Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XVI cycle

life.augmented

Supervisor
Enrico Macii
Massimo Poncino

Implementation of Machine Learning Algorithms on Ultra-Low-Power Hardware for In-Sensor Inference

PhD Candidate: *Francesco Daghero*

1. Introduction / Context

Machine Learning (ML) at the edge has become increasingly popular, since it may lead to **lower latency** and **higher energy efficiency**. Running ML models on resource-constrained devices requires optimizations to reduce their memory and energy footprint with minimal accuracy drops.

2. Goal / Objectives

This PhD thesis, funded by STMicroelectronics, focuses on the introduction of novel Edge-ML optimizations for ultra-low power devices.

3. Method and Results

Edge-ML approaches can be divided in two main categories: i) **static**: the model is changed once and becomes fixed (e.g., quantization; ii) **dynamic** : the model changes at runtime depending on external conditions (e.g. input, battery life, etc.)

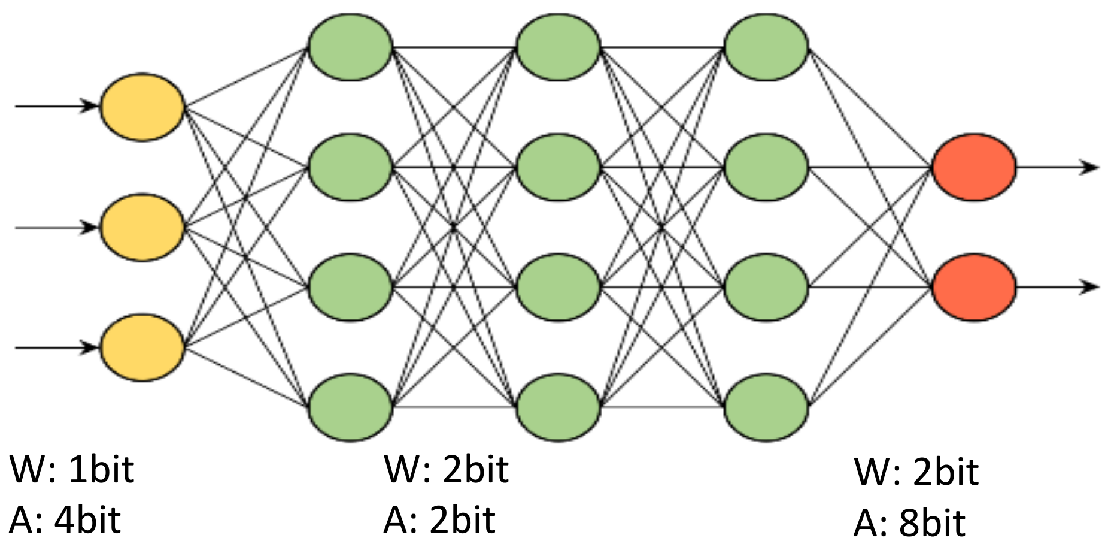

Figure 1 Mixed precision neural network

In [1] we explore sub-byte quantization and mixed precision (Fig. 1) for Human Activity Recognition based on inertial data, showing savings up to **91% memory** w.r.t the standard 8-bit integer quantization.

In [2] we deploy an input-adaptive system (Figure 2) composed of : A fast and inexpensive **Decision Tree** (DT) to predict easy activities. A **Convolutional Neural Network** (CNN) enabled only for complex activities, unrecognized by the DT.

