Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
37th cycle

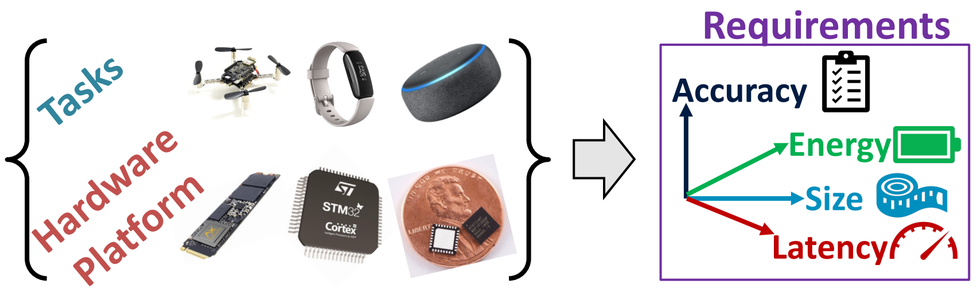
Supervisors
Prof. Massimo Poncino
Prof. Daniele Jahier Pagliari

Automatic hardware-aware design and optimization of deep learning models

PhD Candidate: *Matteo RISSO* Email: matteo.risso@polito.it

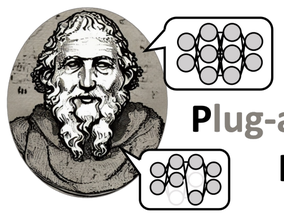
1. Introduction and Goals

- Deep Neural Network Inference is moving from **cloud** to **edge**.



- Given a set **{Task, Hardware Platform}** can we **automate** the design of DNNs to satisfy **requirements**?

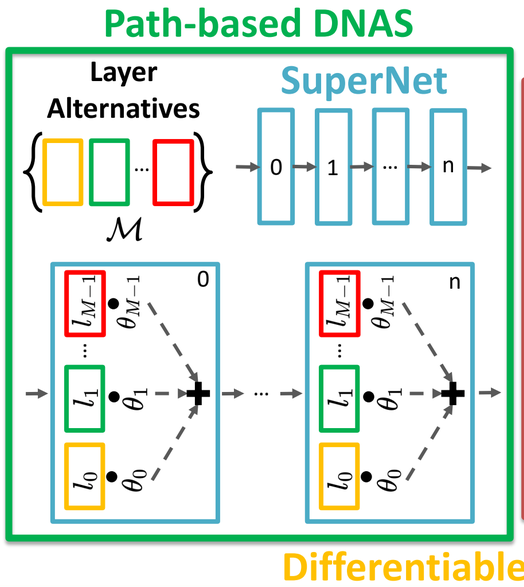
2. Method



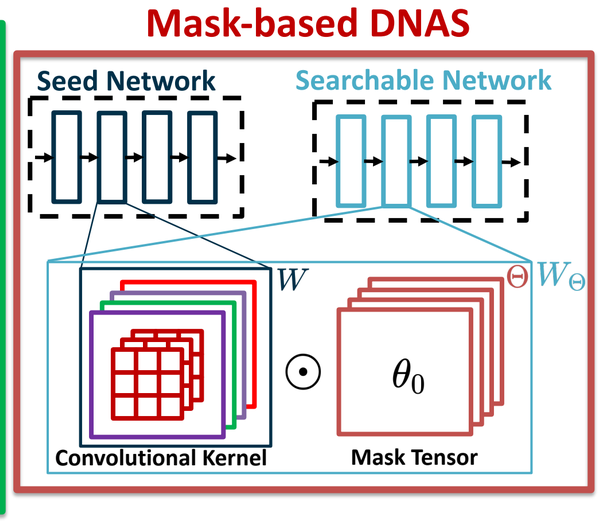
PLiNIO

Plug-and-play **Li**ghtweight **N**eural **I**nference **O**ptimization

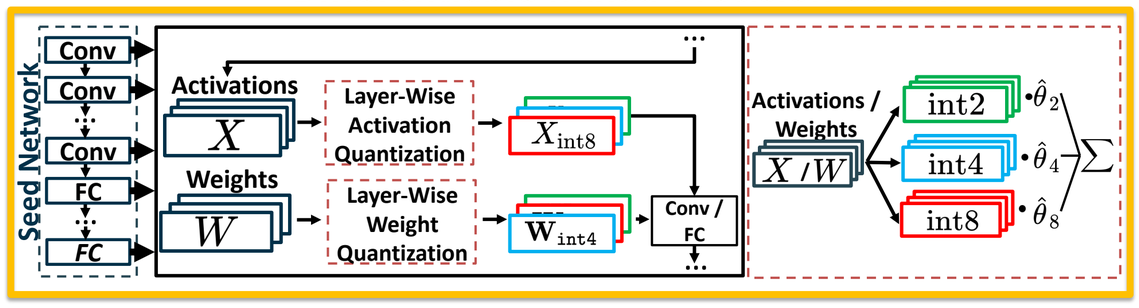
Path-based DNAS



Mask-based DNAS



Differentiable Mixed-Precision Search



- **PLiNIO** [3] is a Python package based on PyTorch that provides a **P**lug-and-play **Li**ghtweight tool for Deep **N**eural networks (DNNs) **I**nference **O**ptimization.
- It supports the following *gradient-based* methods: Path-based Differentiable NAS (DNAS), Mask-based DNAS [1], Differentiable Mixed-Precision Search [2].

