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 Optimizations for Embedded Deep Neural Networks

PhD Candidate: *Luca Mocerino*

1. Intro & Motivations

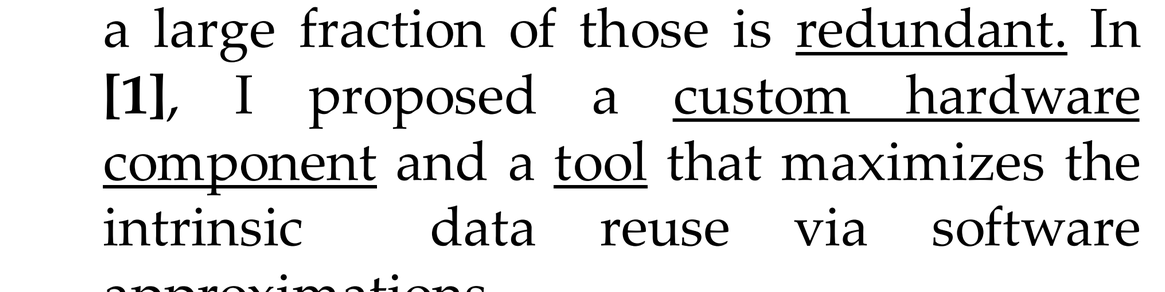
The deployment of **Deep Neural Networks (DNNs)** on **Internet-of-Things** devices reflects the possibility of building distributed services with **low response time, high energy efficiency, and privacy standards.** However, a typical **edge node** is resource- and computationally-constrained while the state-of-the-art DNNs require **massive computational workload, large storage and memory footprint.**

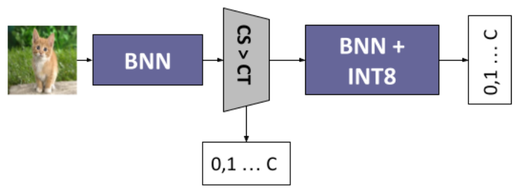
2. Research Goals

Provide a set of tools to deploy **more compact, faster and more energy-efficient** DNNs on **embedded systems**, while preserving their **accuracy**.

3. Tools & Optimizations

- The DNNs workload is dominated by **multiply-and-accumulate** operations but a large fraction of those is redundant. In **[1]**, I proposed a custom hardware component and a tool that maximizes the intrinsic data reuse via software approximations.
- Classical DNNs are designed to spend the same maximal effort on all inputs, causing *a large resource waste on average*. In **[2]**, a **Binary Neural Network (BNNs)** is used for *easy inputs*, while an **8-bit fixed-point (INT8)** is used for *hard inputs*.





- In **[3]**, I proposed an efficient implementation for **Test-time Augmentation**, a technique that improves the accuracy at run time. I implemented a conditional mechanism that runs only the augmented samples that improve accuracy.
- In **[4]**, I presented an architectural template based on the ensemble of BNNs, which reduces the storage requirements with no accuracy loss.

4. Achievements & Results

- In **[1]**, results reveal that the proposed tool achieves up to **77% energy savings with less than 1% of accuracy loss** w.r.t standard arithmetic processing element.
- In **[2]**, results collected on two MCUs reveal, for three computer vision tasks, a **speed-up of 81.49%** and a **gain of accuracy up to 3.8%** w.r.t INT8.
- In **[3]**, results on image classification tasks, on ARM Cortex-A53, show up to **2.2x of speed-up with no accuracy loss**.
- In **[4]**, results show up to **90% of memory reduction w.r.t FP32** model and up to **77% w.r.t SoTa with no quality loss**.

5. References

life.augmented

1. **Mocerino, L.**, et al., *"Energy-efficient convolutional neural networks via recurrent data reuse"*, Proc. DATE 2019.

2. **Mocerino, L.**, and Calimera A., *"Fast and accurate inference on microcontrollers with boosted cooperative convolutional neural networks (bc-net)."* IEEE TCAS-I 2021.

3. **Mocerino, L.** et al., *"AdapTTA: Adaptive Test-Time Augmentation for Reliable Embedded ConvNets"*, Proc. VLSI-SoC 2021.

4. **Mocerino, L.**, and Calimera A., *"TentacleNet: A pseudo-ensemble template for accurate binary convolutional neural networks"*, Proc. AICAS 2020.