TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Electrical and Computer Engineering Vodafone Chair Mobile Communications Systems

Energy Analysis of Accelerators for Deep Neural Networks

Studienarbeit/Project Work

Problem Statement

Deep Neural Networks (DNNs) have become popular in recent years due to their promising results in various application areas. Dedicated hardware accelerators have been developed to exploit the potential of these networks in mobile devices as well [1]. The accelerators mainly consist of Multiply-Accumulate (MAC) units [2]. Moreover, some of them support precision scalable operands as they do not significantly reduce the accuracy of DNNs [3]. However, the energy analysis needs to be extended from a single MAC unit and arrays [4, 5] to complete systems including dispatchers and memories. Your task is to extend an existing set of MAC implementations and memories. In addition, you should set up automated energy analysis after synthesis and/or place&route using state-of-the-art design tools. The result should be a design space exploration of serial and parallel computing hardware designs.

Tasks

- Learning about precision-scalable MAC units and DNN accelerator fundamentals
- Automation of RTL design flow (implementation, simulation, synthesis and place&route)
- Analysis and written documentation of the results in German or English

Expected Skills

- Working with tools in a Linux command line environment (incl. scripting, tcl)
- Experience in hardware design and its description languages (Verilog, VHDL or Chisel/Scala)

Contact Person

Simon Friedrich, simon.friedrich@tu-dresden.de

Please include a recent transcript of records when contacting.

Recommended References

[1] E. Talpes et al., "Compute Solution for Tesla's Full Self-Driving Computer," in IEEE Micro, vol. 40, no. 2, pp. 25-35, 1 March-April 2020, doi: 10.1109/MM.2020.2975764.
[2] V. Sze et al. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.
[3] S. Sharify et al., "Loom: Exploiting Weight and Activation Precisions to Accelerate Convolutional Neural Networks," 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018, pp. 1-6, doi: 10.1109/DAC.2018.8465915.
[4] V. Camus et al., "Review and Benchmarking of Precision-Scalable Multiply-Accumulate Unit Architectures for Embedded Neural-Network Processing," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 4, pp. 697-711, Dec. 2019, doi: 10.1109/JETCAS.2019.2950386.
[5] E. M. Ibrahim et al., "Taxonomy and Benchmarking of Precision-Scalable MAC Arrays Under Enhanced DNN Dataflow Representation," in IEEE Transactions on Circuits and Systems I: Regular Papers, doi: 10.1109/TCSI.2022.3141519.