Contribution ID: **112** Type: **Poster**

Scaling CNN Deployment on FPGAs: The path to 100k parameters with hls4ml

Deploying large CNNs on resource-constrained hardware such as FPGAs poses significant challenges, particularly in balancing high throughput with limited resources and power consumption. To address these challenges, hls4ml was leveraged to accelerate inference through a streaming architecture, in contrast to programmable engines with dedicated instruction sets commonly used to scale to accommodate large neural networks. Despite hls4ml's support for resource-efficient designs and layer-wise streaming via the Vivado HLS compiler, achieving a 100k-parameter CNN deployment remained a significant hurdle. This work details the key advancements implemented in hls4ml to overcome these barriers, with the long-term goal of scaling to 1M parameters. Our approach focuses on three core areas: migrating to the Vitis HLS compiler for enhanced performance, extending the separable convolutional layer implementation to reduce resource demands, and integrating the built-in FIFO depth optimization of Vitis HLS into the hls4ml workflow to efficiently manage resources while maintaining low latency and power efficiency.

Focus areas

Authors: GHIELMETTI, Nicolo (CERN); SUMMERS, Sioni Paris (CERN); TZELEPIS, Stylianos (National Technical Univ. of Athens (GR))

Presenter: TZELEPIS, Stylianos (National Technical Univ. of Athens (GR))