Lightweight DNAS for Inference Optimization on Constrained Embedded Nodes

Matteo Risso

Department of Control and Computer Engineering, Politecnico di Torino

Politecnico di Torino

Agenda

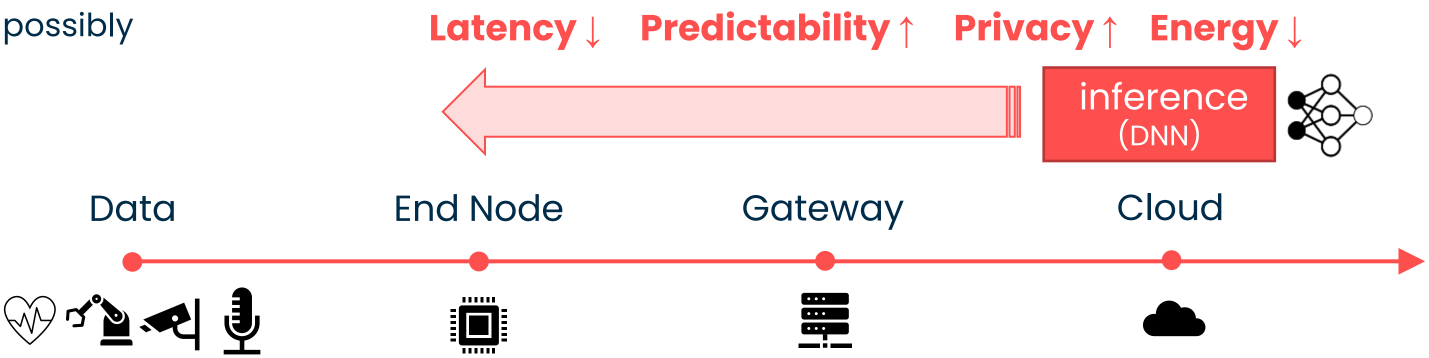
1. AI at the extreme edge: Motivation and General Flow
2. Lightweight Neural Architecture Search
3. Quantization and mixed-precision search

