Exploring Tree-Based Inference Engines for Low-Power Learning Applications
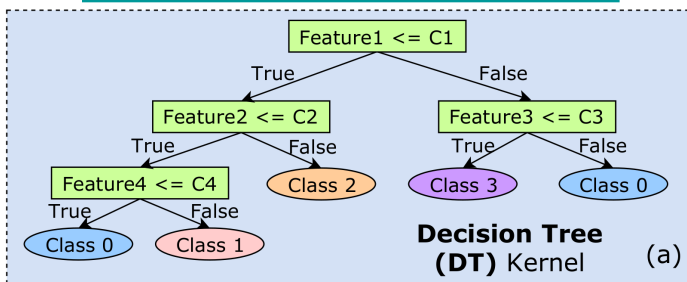
Brunno Abreu, Mateus Grellert, Sergio Bampi - UFRGS

Introduction

- **Machine Learning** has become popularized and is being applied in several fields

- When considering **embedded devices, power dissipation** of ML tools must be considered

- Neural Networks can still be costly in embedded devices

- Simpler models can be used, such as **Decision Trees** (DTs)

> **GOAL**: propose a design flow for analyzing trade-offs of DT models with varying complexity (input quantization/tree depth)

Decision Tree (DT) Kernel (a)

Proposal

- **Design flow** proposed to generate HDL descriptions and obtain power results

- FPUs dissipate too much power; therefore, **quantization** is performed

- **DT training** performed on software using Python Scikit-Learn library

- **Tree-VHDL translation**: translates the Python structure into VHDL synthesizable code

- Architectures synthesized for 65nm ASIC std-cells library, with **Cadence Genus**

Results

- **Input quantization** (approximation) led to accuracy increases of up to 8.7%

- Some models obtained similar accuracy results with **less power dissipation**

- Except for the wearable data set, every model reached a **stable accuracy** with relatively shallower trees

- Input quantization higher than **10 bits** did not present significant improvements in accuracy

- Some **data sets** obtained the best accuracy values with **smaller widths**

Contact: baabreu@inf.ufrgs.br

IEEE CAS/CEDA Electronic Design Automation School 2020