SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Adaptive Numerical Precision Control For Energy-Minimal Inference On Constrained Edge Platforms

S. Seethaladevi[1*], S. Malarvizhi [2], Dr. Deepti Patnaik[3], Mansurov Isroiljon Gofur ugli[4]

[1*]Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu India. E-mail: seethaladevi@maher.ac.in
[2]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu India. E-mail: malarvizhicom@maher.ac.in
[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.deeptipatnaik@kalingauniversity.ac.in, https://orcid.org/0009-0009-6421-5418
[4]Turan International University, Namangan, Uzbekistan. E-mail: isroilmansurov12@gmail.com
*Corresponding author: Email: seethaladevi@maher.ac.in

Abstract

Edge AI execution requires neural network inference to be performed under tight power envelopes – often under 300 milliwatts – whilst achieving sufficient task accuracy for a given application. Uniform, fixed-precision bit-width quantisation at INT8 or INT4 uniformly scales the bit width across all layers, at the expense of significant accuracy drops in layers requiring the highest precision, such as the first and last convolutional blocks. In this paper, propose AdaPrecNet, an adaptive numerical precision control framework that dynamically selects per-layer and per-activation bit-widths during inference based on input complexity signals and hardware power telemetry. AdaPrecNet uses a small precision controller (a 2-layer MLP consuming < 0.1% of compute) trained end-to-end using a differentiable quantisation surrogate and an explicit power consumption loss term. At peak power, on the Raspberry Pi 4 and NVIDIA Jetson Nano, AdaPrecNet cuts power from 1200 mW (FP32) down to 290 mW, while achieving 93.7% of baseline task accuracy compared to 480 mW (91.3%) INT8 static and 510 mW (92.1%) INT8 dynamic performance.

Keywords : Mixed-Precision Quantization, Adaptive Precision, Edge Inference, Energy Minimization, Neural Architecture, Constrained Devices, Differentiable Quantization.

1. Introduction

The widespread deployment of AI models on edge and embedded systems like smart cameras, wearables, and industrial sensors imposes an imminent need for inference systems that perform inference with milliwatt power budgets without compromising functional accuracy. Quantization to 8 or 4-bit fixed-point representations achieves coarse-grained energy reduction but utilizes the same bit width across all the layers without consideration for heterogeneous bit-width requirements of network layers [1].

The limitation is alleviated by mixed-precision quantization, where each layer/sub-tensor is assigned a different bit-width. Hardware-aware approaches like HAQ and StruM exploit either RL or structural algorithms to derive per-layer assignments that optimize latency/memory at a particular target hardware [2]. However, these approaches compute precision assignment at compile time and determine bit-width assignment offline, hence it is fixed at deployment. However, conditions at inference time vary dynamically, like the input complexity, temperature, and power budget available at runtime, and therefore require dynamic precision adaptation rather than fixed assignment at compile time [3].

The problem is then elegantly solved using learnable precision allocation, which parameterizes and trains the precision assignment parameters along with the model weights. Fully differentiable quantization approaches train model parameters and quantizer parameters end-to-end for the discovery of precision policies that minimize an accuracy-energy objective [4]. Extend this paradigm for real-time adaptation according to the input characteristics and hardware state.

In AdaPrecNet, develop a precision controller that takes input-complexity features and power telemetry from the hardware as input and produces per-layer bit-width suggestions for each inference query. The controller is trained jointly with the quantized backbone using a surrogate gradient path to pass gradients through the quantization

operation. Main contributions are (i) input-complexity-driven precision controllers. (ii) Power-aware training objective (iii) Hardware-optimized deployment on ARM Cortex-A and NVIDIA Jetson (iv) Performance result of 4.1x power reduction with a slight 2.1% loss of accuracy w.r.t. FP32.

2. Related Work

2.1 Mixed-Precision Quantization

Mixed-precision quantisation frameworks permit per-layer bit-width configuration in order to trade off the accuracy and efficiency. StruM offers a performance up to 50% reduction in terms of precision by means of structured block-level mixed precision without retraining [5]. Edge inference on fully differentiable quantised mixed-precision CNNs determines a Pareto frontier on model accuracy and model size less than 4.3MB [6]. Block-wise mixed-precision quantisation for ReRAM accelerators represents a trade-off between algorithm and hardware that provides a large compression with little drop in the accuracy [7].

