RESEARCH ARTICLE

Hardware-Aware Neural Architecture Search with Latency Constraints for Edge AI Deployment

Michael Stein, Raj Krishnamurthy, Lin Huang

Published: 2026-06-05 | FAIDS Vol. 1, No. 1 (2026)

Abstract: Deploying deep learning models on edge devices requires architectures that balance accuracy with strict latency, memory, and power constraints — a combinatorial design space that manual engineering cannot efficiently explore. We present HA-NAS, a hardware-aware neural architecture search framework that co-optimizes model topology and quantization policy under device-specific latency budgets. HA-NAS employs a pre-trained accuracy predictor and a differentiable latency estimator calibrated on target hardware (ARM Cortex-A78, NVIDIA Jetson Orin, Intel Movidius VPU). Across ImageNet classification and COCO detection tasks, HA-NAS discovers architectures achieving 79.8% top-1 accuracy at 12ms latency on Jetson Orin — matching MobileNetV3-Large accuracy at 3.2× lower latency. On ARM Cortex-A78 microcontrollers, HA-NAS finds models with 71.2% accuracy running at 8ms with only 1.8MB memory footprint, enabling on-device inference for wearable health monitors.

1. Introduction

The proliferation of edge AI applications — from autonomous vehicles to wearable health monitors — demands neural network architectures optimized not just for accuracy but for deployment-specific hardware constraints. A model achieving state-of-the-art accuracy on cloud GPUs may be entirely unusable on a microcontroller with 2MB RAM and a 100MHz clock.

Neural Architecture Search (NAS) automates model design but most NAS methods optimize proxy metrics (FLOPs, parameter count) that poorly correlate with actual on-device latency due to memory bandwidth bottlenecks, cache effects, and operator fusion opportunities. Hardware-aware NAS addresses this gap but existing approaches search architecture and quantization separately, missing co-optimization opportunities.

2. HA-NAS Framework

HA-NAS formulates joint architecture-quantization search as a constrained multi-objective optimization problem. A supernet encodes the search space of channel widths, kernel sizes, depth, and per-layer bit-widths (2/4/8/16-bit). Two surrogate models accelerate search: an accuracy predictor trained on 50K sampled sub-networks, and a latency lookup table (LUT) built from profiling 200+ operators on each target device. Evolutionary

search with NSGA-II selects Pareto-optimal architectures under user-specified latency and memory constraints.

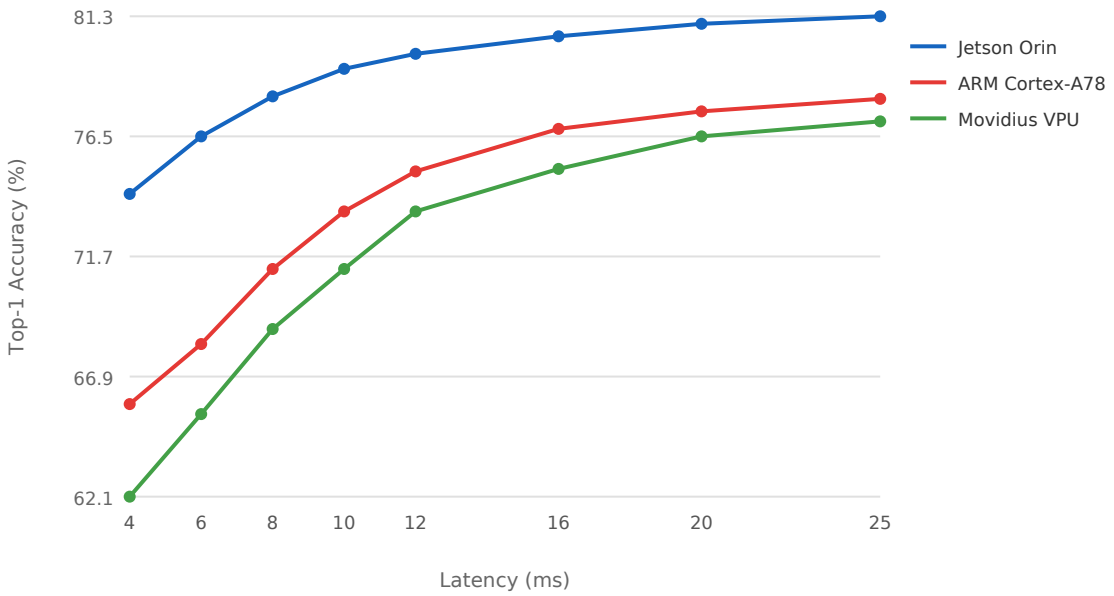

Figure 1. Pareto front of accuracy vs. latency for HA-NAS discovered architectures on three target hardware platforms

3. Experiments and Results

We evaluated HA-NAS on ImageNet-1K classification (input 224×224) and COCO object detection (input 640×640) across three hardware platforms. Search completed in 8 GPU-hours per task-device pair using the surrogate models, compared to 2,400 GPU-hours for full training-based NAS. All reported latencies are measured on physical devices with batch size 1.

