Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

PrototypeNAS: Rapid Design of Deep Neural Networks for Microcontroller Units

ECML PKDD 2026

Mark Deutel, Simon Geis, Axel Plinge | Fraunhofer IIS

Motivation
Rapid Design of Deep Neural Networks (DNNs) for Microcontroller Units (MCUs)

Deep Neural Network Architectures [4]

Metrics	AlexNet	VGG16	ResNet50
Layers	8	16	54
Weights	61M	138M	25.5M
MACs	724M	15.5G	3.9G

Example Microcontroller Units

Metrics	Raspberry Pi Pico	Arduino Nano 33 BLE Sense
Processor	133MHz	64MHz
Flash	2MB	1MB
SRAM	256KB	256KB

Significant gap between DNN requirements and available resources

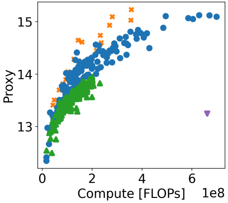
- Low processor speed vs. large number of mathematical operations
- Strict memory limitations vs. large number of weights and big inputs/feature maps
- High precision floating point datatypes vs. hardware often focused on integer arithmetic

ProtototypeNAS

Overview

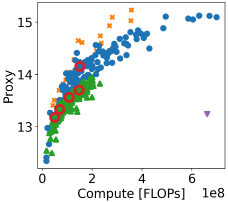
1 Architectural Prototype Exploration

- Multi-objective optimization with FLOPs and zero-shot proxies as objectives
- Combination of architecture, structural, and size (pruning configuration) optimization in a single combined search space



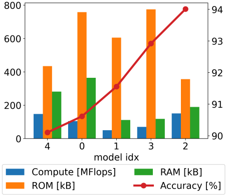
2 Hypervolume Subset Selection

- Top-k selection from the Pareto front of the multiobjective optimization using Hypervolume subset selection.
- Hypervolume subset selection implemented with evolutionary search.

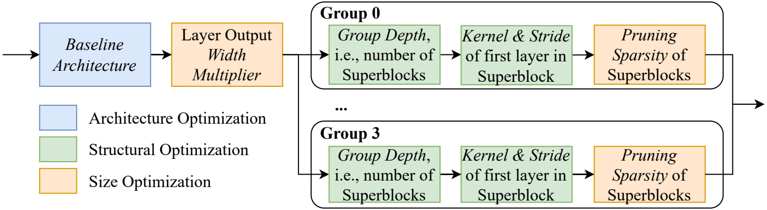


3 Dataset Evaluator

- Top-k selection can be trained, pruned, and quantized for many datasets for the same MCU
- Different trainers can be used to cover a variety of tasks like classification or object detection.



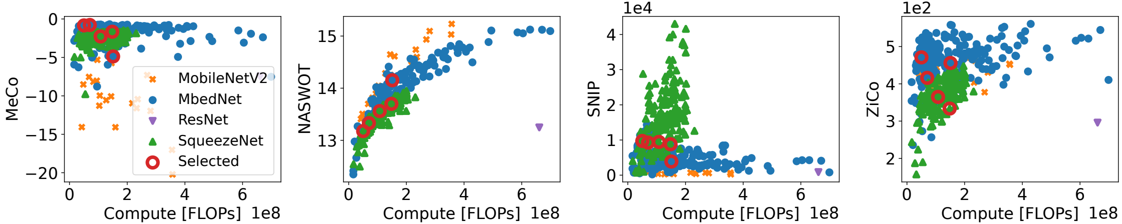
Architectural Prototype Exploration
Search Space Design



Optimization	Hyperparameter	Type	Range/Values
architecture	baseline architecture	categorical	task dependent
structural	group depth i kernel & stride i	categorical categorical	[0, 1, 2, 3] [[3, 2] , [3, 1] , [5, 2] , [5, 1] , [7, 2] , [7, 1]]
size	width multiplier pruning sparsity i	continuous continuous	[0.1, 1.0] [0.1, 0.9]

