[Extended Abstract] Benchmarking AI-methods on Heterogeneous Hardware Resources

Christopher Noel Hesse
christopher.hesse@aptiv.com
Universität Hildesheim, Germany

Holger Eichelberger
eichelberger@sse.uni-hildesheim.de
University of Hildesheim, Germany

Artificial intelligence (AI) is considered as a key enabling technology to address various challenging problems, like steering self-driving cars or acting as intelligent opponents in complex computer games. The increasing capabilities of AI also requires more powerful compute resources, e.g., for training neural networks, graphical processing units (GPU) are utilized as they outperform traditional processor architectures (CPU). In recent time, also further hardware architectures such as tensor processing units (TPU) are applied, in particular to neural networks. Moreover, rewirable processors like field-programmable gate arrays (FPGA) seem to provide a good tradeoff between performance and energy use. While GPUs and TPUs can be programmed through specific software libraries, FPGA-programming is more complex as it requires rather deep hardware knowledge as well as a different development approach. For developing AI-enabled hardware-accelerated applications, which in the extreme case even rely on several hardware architectures, the question arises, which of these architectures shall be applied for which AI method to achieve the best performance. Besides performance measures such as throughput, intensity or latency, also the energy consumption is relevant, in particular to trade-off AI benefits and increasing AI usage with their impact on the environment. Ultimately, also development and integration efforts must be considered.

The HAISEM-lab project (HAISEM=Hardware-optimized AI Applications using modern Software Engineering Methods) [1] is a BMBF-funded AI-lab, in which the University of Hannover and the University of Hildesheim work on the borderlines of AI, software engineering and hardware-acceleration. HAISEM-lab offers (industrial) training courses on these topics and performs research activities at these borderlines.

As part of our work, we are evaluating the performance as well as the energy consumption of such hardware architectures, for both, standalone AI usage as well as hybrid AI-based software systems. Some related evaluations have been published, but usually with different application area, focus or coverage of the hardware architectures. For example, Jouppi et al. compare in [4] compute and energy performance of CPU, GPU and TPU processing for three neural network applications realized in TensorFlow. Patterson [3] analyzes the compute performance of convolutional neural networks on GPU and FPGA architectures. Qasaimeh et al. report in [2] on a performance and energy evaluation of open source computer vision kernels for CPU, GPU and FPGA architectures. Moreover, there are several (industrial) macro-benchmarking suites such as MLPerf [5], cnn-benchmarks [6], Yolo [7], or DeepSpeech [8] as well as the micro-benchmark mixbench [9]. However, these approaches either do not cover all mentioned hardware resources or they impose limitations, e.g., through the programming approach or the considered type of computational loads.

To perform a systematic comparison among the major types of programmable hardware acceleration for AI (there are also application-specific chips that are out of our scope), we are currently creating a micro-benchmarking approach based on the building blocks of modern AI methods (e.g., convolutions) that can be executed on the mentioned hardware architectures in comparable manner. As the required hardware is installed on the premises of both partners, also energy measurements are possible. Where feasible, we plan to compare the results delivered by existing benchmarks to our approach. In our talk, we will motivate the problem and the need for such hardware. We will introduce our approach to combined software/hardware-benchmarking and present experiences as well as preliminary results.

1 References

[1] http://haisem-lab.de/
[2] M. Qasaimeh, K. Denolfy, J. Loy, K. Vissersy, J. Zambreno, and P. H. Jones, Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels, ICESS'19, 1-8, 2019
[3] L. Patterson, Convolutional Neural Networks on FPGA and GPU on the Edge: A Comparison, Upp-

sala University, 2020
[4] N. P. Jouppi, C. Young, N. Patil, et al., In-Datacenter Performance Analysis of a Tensor Processing Unit, ICSA'17, 1-2, 2017
[5] https://mlperf.org/
[6]https://github.com/jcjohnson/cnn-benchmarks
[7] https://github.com/pjreddie/darknet
[8] https://github.com/mozilla/STT
[9] https://github.com/ekondis/mixbench