1. AI at the Extreme Edge

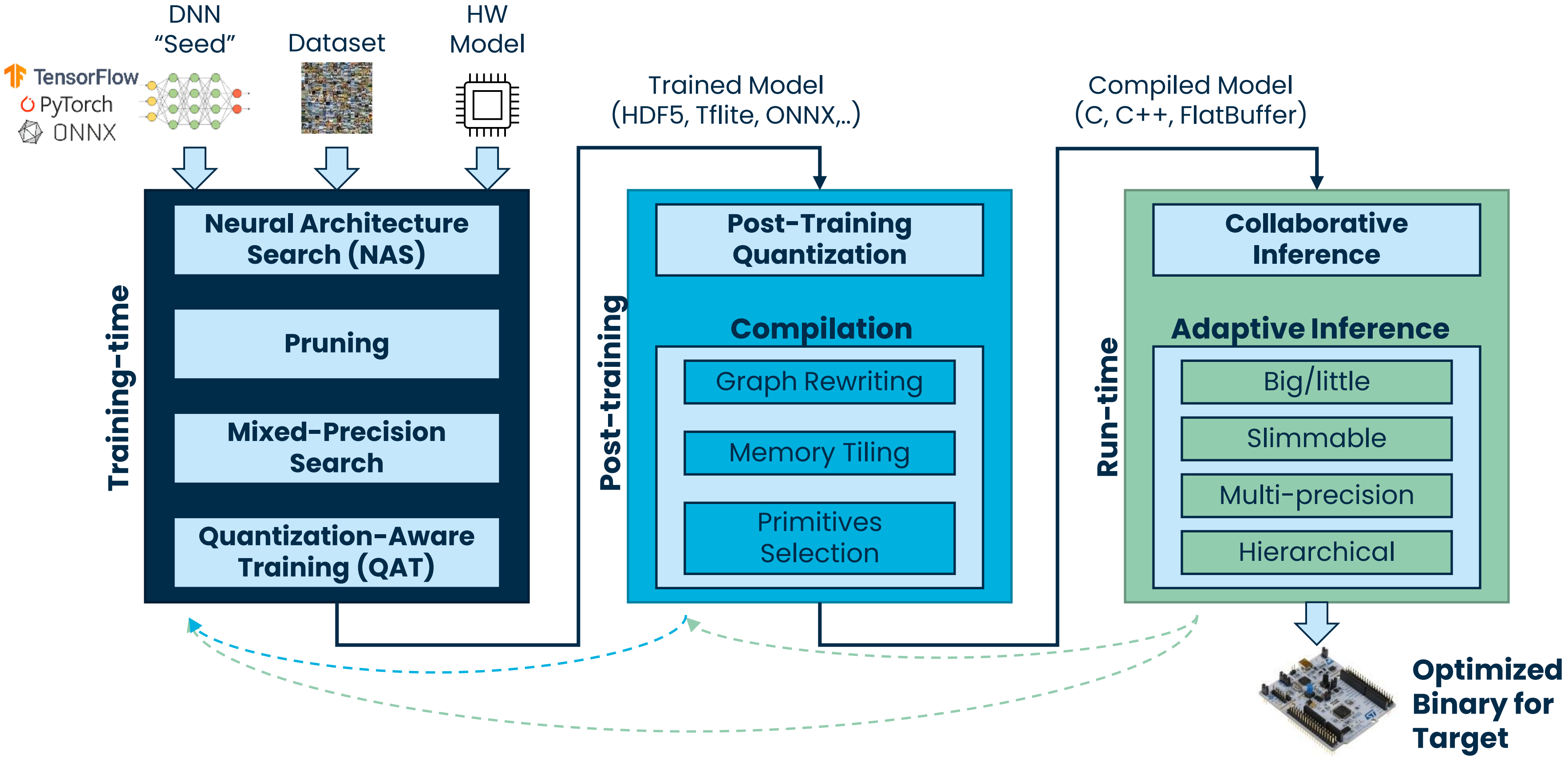
DNNs at the Extreme Edge

- Near-sensor DNN inference has several potential benefits w.r.t. a traditional cloud-centric approach:
  1. More predictable and lower (*) latency
  2. Data privacy
  3. Lower energy consumption (*)