Evaluation
Architectural Prototype Exploration

Optimization results for an ensemble of four zero-shot proxies and FLOPs, 500 samples



$$\min_{x \in X} \quad \text{flops}(x), -\text{prox}_1(x), \ldots, -\text{prox}_3(x)$$

$$\begin{aligned} \text{s.t.} \quad \text{ram}(x) &\leq \text{ram}_{max} && (\text{ram}_{max} = 480\text{KB}) \\ \text{rom}(x) &\leq \text{rom}_{max} && (\text{rom}_{max} = 2\text{MB}) \\ \text{flops}(x) &\leq \text{flops}_{max} && (\text{flops}_{max} = 1\text{e}9) \end{aligned}$$

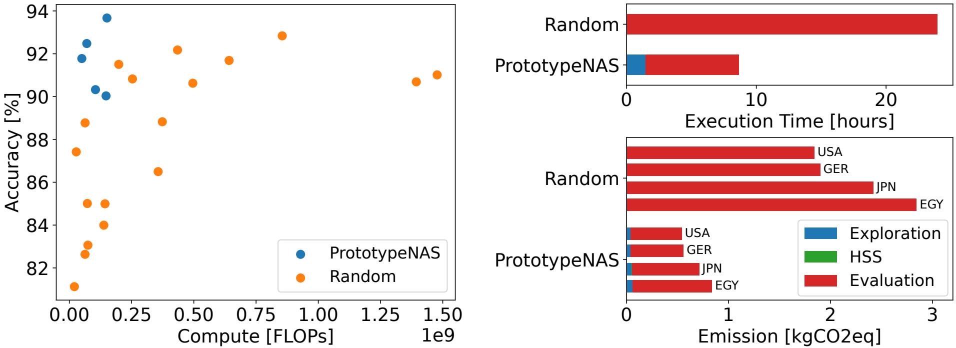
- Zero-shot proxies are known to suffer from structural and feature biases, i.e., they often favor particular network properties [1]
- We selected MeCo, NASWOT, SNIP, and ZiCO because they consider different DNN features [3]
- We use Hypervolume subset selection to choose five DNNs from the set of non-dominated samples

Evaluation
Training on Vision Datasets

Optim. Index	210	283	190	311	237
Architecture	MBedNet	MBedNet	SqueezeNet	SqueezeNet	MBedNet
Quantized Test Accuracy [%] (Post Training Static Quantization)					
CIFAR10	91.8	92.5	90.3	90.0	93.7
ArxPhotos314	90.8	95.4	97.4	79.5	93.6
Flowers	88.9	91.6	82.9	86.3	93.3
Pets	99.8	99.8	97.5	96.1	99.9
GTSRB	96.3	96.1	95.8	97.0	95.5
Compute [MFLOPs]	50.1	70.0	105.0	147.5	151.0
ROM [KB]	605.4	774.9	758.4	434.7	356.5
RAM [KB]	111.7	118.3	364.7	281.6	189.7
Latency [ms]	224.1 ± 0.1	312.9 ± 0.1	367.1 ± 0.1	447.2 ± 0.1	753.2 ± 0.1
Energy [mJ]	77.6 ± 8.6	108.6 ± 9.9	126.0 ± 12.7	156.3 ± 16.8	254.5 ± 13.7

- We measured latency and energy on an iMXRT1062 Cortex-M7 MCU with a Joulescope energy analyzer
- Input image resolution 128×128, 100 training epochs + 50 epochs of ImageNet pretraining

Evaluation
Execution Time and Emission Analysis



- Performance and emission comparison of PrototypeNAS and randomly selected models on CIFAR10
- We used CodeCarbon [2] to assess the execution time and emissions on a desktop computer with 16 GB RAM, an eight-core AMD Ryzen 7 2700 processor, and a GeForce RTX 2070 GPU

References

[1] Kartikeya Bhardwaj et al. "ZiCo-BC: A bias corrected zero-shot NAS for vision tasks". In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE. 2023, pp. 1345–1349.

[2] Benoît Courty et al. mlco2/codecarbon: v3.2.7. 2026. DOI: 10.5281/zenodo.20257903.

[3] Junhao Huang et al. "Evolving comprehensive proxies for zero-shot neural architecture search". In: Proceedings of the Genetic and Evolutionary Computation Conference. 2025, pp. 1246–1254.

[4] Vivienne Sze et al. "Efficient processing of deep neural networks: A tutorial and survey". In: Proceedings of the IEEE 105.12 (2017), pp. 2295–2329.

This work was funded by the European Commission as part of the MANOLO project under the Horizon Europe programme Grant Agreement No.101135782

Thank you!

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS