LLM-assisted RTL generation for a pretrained deep neural network

Project context

State-of-the-art neural networks (e.g., object detection or image classification) are typically developed and distributed as **pretrained models** in PyTorch/TensorFlow. Deploying these models efficiently in custom silicon requires translating the high-level network description into a **hardware implementation** that can be synthesized and evaluated for **power, performance, and area (PPA)** in an advanced technology node.

This project investigates a workflow in which a **large language model (LLM)** is used as an engineering assistant during the conversion from pretrained model → hardware architecture

→ **synthesizable RTL**, followed by PPA evaluation in the **imec N2 Pathfinder** technology environment.

Goal and research questions

Goal: Take a **complex pretrained network** (e.g., YOLO-class detector or ResNet-class classifier) and produce a **synthesizable RTL implementation** (full network or a well-defined, compute-dominant subset) suitable for PPA evaluation in imec N2 Pathfinder.

Research questions include:

- Which stages of model-to-RTL translation benefit most from LLM assistance (architecture mapping, RTL generation, verification, scripting, debug)?
- What is the resulting implementation quality in terms of **PPA** compared to a conventional/manual workflow?
- What are the dominant bottlenecks in advanced-node implementation (memory, bandwidth, MAC utilization, control overhead, timing closure)?

Reference: Tomlinson, Michael, Joe Li, and Andreas Andreou. "Designing silicon brains using LLM: Leveraging ChatGPT for automated description of a spiking neuron array." *2024 Argentine Conference on Electronics (CAE)*. IEEE, 2024.

Supervisor: Amirreza Yousefzadeh, CAES research group, a.yousefzadeh@utwente.nl

Coordinator: Ghayoor Gillani, CAES research group, s.ghayoor.gillani@utwente.nl