PrototypeNAS: Rapid Design of Deep Neural Networks for Microcontroller Units

Mark Deutel, Simon Geis, Axel Plinge
Fraunhofer IIS | ECML PKDD 2026

Fraunhofer-Institut für Integrierte Schaltungen IIS

1 Architectural Prototype Exploration

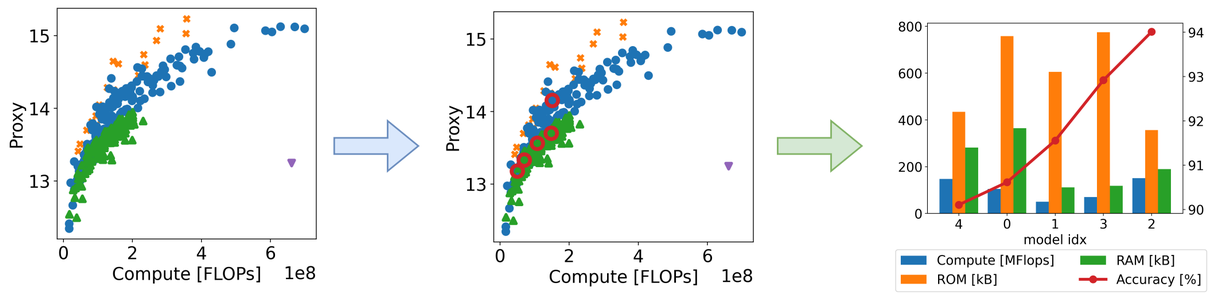
- Multi-objective optimization with FLOPs and zero-shot proxies as objectives
- Combination of architecture, structural, and size (pruning configuration) optimization in a single combined search space

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.

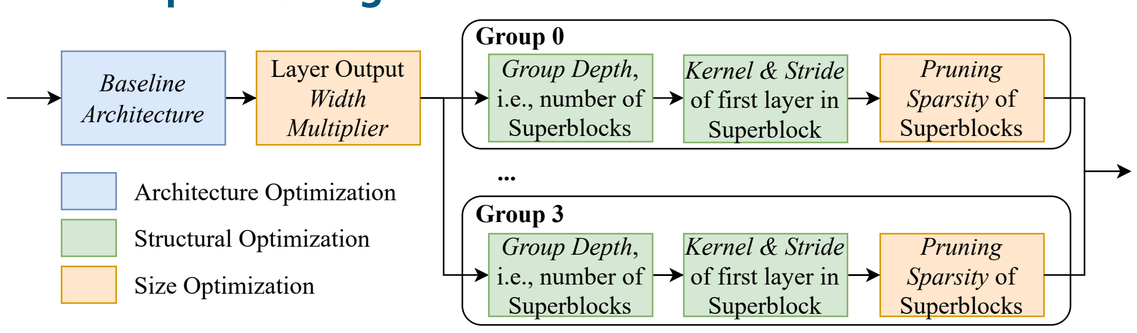
3 Dataset Evaluator

- 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.



- Designing deep neural networks (DNNs) for microcontroller units (MCUs) that have different hardware constraints is challenging.
- It requires solving a multi-objective optimization problem between the accuracy and resource requirements of the DNN candidates.
- Neural architecture search (NAS) can automate the process of designing and fine-tuning DNNs for MCU deployment.

Search Space Design

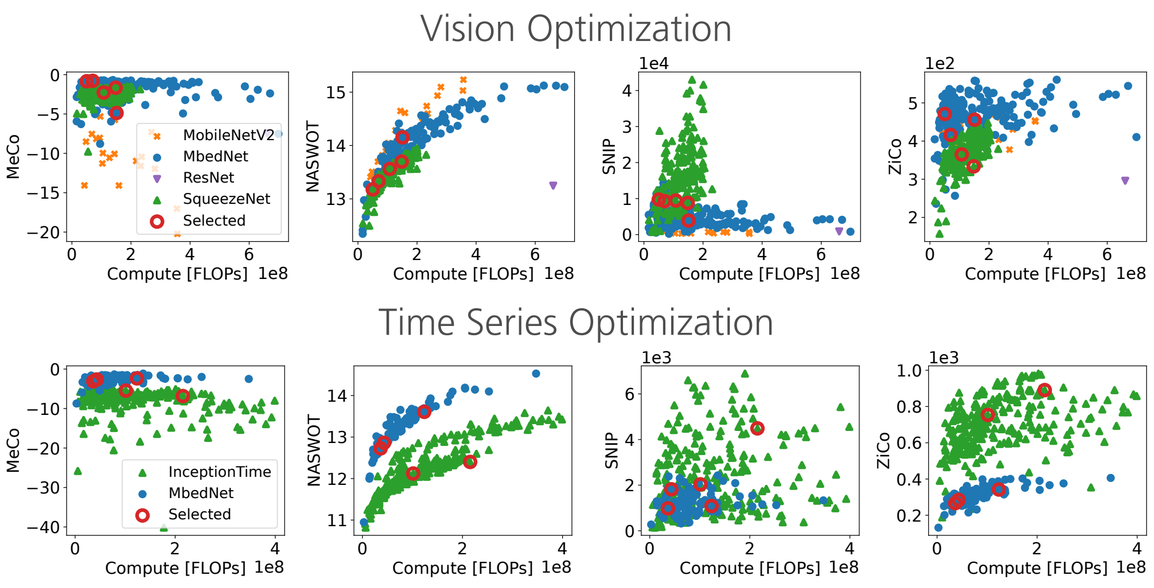


The proposed search space and how DNN prototypes can be created from it. Each baseline architecture consists of repeatable superblocks, i.e., a predefined pattern of layers, organized in the search space into four groups to be optimized separately.