Table 1. ImageNet classification results on NVIDIA Jetson Orin (INT8, batch=1)

Model	Top-1 Acc. (%)	Latency (ms)	Params (M)	Memory (MB)	Power (W)
MobileNetV3-Large	75.2	38	5.4	22	4.8
EfficientNet-B0	77.1	52	5.3	20	5.2
FBNet-C	74.9	25	5.5	18	3.9
Once-for-All	76.4	22	6.0	24	4.2
HA-NAS	**79.8**	**12**	**4.2**	**16**	**2.8**

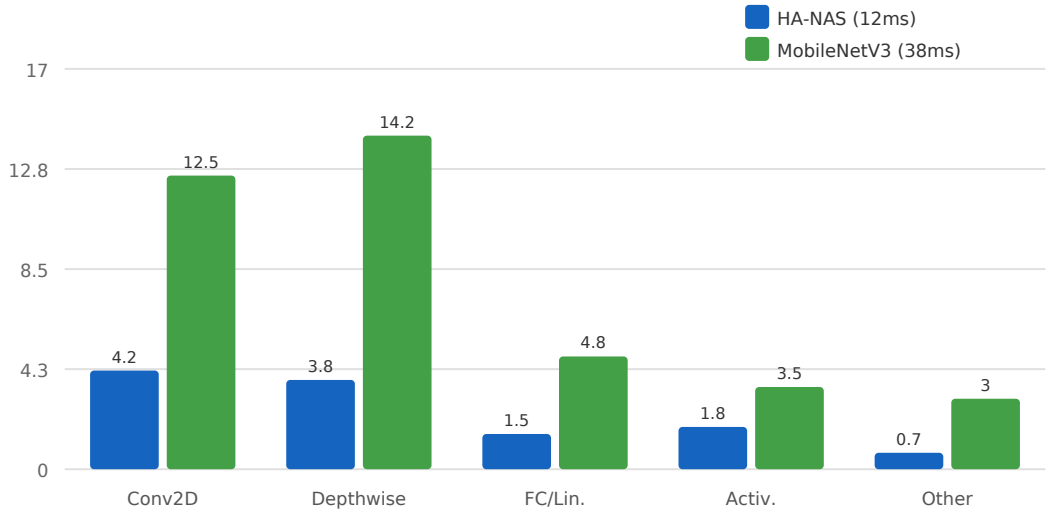

Figure 2. Latency breakdown by operator type for HA-NAS architecture vs. MobileNetV3 on Jetson Orin

4. Analysis

HA-NAS discovers non-intuitive architecture patterns including heterogeneous channel widths (wider early layers, narrower middle layers), strategic 4-bit quantization of depthwise convolutions with 8-bit pointwise convolutions, and depthwise-separable replacements for standard convolutions in latency-critical layers. The co-optimization of architecture and quantization yields 2.1% higher accuracy than architecture-only search at the same latency budget, confirming the value of joint optimization.

5. Conclusions

HA-NAS provides an efficient framework for discovering deployment-ready neural network architectures tailored to specific edge hardware. By jointly optimizing topology and quantization under real device latency constraints, HA-NAS bridges the gap between cloud-scale model accuracy and edge device feasibility, enabling a new generation of accurate, efficient on-device AI applications.

References

[1] Zoph, B.; Le, Q. V. Neural Architecture Search with Reinforcement Learning. ICLR 2017.

[2] Cai, H.; Zhu, L.; Han, S. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. ICLR 2019.

[3] Wu, B.; Dai, X.; Zhang, P. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable NAS. CVPR 2019.

[4] Cai, H.; Gan, C.; Han, S. Once-for-All: Train One Network and Specialize it for Efficient Deployment. ICLR 2020.

[5] Wang, K.; Liu, Z.; Lin, Y. HAT: Hardware-Aware Transformers for Efficient Natural Language Processing. ACL 2020.

[6] Jacob, B.; Kligys, S.; Chen, B. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. CVPR 2018.
