei.ethz.ch *(corresponding author)*

Background. Neural networks deployed on resource-constrained edge devices in industrial environments must satisfy strict power budgets (≤ 5 mW) while sustaining inference latency under 10 ms per sample. Conventional post-training quantization degrades accuracy by up to 8% on domain-specific time-series data, motivating architecture-level solutions.

Methods. We propose *AdaptiQ-Edge*, a hardware-aware neural architecture search (NAS) framework that co-optimises bit-width allocation and operator selection for mixed-precision inference on ARM Cortex-M55 and RISC-V-based MCUs. The search space spans 4-/8-/16-bit integer quantization per layer, depthwise-separable convolutions, and lightweight attention blocks. Training uses a differentiable proxy loss combining cross-entropy with a latency–energy Pareto term calibrated via device profiling. Experiments were conducted on three industrial sensor benchmarks: *SmartFactory-TS*, *PredMaint-12k*, and the public CWRU bearing-fault dataset.

Results. AdaptiQ-Edge achieves a top-1 accuracy of **94.7%** on SmartFactory-TS, surpassing full-precision MobileNetV3-Small by $+1.3$ pp while reducing energy per inference by **68%** (from 18.4 μJ to 5.9 μJ). On CWRU, F1-score reaches 0.983 at an average latency of 6.8 ms—within budget— compared to 0.961 for a quantization-unaware baseline. Model size is reduced by $4.2\times$ (from 1.8 MB to 0.43 MB).

Conclusions. Joint architecture search and quantization-aware training yields models that are simultaneously more accurate, smaller, and faster than post-training quantization approaches. AdaptiQ-Edge is publicly available, and our device-profiling API generalises to heterogeneous MCU families, enabling reproducible edge-AI deployment across industrial IoT sectors.

Keywords: neural architecture search; quantization-aware training; edge inference; industrial IoT; mixed-precision; energy efficiency

Submission Details

Track	Hardware-Efficient AI & Embedded Systems
Presentation Type	Oral (20 min + 5 min Q&A)
Topic Area	Neural Architecture Search / Model Compression
Submission ID	NeurArch2026-0417
Conflict of Interest	The authors declare no competing financial interests.