Hypervolume Subset Selection

- We use Hypervolume subset selection to identify a top-k (k = 5) subset of DNN candidates from the set of non-dominated samples.
- Using evolutionary programing, we search for a k-subset whose Hypervolume approximates that of the complete non-dominated set as closely as possible.
- If the set of non-dominated samples is equal to or less than k, we use the non-dominated set directly.

Prototype Exploration



- 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[2].
- We use Hypervolume subset selection to choose five DNNs from the set of non-dominated samples.

Training on Target Datasets

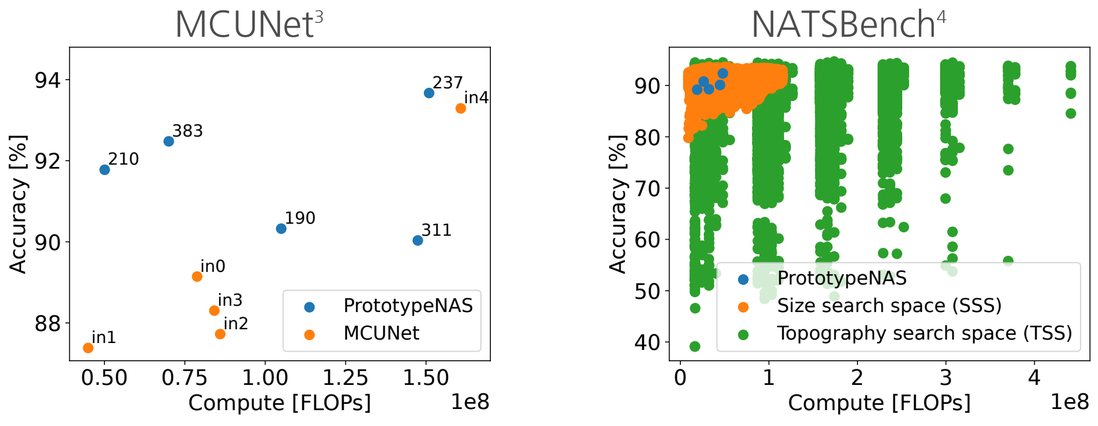
Vision Datasets (100 epochs, 50 epochs of ImageNet pre-training)

Optim. Index	210	283	190	311	237
Architecture	MBedNet	MBedNet	SqueezeNet	SqueezeNet	MBedNet
Quantized Test Accuracy [%] (Post Training Static Qunatization (PTQ))					
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
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

Time Series Datasets (100 epochs)

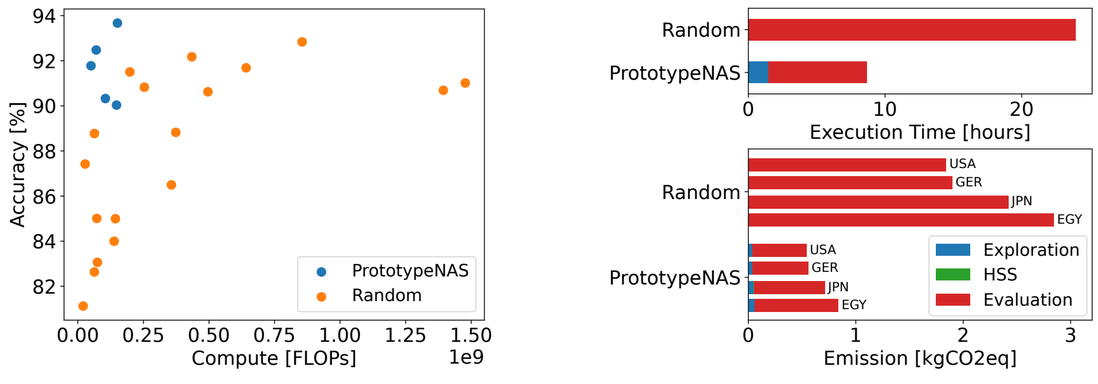
Optim. Index	224	226	73	273	347
Architecture	MBedNet	MBedNet	Inception	MBedNet	Inception
Quantized Test Accuracy [%] (Post Training Static Qunatization (PTQ))					
BitBrain	83.7	87.2	81.7	88.0	78.9
Mafaulda	96.8	98.3	97.4	97.8	98.5
Daliac	95.9	97.1	97.2	97.0	96.1
Compute [MFLOPs]	36.3	43.2	101.5	123.5	215.3
ROM [KB]	149.0	231.0	571.0	337.0	978.0
RAM [KB]	246.0	244.0	251.0	252.0	250.0
Latency [ms]	162.3 ± 0.1	183.6 ± 0.1	447.2 ± 0.1	467.9 ± 0.3	635.0 ± 1.4
Energy [mJ]	55.8 ± 5.7	63.7 ± 6.4	156.0 ± 16.7	165.8 ± 15.9	227.6 ± 17.6

Comparison to Related Work



Execution Time and Emission Analysis

Emission and execution time analysis with random models on CIFAR10.



[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] 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
[3] Ji Lin et al. "Mcunet: Tiny deep learning on IoT devices". In: Advances in Neural Information Processing Systems 33 (2020), pp. 11711–11722
[4] Xuanyi Dong et al. "NATS-bench: Benchmarking NAS algorithms for architecture topology and size". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 44.7 (2021), pp. 3634–3646

MANOLO CLOUD-EDGE EFFICIENT & TRUSTWORTHY AI