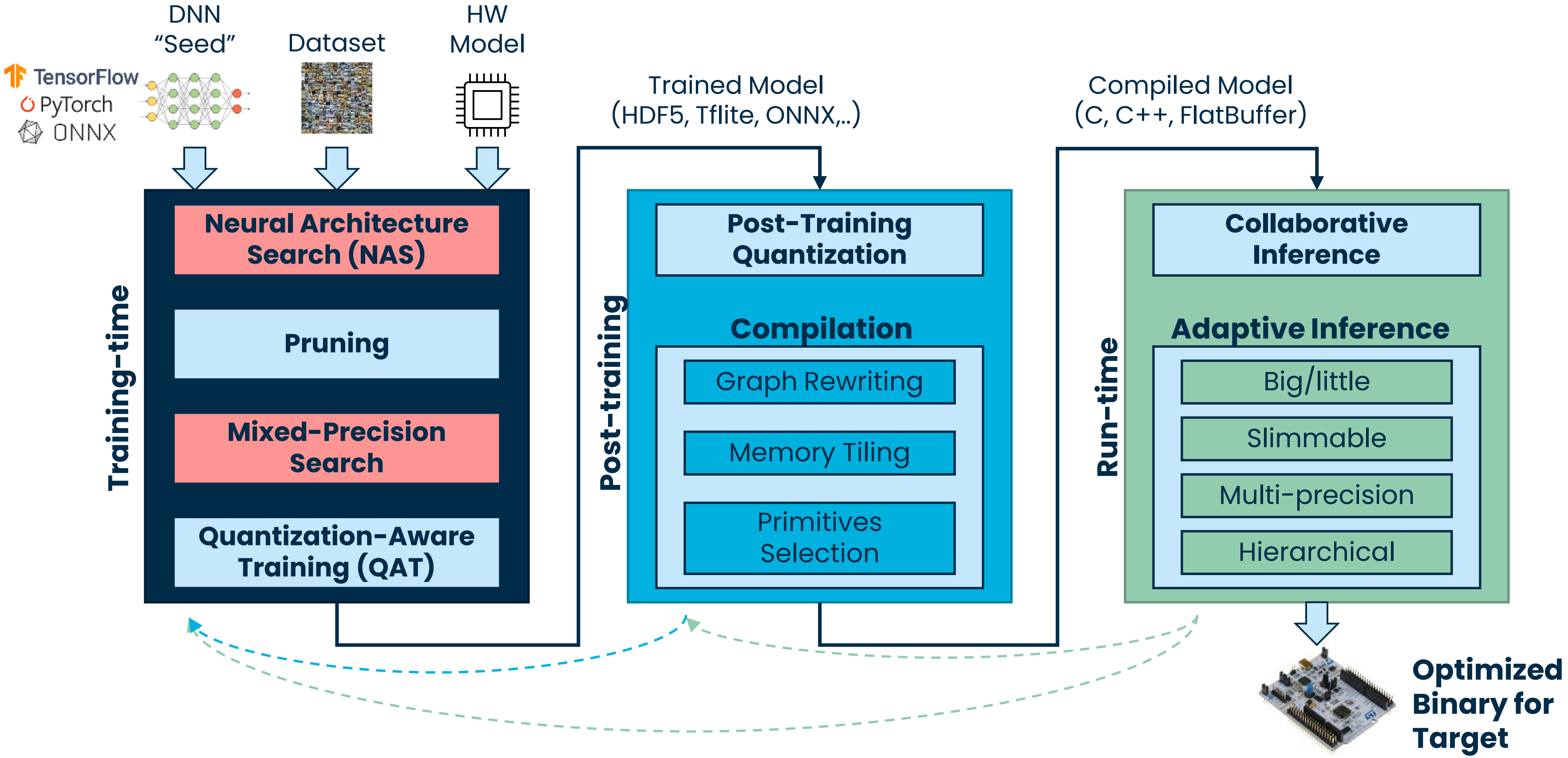
(*) possibly



DNN Deployment Flow



DNN Deployment Flow



2. Lightweight Neural Architecture Search

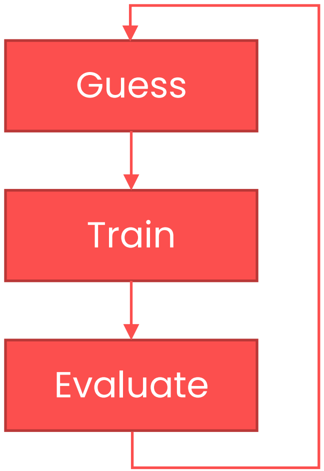
Neural Architecture Search

- Picking hyper-parameters manually is tricky:
  - Biases (rules of thumb, traditions, etc.)
  - Fragmented and coarse design space explorations (e.g., width/res mult in MobileNets)
  - **Classic ML: hand-craft features, DL: hand-craft feature extractors**!
- Neural Architecture Search (NAS):
  - Automatic optimization of the network topology, exploring a large and fine-grain design space of hyper-parameter settings
  - Typically **multi-objective**: co-optimize accuracy and model complexity
    - Model size/#MACs….
    - …or better, **latency/energy directly** (requires models)!

Classic NAS

- Procedure:

Propose 1+ new architecture(s)



Feedback to drive the search

- Key steps:

1. Define the search space:
   - Design variables (topology, cardinality, precision)
   - Discretization of each variable (e.g., #filters in {32, 64, 128}, K in {3, 5, 7}, etc.)
2. Define a search engine:
   - RL, Evolutionary, Bayesian, others…
3. Build a performance estimator:
   - The actual bottleneck!
   - Accuracy estimation encompasses **training**
   - Extra-functional metrics are HW-dependent (**deployment** or accurate model)

- ***Thousands of GPU-hours per search!***

POLITO’s Lightweight NAS

- **Mask-based Differentiable NAS (DNAS):**
  - Relax the search space to make it **continuous and differentiable**
  - Optimize the topology by gradient descent **while training the network**
    - Greatly reduce search costs

- Working principle:
  - Search the architecture hyperparameters “by subtraction”, starting from a large **seed model**
  - **Shrink the seed layers** (e.g., eliminate some channels, reduce the filter size, etc)
  - Similar to structured pruning…

