Contribution ID: **9** Type: **Standard Talk**

Reconfigurable Fused and Branched CNN Accelerator

Wednesday 27 September 2023 14:45 (15 minutes)

There has been a growing trend of Multi-Modal AI models capable of gathering data from multiple sensor modalities (cameras, lidars, radars, etc.) and processing it to give more comprehensive output and predictions. Neural Network models, such as Transformers, Convolutional neural networks (CNNs), etc., exhibit the property to process data from multiple modalities and have enhanced various applications, ranging from consumer devices, medical equipment, and safety-critical systems.

CNNs have shown remarkable performance, especially in vision-based applications, ranging from performing one classification task to more intricate and extensive tasks involving multiple modalities and sub-tasks. They do so by learning the low and high-level features of the images. Most images have common lower-level features, which are learned by the lower layers of the network. As we advance deeper into the network, the layers acquire higher-level or more abstract features.

The proposed methodology harnesses the fundamental capabilities of CNNs to learn patterns and perform multiple un-correlated tasks (radar hand gestures, modified MNIST, SVHN) using a single CNN accelerator. In this way, various tasks can share all the CNN layers (fused model) or some layers (branched model) and maintain, on average more than 90\% accuracy. In the hls4ml-generated accelerator, sharing layers translates to sharing hardware resources. Thus, the suggested approach leads to considerable savings in hardware resources and energy, which would otherwise require separate accelerators for separate tasks. Two architectures are proposed. 1) Fused Model (FM): All the tasks share all the layers, and the task-specific classes get activated in the last layer of the model. 2) Branched Model (BM): It consists of tasks-specific branches and shares only specific layers, and supports sub-tasks classification.

Due to the varying AI requirements and workload, hardware resource utilization and energy budget reach a threshold quickly. The proposed approach is further leveraged to introduce a reconfigurable CNN accelerator that adapts to the application's needs. Three identical instances of an FM/BM accelerator can be configured in a Fault Tolerant mode (high-reliability mode consisting of TMR design), High performance (parallel processing of multiple tasks to deliver maximum performance), and De-Stress (switching off one or more accelerator instances by clock/power-gating to reduce aging and power consumption) mode. This work forms the basis for a fully reconfigurable AI processing system comprising reconfigurable quad-core RISC-V cores, on-chip sensors, reconfigurable AI accelerators, and reconfigurable hardware (i.e., FPGAs).

Author: Mr SYED, Rizwan Tariq (IHP - Leibniz-Institut für innovative Mikroelektronik)

Co-authors: Dr ANDJELKOVIC, Marko (IHP - Leibniz-Institut für innovative Mikroelektronik); Dr ULBRICHT, Markus (IHP - Leibniz-Institut für innovative Mikroelektronik); Prof. KRSTIC, Milos (IHP - Leibniz-Institut für innovative Mikroelektronik)

Presenters: Mr SYED, Rizwan Tariq (IHP - Leibniz-Institut für innovative Mikroelektronik); Dr ANDJELKOVIC, Marko (IHP - Leibniz-Institut für innovative Mikroelektronik); Dr ULBRICHT, Markus (IHP - Leibniz-Institut für innovative Mikroelektronik); Prof. KRSTIC, Milos (IHP - Leibniz-Institut für innovative Mikroelektronik)

Session Classification: Contributed Talks

Track Classification: Contributed Talks