Assignment 1 – Paper Review

Paper Title

A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim

Authors

Ananda Samajdar, Jan Moritz Joseph, Yuhao Zhu, Paul Whatmough, Matthew Mattina, Tushar Krishna

PC Member

NA

Evaluation

Novelty: Some Novelty

Quality: Good

Repeatability: Yes

Presentation: Yes

Overall Evaluation: I fully champion and believe this is in top 10% of papers in KDD

Reviewer's Confidence: I have read key papers in the area

Rank: No

Strengths:

1. Cycle accurate systolic array simulator with configurable architectural parameters and can handle various neural network topologies.
2. Introduces three dataflow techniques that can help study memory access patterns and cache behavior, and search for the best architectural configuration.
3. Analytical timing model for all the dataflows as well as scale-out and scale-up systems.

Weaknesses:

1. The authors do not yet model the complex function unit present in several systolic array based commercial AI accelerator chips (e.g. Google TPU, IBM AIU). The complex function unit handles activation functions like ReLU, softmax, and sigmoid etc. It also evaluates other operations like batch normalization and max pooling, that are important components of a neural network architecture and should be modeled in a simulator.
2. The tool introduced by the authors does not model the memory system fully to consider the effects of low memory bandwidth when fetching data from external memory sources. This is a major issue with today's AI chips that needs to be further investigated and researched.

Detailed Review:

The paper titled "A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim" introduces a cycle accurate systolic array simulator to gather statistics and traces regarding running neural network topologies on a systolic array.

Systolic arrays are ubiquitous in present AI accelerator chips so it is important to have tools that can help judge what designs are important and efficient when running a deep neural network.

The authors also introduce three important dataflow techniques (weight, input, and output stationary) where the term "stationary" pertains to the matrices that spend most time in the systolic array. These dataflow patterns affect the frequency of memory access and are an important parameter to consider while determining the optimal design for an accelerator.

The authors also introduce an analytical timing model that can calculate the number of cycles a given neural network will take to execute on a given hardware configuration. Although the analytical model has been formulated very precisely, there is an underlying assumption that the execution is stall-free which might not be the case in the real world due to memory bandwidth limitations.

Overall, it's a very innovative paper, with substantive novel work particularly with the introduction of dataflow techniques and the timing model and is worthy of this conference.

Suggestions:

1. It would be useful to include an option to enable the complex function unit in the systolic array design as that will better help to understand the performance implications of the hardware on the execution of neural network topologies. It will make the work more "complete".
2. More work is required in modeling memory bandwidth limitations when interacting with external memory sources like HBMs and DRAMs.
3. More detailed insight into how operand, prefetch, and demand matrices were generated in the code. These details should be added to the Implementation Details or the Appendix section of the paper.

Confidential remarks for the PC:

NA