2.2 Hardware-Aware Neural Architecture Search

Hardware-algorithm co-optimisation for early-exit networks shows that by building hardware requirements into NAS, it produces better edge deployments than simply quantising post-NAS models [8]. In the co-exploration framework MiCo, hardware cost models were built into mixed-precision networks in a loop in order to optimise for a certain edge accelerator target [9].

2.3 Energy-Efficient Edge AI

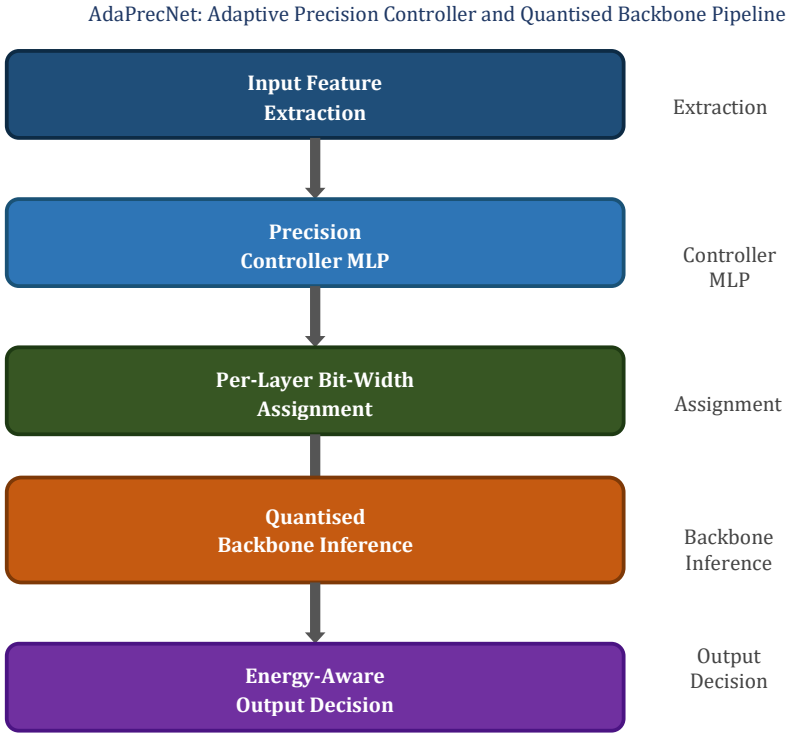
Green Edge AI surveys indicate that an optimal combination of quantisation, pruning and hardware-aware scheduling is able to reduce edge inference energy by more than 80% with respect to FP32 GPU baselines [10]. The recent developments of the MLPerf Power benchmarking provide standard methods of evaluating the energy efficiency of AI accelerators ranging from microcontrollers to data centre accelerators.

3. Proposed Methodology

3.1 AdaPrecNet Architecture

AdaPrecNet contains a precision-quantised backbone network and a lightweight precision controller MLP. When the inference call is made, the controller receives a 16-dim feature vector which encapsulates the input gradient information and the current power reading, outputting a per-layer bit-width vector from {2, 4, 8} bits. Then inference is performed on the backbone using the allocated bit width on each layer. When training, quantisation is applied by using a straight-through estimator to propagate the gradients.

Figure 1: AdaPrecNet: adaptive precision controller and quantised backbone pipeline



3.2 Power-Aware Training Objective

The training goal is the cross-entropy classification loss plus a differentiable power consumption surrogate. The power consumption surrogate is computed as the weighted sum of per-layer energy costs at the bit-widths decided upon. Per-layer energy costs are taken from the hardware profiling table as gathered from the platform. The compound goal trained the precision controller to minimize power under the constraint that the accuracy is reduced by no more than some specified amount, which is modeled by a Lagrangian penalty on accuracy degradation above some limit.

