Contribution ID: **121** Type: **Standard 15 min talk**

Bit-Width Optimization of Power-Efficient Hardware Accelerators for Neural Networks using Catapult AI NN

Wednesday 16 October 2024 17:45 (15 minutes)

Nowadays, the application of neural networks (NNs) has expanded across different industries (e.g., autonomous vehicles, manufacturing, natural-language processing, etc.) due to their improved accuracy results. This was made possible because of the increased complexity of these networks which requires higher computational efforts and memory consumption. As a result, there is more demand for specialized NN hardware accelerators that can be used for efficient inference tasks, especially on resource-constrained edge devices (e.g., wearable devices). NNs are typically modeled and trained using high-level languages like Python. To implement NNs in hardware platforms such as field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs), the Python-level code needs to be translated into register-transfer level (RTL) designs. This code transformation requires significant human effort and strong hardware expertise that can be challenging, especially when the NN architecture is not fixed. Catapult AI NN, an extension to the HLS4ML open-source project developed by Fermilab, converts the NN Python code into synthesizable C++ which after Catapult' s high level synthesis downstream flow, produces the RTL code at the end. Compared to other backends of HLS4ML (e.g., Vivado), Catapult allows designers to target ASIC platforms. By automating this transformation, designers can save their time and focus more on tuning the hardware related parameters (such as the amount of parallelism) to explore the design space and obtain the most optimal design for power, performance and area (PPA) in shorter time. We present a case study using Catapult AI NN to synthesize the design. When converting the floating-point data types into bit-level fixed point representation through Quantized Aware Training, value range analysis is performed to validate that optimal bit-widths are chosen and no overflow or saturation errors are present.

Focus areas

Authors: VAEZ TORSHIZI, Marzieh (Siemens EDA); CONDRAT, Christopher (Siemens EDA); BURNETTE, David (Siemens EDA)

Presenter: VAEZ TORSHIZI, Marzieh (Siemens EDA)

Session Classification: Contributed talks