Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXIV cycle

Supervisor
Prof. Andrea Calimera
Prof. Enrico Macii

HW/SW Co-Design and Optimization for Intelligent Embedded Systems

PhD Candidate: *Antonio Cipolletta*

1. Introduction

- Modern **Deep Neural Networks** (DNNs) represent the backbones of several computer vision, audio and natural language processing applications.
- The **deployment** of DNNs on **embedded systems** enables us to sense the physical world with high privacy standards, low latency, and high energy efficiency.
- Unfortunately, modern DNNs require **huge computational power**, **large storage** and **memory footprint**. However, embedded systems have **limited computational power**, **low storage capacity,** and **small on-chip memory**.

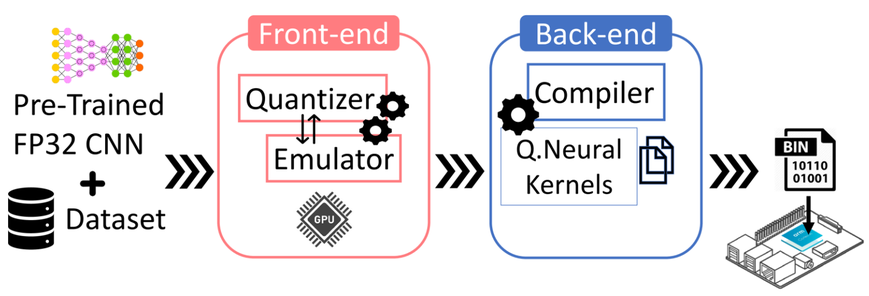
2. Research Goals

- Provide application designers with a **toolbox of optimization techniques** to **make DNNs** not only **accurate** ⌖ but also:
  - ☑ **fast**
  - ☑ **energy-efficient**
  - ☑ **small**
  - ☑ **adaptable**

  to satisfy the physical constraints and the requirements of embedded systems.

3. The Optimization Toolbox

- To make DNNs faster and more energy-efficient, in [1], we proposed an **end-to-end optimization flow** consisting of an HW-aware quantization process and optimized fixed-point convolutional routines.



- To build small but accurate DNNs, in [2], we proposed combining **input resolution scaling** with HW-aware **neural architectural design** and **training pipeline** to bring DNNs on tiny devices powered by microcontroller units (MCU).
- To reduce the activation footprint of a DNN, in [3], we proposed a compiler pass that identifies the sub-graphs with the highest memory requirements and then applies a **functional-preserving topology restructuring** to reduce the peak memory consumption.

4. Results

- On a RaspberryPi 3B, our solution [1] was up to **3x faster** and **more energy-efficient** than SOTA with similar accuracy.
- On an STM32-F7 powered by an ARM CortexM7, uPyD-Net [2] processes **3 FPS** within a **350mW** power budget.
- Results collected in [3] on several DNNs show that the topology restructuring achieves **remarkable memory savings** (62.9% on avg.) with **low computational overhead** (8.6% on avg.).
- In [4], we demonstrated that **composing sparsity and graph transformations** makes highly accurate DNNs feasible on devices with minimal memory resources (512KB of RAM and 1MB of FLASH).

5. References

1. Peluso, V., Cipolletta, A., et al. "Enabling energy-efficient unsupervised monocular depth estimation on armv7-based platforms." Proc. DATE 2019.
2. Peluso, V., Cipolletta, A., et al. “Monocular Depth Perception on Microcontrollers for Edge Applications.” IEEE TCSVT 2021.
3. Cipolletta, A., et al. “Dataflow Restructuring for Active Memory Reduction in Deep Neural Networks.” Proc. DATE 2021.
4. Cipolletta, A., et al. “On the efficiency of Sparse-Tiled Tensor Graph Processing for Low Memory Devices.” Proc. DAC 2021.