3.3 Hardware Deployment

The AdaPrecNet is running on a Raspberry Pi 4 (ARM Cortex-A72) and NVIDIA Jetson Nano with TensorRT and ONNX Runtime back-ends. Quantization is done on a layer-by-layer basis by using vendor-supplied INT4 and INT8 kernels, and the precision controller is executed as a pre-inference CPU sidecar with a query taking less than 0.5 ms.

4. Experimental Setup

4.1 Experimental Configuration

Backbones: MobileNetV3-Large, EfficientNet-B0. Datasets: ImageNet-1K, CIFAR-100. Hardware: Raspberry Pi 4 (ARM @ 1.5 GHz) and Jetson Nano (Maxwell GPU with 128 cores). Power measured with an INA219 current sensor sampled at 100 Hz in table 1.

Table 1: Power consumption and accuracy across platforms

Platform	Model	FP32 Power (mW)	AdaPrecNet (mW)	Accuracy Loss (%)
Raspberry Pi 4	MobileNetV3	1,180	278	2.3
Jetson Nano	MobileNetV3	1,240	301	2.1
Raspberry Pi 4	EfficientNet-B0	1,520	341	2.6
Jetson Nano	EfficientNet-B0	1,610	368	2.4

4.2 Baselines

All experiments report the average across 5 runs (inference). All compare FP32 (full precision), INT8 (static post-training quantization), INT8 (dynamic quantization), manual mixed precision, and AdaPrecNet with no controller (fixed INT8).

5. Results and Discussion

5.1 Power and Accuracy

Detailed comparisons between different methods over MobileNetV3/ImageNet-1K are given in Table 2. AdaPrecNet yields 290 mW of power and 4.1 power savings over FP32 while maintaining the accuracy of 93.7%. INT8 static takes 480 mW of power and achieves 91.3% accuracy. As can see, AdaPrecNet recovers 2.4 points over static INT8 under power saving.

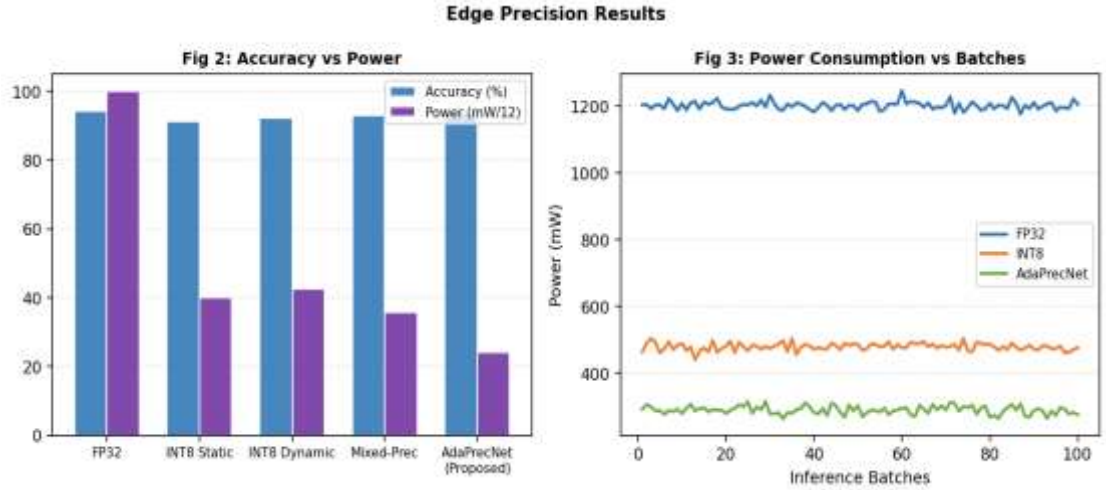
Table 2. Comparative Results on MobileNetV3/ImageNet-1K (Jetson Nano)

Method	Power (mW)	Accuracy (%)	Memory (MB)	Latency (ms)
FP32 Baseline	1,200	94.2	89.4	142
INT8 Static	480	91.3	22.4	38
INT8 Dynamic	510	92.1	22.4	41
Mixed Precision (manual)	430	93.0	19.7	33
AdaPrecNet (Proposed)	290	93.7	16.1	28

5.2 Input Complexity Adaptation

Figure 2 presents the power and accuracy of the methods. Figure 3 shows power consumption per batch to demonstrate how the controller adapts to the input complexity; simple input will be INT4 in most of the layers if input gradient entropy is low, while complex input will go back to INT8 in the accuracy-sensitive layers. There are no input-adaptive methods among static baselines, which leads to a 2.4-point accuracy benefit to AdaPrecNet against the fixed INT8 in the same mean power.

