Contribution ID: **96** Type: **Standard 15 min talk**

Large Neural Network Partitionning for Distributed Inference on FPGAs

Thursday 17 October 2024 13:25 (15 minutes)

Ultra-high-speed detectors are crucial in scientific and healthcare fields, such as medical imaging, particle accelerators and astrophysics. Consequently, upcoming large dark matter experiments, like the ARGO detector with an anticipated 200 m² detector surface, are generating massive amounts of data across a large quantity of channels that increase hardware, energy and environmental costs. Simultaneously, there are also increasing concerns about cybersecurity for edge devices, such as Internet of Things, which currently cannot detect attacks in real-time while maintaining normal operations. Many of these devices do not have the compute power to host complex cybersecurity algorithms. To address these challenges, future experiments and systems need effective real-time computation on edge devices. Many new approaches utilize machine learning (ML) algorithms at the source to analyze data in real-time on field-programmable gate arrays (FPGAs) using tools like HLS4ML. However, the complexity and size of the models often do not fit on a single FPGA hence the need for a distributed approach across multiple FPGAs.

This work introduces a method to divide and distribute large neural network models across multiple FPGAs for inference. By decomposing the network layer by layer, we address the limitations of fitting expansive models on a single FPGA. When a layer is too large, it can be divided into multiple parallel components. We employ a partitioning tool using rule4ml to accelerate this process, ensuring efficient resource allocation and allowing for low-latency optimization. Alternatively, the method can be applied manually for more customized splitting and distribution. Utilizing a pipelined architecture, we mitigate the network-induced latency between each node. As a proof of concept, we implemented this approach by deploying a fully connected neural network (FCNN) for the MNIST dataset and a convolutional neural network (CNN) for a cybersecurity classifier on five small PYNQ-Z2 boards, handling models with 12k and 73k parameters, respectively. This technique not only accommodates large models but also reduce the model to FPGA tuning, making it ideal for applications requiring fast development cycles.

This presentation will discuss the necessity of a distributed approach for large ML models on FPGAs and detail the methodology to split and distribute a large neural network models across multiple FPGAs, showing a quick demonstration of the process from start to finish on FPGAs. The tested models show latency ranging from milli to microsecond range with no loss of accuracy when compared to inference on CPU. Finally, we will discuss future directions for scaling this method to accommodate even larger models and more complex neural network architectures.

Focus areas

Primary author: GRANGER, Charles-Étienne (Université de Sherbrooke)

Co-authors: Prof. CORBEIL THERRIEN, Audrey (Université de Sherbrooke); Mr ZAMAÏ, Enzo (Université de Sherbrooke); EZZAOUI RAHALI, Hamza (University of Sherbrooke); RAHIMIFAR, Mohammad Mehdi (University of Sherbrooke)

Presenter: GRANGER, Charles-Étienne (Université de Sherbrooke)

Session Classification: Contributed talks