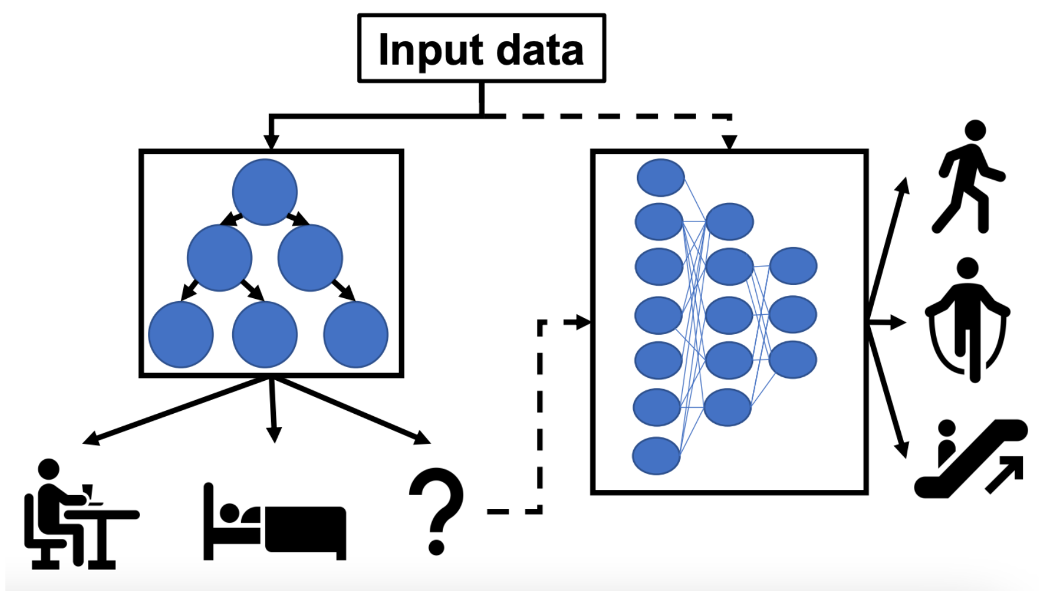

Figure 2 Adaptive Inference for Human Activity Recognition

This approach saves up to **32% energy** per inference, while also reducing the memory by **up to 64%** at iso-accuracy.

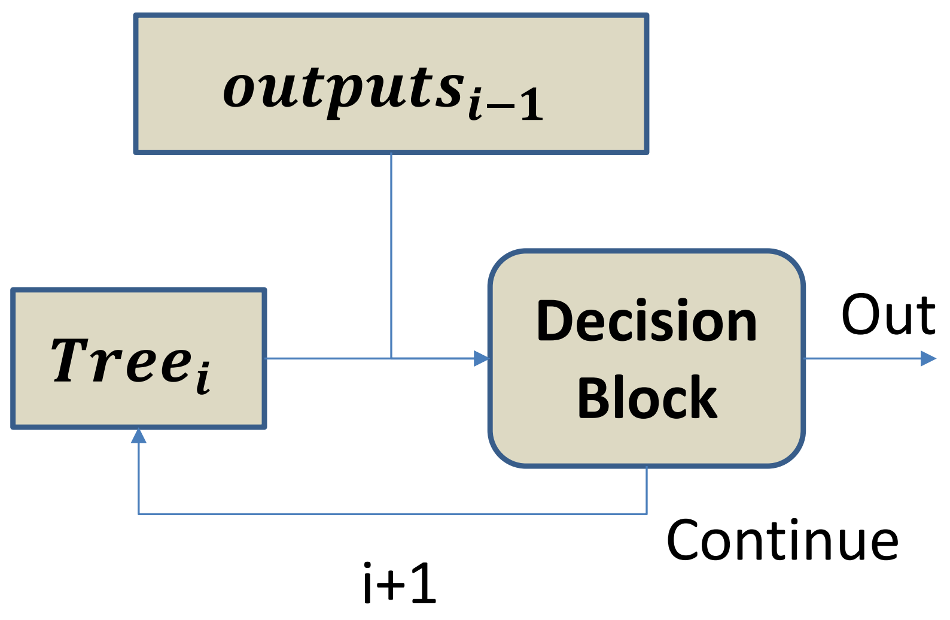

Figure 3 Adaptive Inference for Random Forests

In [3-4] we design an early-stopping (Fig. 3) approach for Random Forests (RFs). The dynamic model will run as many DTs as needed to satisfy the stopping criterium implemented by the Decision Block, halting the execution sooner when the condition is met (e.g., high prediction confidence). This approach saves up to 91% energy per inference at iso-accuracy.

4. References

1. Daghero, Francesco et al. "Human Activity Recognition on Microcontrollers with Quantized and Adaptive Deep Neural Networks." ACM TECS 2022
2. Daghero, Francesco et al.,, "Two-Stage Human Activity Recognition on Microcontrollers with Decision Trees and CNNs." , IEEE PRIME 2022.
3. Francesco Daghero et al., Adaptive Random Forests for Energy-Efficient Inference on Microcontrollers, VLSI-SOC 2021
4. Daghero, Francesco, et al. "Low-Overhead Early-Stopping Policies for Efficient Random Forests Inference on Microcontrollers." VLSI-SoC: Extended Selected Papers. Cham: Springer Nature Switzerland, 2022.