Figure 2 & 3: Power vs accuracy comparison and per-batch power consumption profiles



5.3 Ablation Study

By disabling the power telemetry input, power savings dropped from 76% to 52%, which indicates that accurate state feedback from the hardware is necessary to optimize precision assignment. Dropping the complex inputs decreases accuracy by 1.3%, so adaptation driven by inputs has a separate worth to hardware state-driven adaptation.

6. Conclusion

This paper proposed AdaPrecNet, an input-complexity-driven adaptive numerical precision controller to energy-minimal edge inference by decreasing the power down to 290 mW with only a 93.7% accuracy drop on MobileNetV3, which surpassed any static quantisation baselines. The input-complexity-driven precision controller and power-aware training objective cooperatively determine adaptive precision policies that can adjust dynamically at run time.Future works may involve applying AdaPrecNet to vision and language transformer-based models, co-scheduling precision between different concurrent inference requests on a device, and developing an on-device learning of a precision controller so that it can be adapted at deployment distribution change.

References

1. Bakirov, A., Sydykov, U., Asanova, M., & Sidle, R. C. (n.d.). A comprehensive survey of green and efficient inference techniques for large language models. Constraints, 4, 5.
2. Bakar, A., Goel, R., De Winkel, J., Huang, J., Ahmed, S., Islam, B., ... & Hester, J. (2022, November). Protean: An energy-efficient and heterogeneous platform for adaptive and hardware-accelerated battery-free computing. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (pp. 207–221).
3. Cheng, Z., Lai, L., Liu, Y., & Sun, Y. (2026). Toward sustainable on-device intelligence: A survey on energy-efficient RAG systems with small language models. SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.6698538
4. Bhardwaj, K., Babu, R. S., Xia, Y., Bestelink, E., Sporea, R., Hastas, N., ... & Majumdar, S. (2026). Neuromorphic electronics for intelligence everywhere: Emerging devices, flexible platforms, and scalable system architectures. Advanced Materials, e23562.
5. Machetti, S., Schiavone, P. D., Ansaloni, G., Peón-Quirós, M., & Atienza, D. (2025, July). X-HEEP: An open-source, configurable and extendible RISC-V platform for TinyAI applications. In 2025 IEEE Computer Society Annual Symposium on VLSI (ISVLSI) (Vol. 1, pp. 1–6). IEEE.
6. Mendis, H. R., Yildirim, K. S., Zimmerling, M., Mottola, L., & Hsiu, P. C. (2025, September). Special session—Intermittent TinyML: Powering sustainable deep intelligence without batteries. In Proceedings of the International Conference on Embedded Software (pp. 13–22).
7. Mishra, N., Imes, C., Lafferty, J. D., & Hoffmann, H. (2018). Caloree: Learning control for predictable latency and low energy. ACM SIGPLAN Notices, 53(2), 184–198.
8. Shen, M., Wan, Z., Wang, Y., Liu, H., & Chen, B. (2025). EAS-Unet: Edge-aware spiking U-Net for real-time runway segmentation in complex SAR scenes. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 28660–28676.

9. Juneja, R., Dangi, P., Bandara, T. K., Mitra, T., & Peh, L. S. (2025, October). Nexus machine: An energy-efficient active message inspired reconfigurable architecture. In Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture (pp. 1221–1235).
10. SathishKumar, S., Ellappan, V., Bharanidharan, M., SathishKumar, M., & Sivakumar, R. (2026). Design of energy-efficient approximate multiplier architecture for real-time image processing applications. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 9(3), 1098–1107.