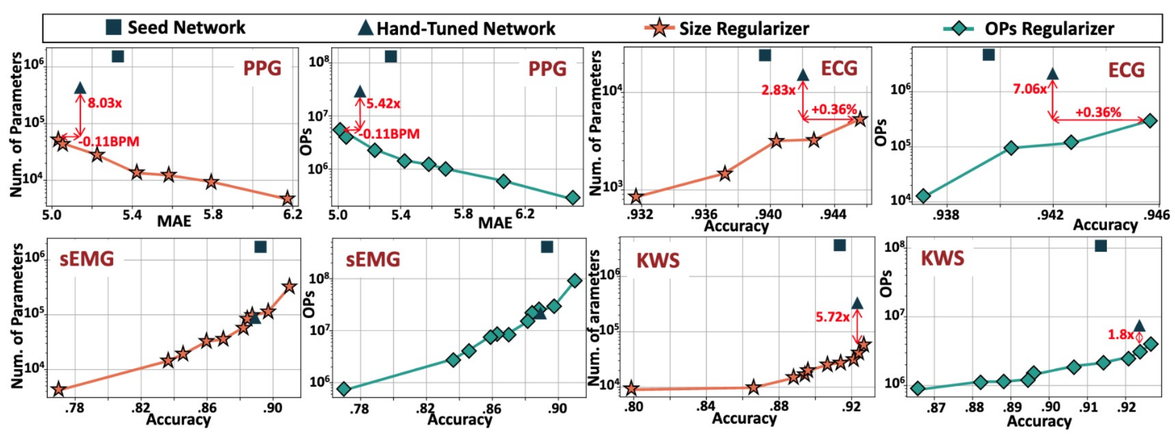
- In general, the the considered optimization problem is:

$$\min_{W,\theta} \mathcal{L}_{\mathcal{T}}(W;\theta) + \lambda \mathcal{L}_{\mathcal{R}}(\theta)$$

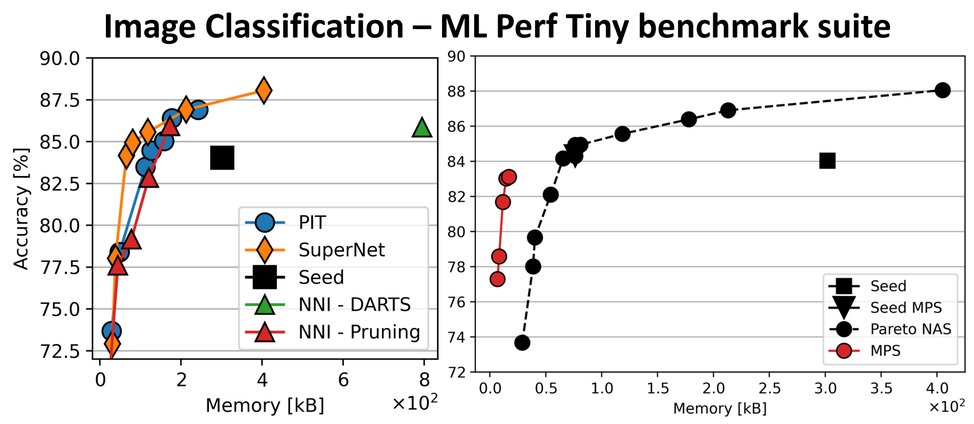
- E.g., in DUCCIO [4] we considered the case of multiple constraints (e.g., latency, size, etc.):

$$\mathcal{L}_{\text{task}}(W,\theta) + \textstyle\sum_j \lambda_j \max(0, \mathcal{R}_j(\theta) - T_j)$$

3. Results



- Rich collection of **Pareto optimal** architectures using **PIT** (mask-based DNAS) [1].



- Rich collection of **Pareto optimal** architectures with up-to **94%** smaller memory footprint @ <1% accuracy drop with respect to baseline *using all PLiNIO optimizations in cascade*.

4. References

1. **M. Risso,** et Al., "Lightweight Neural Architecture Search for Temporal Convolutional Networks at the Edge," in IEEE TCOMP.
2. **M. Risso,** et Al., "Channel-wise Mixed-precision Assignment for DNN Inference on Constrained Edge Nodes," IEEE IGSC.
3. D. J. Pagliari, **M. Risso**, et Al., "PLiNIO: A User-Friendly Library of Gradient-Based Methods for Complexity-Aware DNN Optimization," FDL.
4. A. Burrello, **M. Risso**, et Al., "Enhancing Neural Architecture Search with Multiple Hardware Constraints for Deep Learning Model Deployment on Tiny IoT Devices," in IEEE TETC.