Contribution ID: **51**

Type: **Standard 15 min talk**

rule4ml: An Open-Source Tool for Resource Utilization and Latency Estimation for ML Models on FPGA

Thursday 17 October 2024 13:10 (15 minutes)

Deploying Machine Learning (ML) models on Field-Programmable Gate Arrays (FPGAs) is becoming increasingly popular across various domains as a low-latency and low-power solution that helps manage large data rates generated by continuously improving detectors. However, developing ML models for FPGA deployment is often hindered by the time-consuming synthesis procedure required to evaluate resource usage and latency. In particular, the synthesis has a chance of failing depending on the ML architecture, especially if the required resources exceed the capacity of the target FPGA, which in turn makes the development process slow and repetitive.

To accelerate this development, we introduce rule4ml, an open-source tool designed to predict the resource utilization and inference latency of Neural Networks (NNs) before their synthesis and implementation on FPGA. We leverage hls4ml, a framework that helps translate NNs into high-level synthesis (HLS) code, to synthesize a diverse dataset of NN architectures and train resource utilization and inference latency predictors. While hls4ml requires full synthesis to obtain resource and latency insights, our method uses trained regression models for immediate pre-synthesis predictions. The prediction models estimate key FPGA metrics, including the usage of Block RAM (BRAM), Digital Signal Processors (DSP), Flip-Flops (FF), and Look-Up Tables (LUT), as well as the inference clock cycles. Evaluation on both synthetic and benchmark NN architectures demonstrates high prediction accuracy, with R^2 scores between 0.8 and 0.98, and sMAPE values ranging from 10% to 30%.

This presentation will focus on introducing rule4ml, showcasing how this tool allows immediate assessment of the feasibility and performance of NNs on FPGAs. We will also explore the data generation and the regression models' training and validation, presenting the predictive performance and the current limitations of our approach as well as potential future improvements. By providing these insights, we aim to show how rule4ml can significantly streamline the deployment of ML models in real-time applications, ultimately reducing development time and enhancing productivity in workflows that rely on NN-to-FPGA frameworks such as hls4ml.

Focus areas

Primary author: Mr RAHIMIFAR, Mohammad Mehdi (University of Sherbrooke)

Co-authors: Prof. CORBEIL THERRIEN, Audrey (University of Sherbrooke); EZZAOUI RAHALI, Hamza (University of Sherbrooke)

Presenter: EZZAOUI RAHALI, Hamza (University of Sherbrooke)

Session Classification: Contributed talks