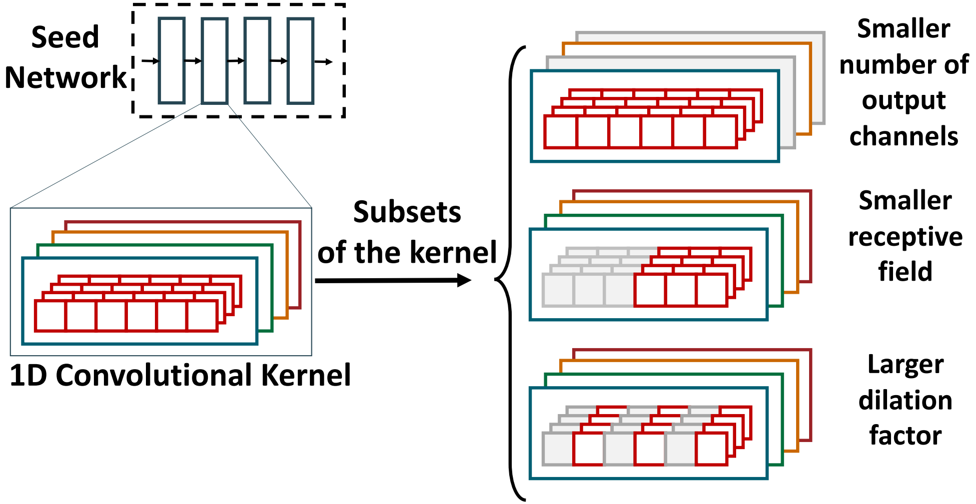
POLITO’s Lightweight NAS

- Named **”Pruning In Time” (PIT):**
  - Hybrid between NAS and pruning
  - Focuses mostly on 1D CNNs for processing time-series.

- Recently applied also to 2D CNNs for vision on nano-drones, with excellent results…

PIT

- **Search space:** For each Convolutional or Fully-Connected layer



PIT

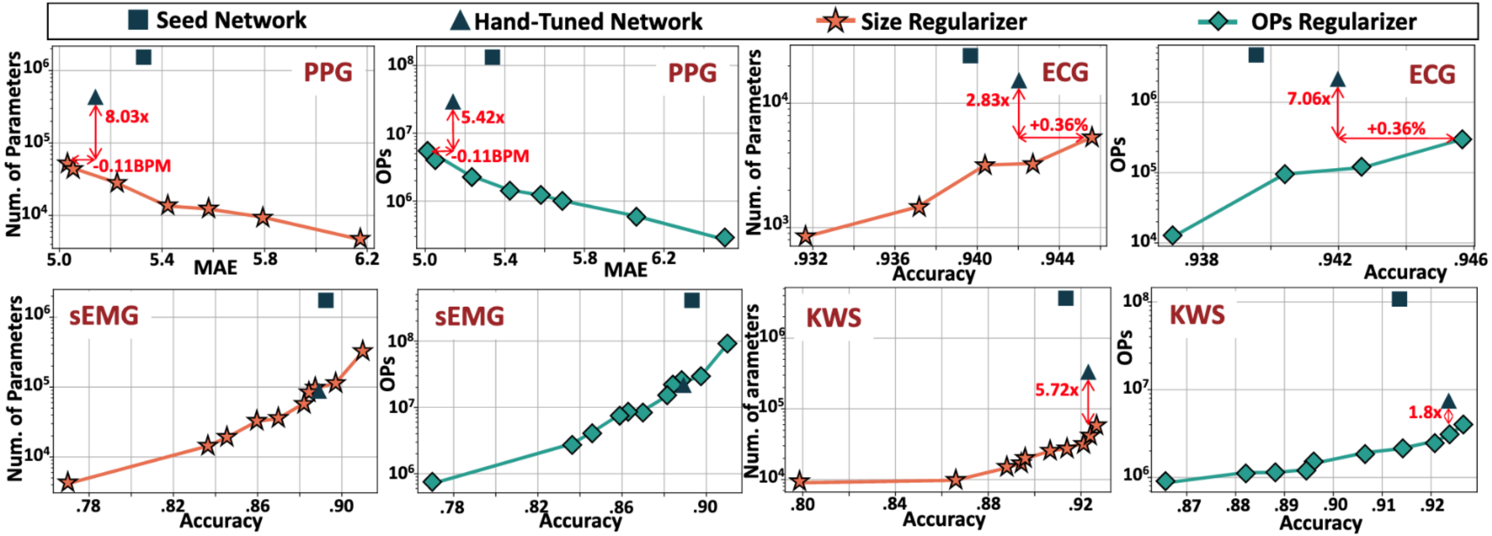
- Add a **L1 regularization term** to the training loss function that brings masks to 0
  - More 0-valued masks → smaller network
  - Must model network complexity in a **differentiable** way
  - Practical regularizers:
    - **N. of weights**, correlates with memory occupation
    - **N. of MACs**, correlates with latency/energy

- **Final Loss Function:** $\min_{W,\theta} \mathcal{L}(W;\theta) + \lambda \mathcal{R}(\theta)$ → Regularizer, function of Trainable binary masks

- Changing the λ yields different trade-offs between accuracy and cost

PIT Results

- 4 edge-relevant benchmarks (biosignals, keyword spotting).
- Up to 8x smaller and 7x faster models at iso-performance



PIT Results

- Up to 5.5x energy reduction with respect to hand-tuned state-of-the-art models when deployed on two different extreme edge devices (GWT GAP8 and STM32H7).

				GAP8		STM32	
Task	TCN	Perf. int8 (float32)	Mem. [kB]	Lat. [ms]	En. [mJ]	Lat. [ms]	En. [mJ]
PPG	HT	5.01 (5.14) BPM	423	23.2	1.2	58.3	13.6
	S	5.71 (6.17) BPM	4.7	1.18	0.06	3.2	0.75
	L	5.01 (5.03) BPM	53.2	4.25	0.22	15.2	3.56
ECG	HT	94.2 (94.2) %	15.2	2.69	0.14	6.66	1.56
	S	92.84 (93.16) %	0.9	0.78	0.04	1.8	0.42
	L	94.13 (94.13) %	5.4	1.26	0.06	2.84	0.66
sEMG	HT	88.89 (88.87) %	88.8	61.0	3.11	291	68.1
	S	86.97 (86.98) %	35.4	39.6	2.02	169	39.5
	L	91.2 (90.99) %	317.8	238	12.1	960	225
KWS	HT	92 (92.31) %	323.4	13.4	0.68	30.7	7.17
	S	87 (86.58) %	9.8	1.40	0.07	2.66	0.62
	L	92.16 (92.64) %	56.5	3.74	0.19	10.6	2.48

PIT Latest Developments

- PIT has been now extended to 2D networks for vision.
  - Example: drone-to-human pose estimation in low-power nanodrones
  - **Same results** of previous hand-tuned network with **3x less memory**, thanks to PIT
  - Collaboration with UNIBO + ETHZ + IDSIA (Lugano)
  - Paper submitted @ ICRA23

PIT Latest Developments

- Real objective: **Minimize latency/energy and maximize accuracy under max memory constraint**
- **Solution:** new loss formulation:

$$\min_{W,\theta} \mathcal{L}(W;\theta) + \lambda|\mathcal{S}(\theta) - s^*| + \mu\mathcal{O}(\theta)$$

  - *S* = size regularizer
  - *O* = ops/latency/energy regularizer
  - s* = size constraint (HW-dependent)
- Sweep μ to trade-off accuracy and latency

PIT Latest Developments

- Tested on 2D CNN, searching the n. of output channels only:
  - IC = image classification, VWW = visual wake word, KWS = keyword spotting
  - Same color points correspond to same s*
  - Almost 1 order of magnitude span in OPs and ± 5% accuracy for the same size

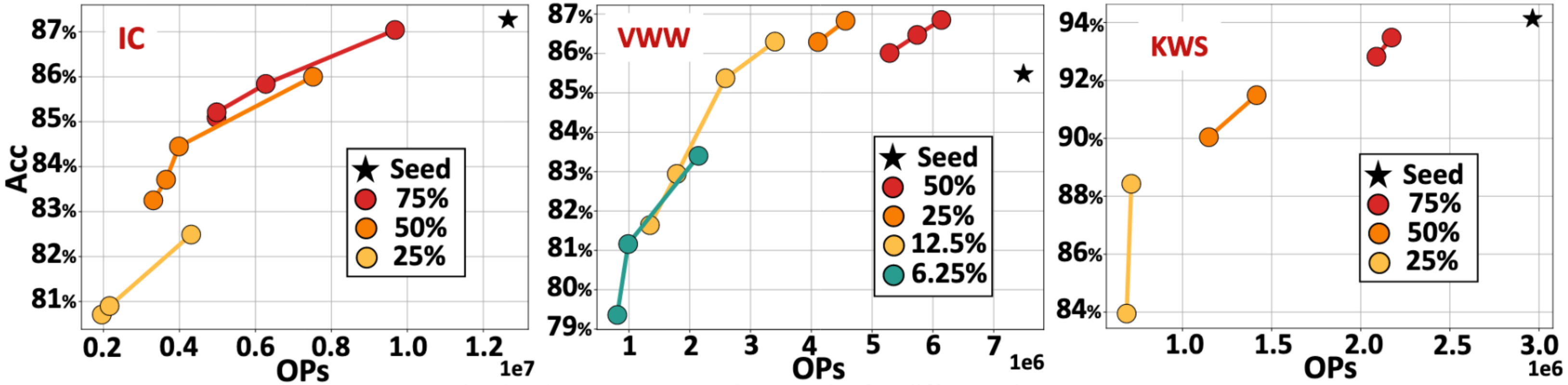


Fig. 3. Accuracy versus OPs results for different size targets.

NAS@POLITO References

- PIT:
  - M. Risso et al, *"Lightweight Neural Architecture Search for Temporal Convolutional Networks at the Edge"*, IEEE Trans. on Computers 2022
  - M. Risso et al, *"Pruning In Time (PIT): A Lightweight Network Architecture Optimizer for Temporal Convolutional Networks"*, Proc. ACM/IEEE DAC 2021
- Multi-regularization:
  - M. Risso et al, *"Multi-Complexity-Loss DNAS for Energy-Efficient and Memory-Constrained Deep Neural Networks"*, Accepted at ISLPED 2022.
- Application to PPG-based HR Monitoring:
  - A. Burrello et al, *"Q-PPG: Energy-Efficient PPG-based Heart Rate Monitoring on Wearable Devices"*, IEEE Trans. on BioCAS, 2021
  - M. Risso et al, *"Robust and energy-efficient PPG-based heart-rate monitoring"*, Proc. IEEE ISCAS 2021
- **Code:** https://github.com/EmbeddedML-EDAGroup

NAS@POLITO Future Work

- Combine mask-based approach with other types of NAS to support a wider search space
- Joint NAS and mixed-precision search (see next)
- Better HW-aware models.

3. Quantization and Mixed-Precision Search

Quantization

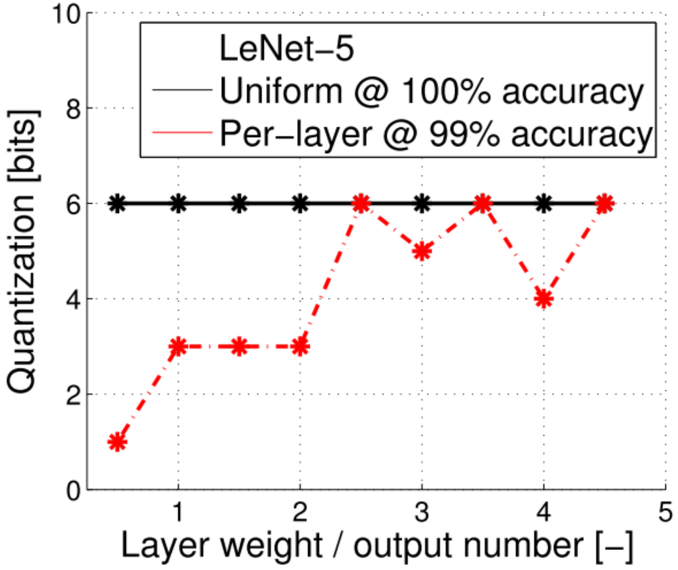
- DNNs are very tolerant to the use of low-precision data representations for weights & activations
  - **For extreme edge, quantization can be mandatory (no FPU).**

- Edge de facto standard: **8bit integer quantization**
  - Well supported by HW ISAs
  - Little degradation in accuracy, especially with QAT (empirical “sweet spot”)

Fixed- vs Mixed-Precision

- **Fixed-precision**: while Δ and z change per-tensor (or channel), the bit-width N is fixed for the entire network

- **Mixed-precision:** uses a different N layer-wise or channel-wise.
  - Typically **1, 2, 4, 8-bit**
  - N_w (weights) can differ from N_x (activations)
  - Possibly higher compression for the same accuracy



[Source] B. Moons. Energy-efficient ConvNets through approximate computing, 2016

Mixed-Precision Quantization

- **Bit-width assignment** problem:
  - How to assign N to different layers?
  - Huge search space: $((N_{prec})^2)^{N_{layers}}$

