Energy–Accuracy Trade-offs in Edge AI Accelerators for Real-Time Computer Vision

Supervisor: *Sifat Rezwan, Juan Cabrera*

Deploying AI workloads at the edge demands a balance between model accuracy, inference latency, and energy consumption. Green AI emphasizes minimizing the environmental and energy footprint of AI while retaining reliable performance. With diverse hardware platforms available (Axelera AI accelerator, Hailo AI accelerator, Nvidia Jetson Nano, Asus AI accelerator, and Nvidia GPUs), this project explores how computer vision models behave under varying constraints, providing a systematic study of the energy–accuracy trade-offs across hardware.

What you will do

- Implement a baseline pipeline with state-of-the-art deep learning models (e.g., ResNet, MobileNet, YOLO).
- Deploy the models across heterogeneous hardware (Axelera, Hailo, Jetson Nano, Asus, GPU).
- Measure and benchmark:
    - Accuracy: Top-N classification accuracy, mAP for detection.
    - Latency: average inference time per sample.
    - Throughput: effective frames per second (FPS).
    - Energy: Joules per inference.
    - Energy–accuracy Pareto front analysis, if possible.
- Investigate optimization techniques: quantization, pruning, model compression.
- Conduct a multi-objective evaluation of accuracy vs. efficiency trade-offs.
- Provide guidelines for choosing accelerators for specific energy/accuracy constraints.

Why these matters

This thesis will provide concrete insights into the environmental and performance impact of deploying AI models on different accelerators. Results will guide both industry and academia in sustainable deployment strategies for AI at the edge, supporting the broader goals of Green AI.

Required skills

- Python (must), ML frameworks (PyTorch/TensorFlow).
- Familiarity with computer vision architectures.
- Basic knowledge of Linux and hardware benchmarking.