Contribution ID: **103** Type: **Standard 15 min talk**

[Remote] Machine Learning Inference on FPGAs Using HLS4ML with oneAPI Backend

Thursday 17 October 2024 14:10 (20 minutes)

The increasing demand for efficient machine learning (ML) acceleration has intensified the need for user-friendly yet flexible solutions, particularly for edge computing. Field Programmable Gate Arrays (FPGAs), with their high configurability and low-latency processing, offer a compelling platform for this challenge. Our presentation gives update to an end-to-end ML acceleration flow utilizing the oneAPI backend for the HLS4ML compiler to translate models from open-source frameworks such as Keras and PyTorch into FPGA-ready kernels. These kernels, once synthesized, generate optimized bitstreams that implement core ML operations such as layers, activation functions, and normalization, orchestrated by the host for real-time inference. The key challenge of optimizing ML inference on FPGA lies in the architectural differences compared to traditional CPUs and GPUs. Our approach leverages domain-specific optimizations, including pipelined kernel execution, input streaming, fine-grained parallelism control, and improved memory organization. These techniques are critical to achieving superior results in terms of reduced resource utilization, higher maximum clock frequency (fMAX), and lower latency, as demonstrated in synthesis reports targeting Agilex™ FPGAs. Though hardware-based benchmarks are still in progress, we will present preliminary performance estimates and sample outputs from the HLS4ML compilation with the oneAPI backend. The goal of this presentation is to introduce how FPGAs can be used effectively for machine learning tasks, focusing on the oneAPI backend for HLS4ML. We aim to show how this approach simplifies the process of running ML models on FPGAs, making it easier for developers to prototype and deploy solutions. By integrating familiar software frameworks with FPGA hardware, this work provides a practical path toward fast ML inference on edge devices.

Focus areas

Primary author: WANG, Haoyan (Intel Corporation)

Presenter: WANG, Haoyan (Intel Corporation)

Session Classification: Contributed talks