- Classical solutions:
  - Black-box meta-heuristics (e.g., genetic algorithms)
  - Greedy
  - Simulated Annealing

- Can be approached with a method **similar to DNAS!**

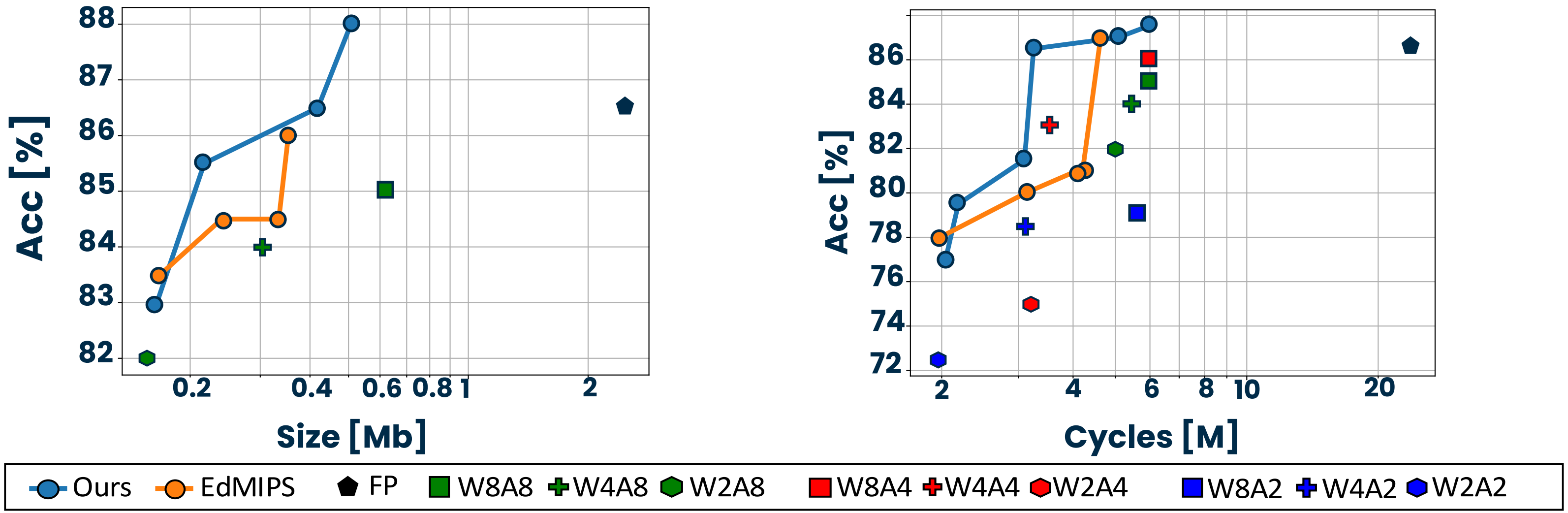
Mixed Precision @ POLITO

- SoTA:
  - Per-channels quantization parameters (Δ and z)
  - Per-layer bit-width (mixed-precision)

- How to achieve further compression? → **per-channel bit-width**
  - Currently applied only to weights

Mixed Precision Results

- CIFAR-10 + ResNet8:
  - deployed on **MPIC**: RISC-V PULP core with support for 1/2/4/8-bit MACs
  - Up to **54% memory reduction** and **36% cycles reduction** at **iso accuracy** w.r.t . EdMIPS



Mixed Precision @ POLITO References

- Search tool:
  - M. Risso et al, *"Fine-grained mixed-precision quantization through efficient DNAS for memory-constrained MCUs", arXiv preprint arXiv:2206.08852.*

- Applications of EdMIPS to extreme edge tasks:
  - A. Burrello et al, *"Q-PPG: Energy-Efficient PPG-based Heart Rate Monitoring on Wearable Devices"*, IEEE Trans. on BioCAS, 2021
  - F. Daghero et al, *"Human Activity Recognition on Microcontrollers with Quantized and Adaptive Deep Neural Networks"*, ACM Trans. on Embedded Systems, 2022.

Mixed Precision @ POLITO Future Works

- Joint NAS and mixed-precision search
- Target domain-specific accelerators that support peculiar quantization formats (e.g., Analog In-Memory Computing)