TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Electrical and Computer Engineering Vodafone Chair Mobile Communications Systems

Latency Reduction of Thread Changes of Layer-Fusion of a Neural Network Accelerator

Studienarbeit/Project Work

Problem Statement

In recent years, many accelerators have been developed to make the execution of Deep Neural Networks (DNNs) on edge devices possible. Due to the limited on-chip memory of these systems, several techniques for increasing the data locality have been introduced in the literature. Hereby, layer-fusion reduces the amount of off-chip transfer by directly re-using intermediate results instead of subsequently executing the network layers However, this method introduces additional changes between layers, called thread changes. Each change has to be handled by the control module of an existing accelerator architecture and introduces additional latency. Therefore, the task of this project work is to accelerate the thread changes by reducing the latency due to the integration of a dedicated hardware module.

Tasks

- Familiarization with dataflows and layer fusion for DNN accelerators
- Analysis of the latency of thread changes and layer-fusion
- Design and implementation of a concept for accelerating the thread changes
- Integration of designed hardware module in existing accelerator structure
- Evaluation of the speed-up and the costs of the implementation

Expected Skills

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

Contact Person

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

Please include a recent transcript of records when contacting.

Recommended References

- Manoj Alwani, Han Chen, Michael Ferdman, and Peter Milder. 2016. Fused-layer CNN accelerators. In The 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-49). IEEE Press, Article 22, 1–12.
- S. Friedrich, S. Balamuthu Sampath, R. Wittig, M. Rohit Vemparala, N. Fasfous, E. Matúš, W. Stechele and G. Fettweis, “Lightweight Instruction Set for Flexible Dilated Convolutions and Mixed-Precision Operands,” in Proceedings of 24th International Symposium on Quality Electronic Design (ISQED 2023), San Francisco, USA, Apr 2023.
- S. Friedrich, R. Wittig, E. Matúš and G. Fettweis, “Energy-based Optimization for Resource Limited Neural Network Accelerators with Fused-Layer Support,” in Proceedings of 3 rd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2021), Porto, Portugal, (pp. 41-45), Nov 2021.