Chair of Integrated Systems
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Master Thesis Topic
Advisor: Nael Fasfous

Flexible On-Chip Networks for Convolutional Neural Network Accelerators

Topic Description

Convolutional Neural Networks (CNNs) have become the state-of-the-art for many computer vision tasks. Their highly parallel computation graph structure offers many optimization possibilities on hardware. The major part of CNN execution can be formulated as a nested for-loop of a core instruction, the Multiply-Accumulate operation. These loops which iterate over spatial dimensions, input channels, filters, and batches allow for many scheduling schemes [1]. A schedule for a CNN execution can involve single or multiple dataflows, which manipulate the unrolling of the for-loops over the available computation elements [2,3]. In order to support schedules with dynamic dataflows, an efficient communication infrastructure can be exploited.

Due to the deterministic nature of CNN execution, the communication does not necessarily require complex routing algorithms typically supported by Network-on-Chips (NoCs). A light and efficient on-chip interconnect can be realized, supporting broadcast, multicast and unicast transfers between the memory and the processing elements.

[1] A. Parashar, P. Raina, Y. S. Shao, et al., “Timeloop: A Systematic Approach to DNN Accelerator Evaluation,” in ISPASS, 2019
[2] M. Gao, X. Yang, J. Pu, et al., “Tangram: Optimized Coarse-Grained Dataflow for Scalable NN Accelerators,” in ASPLOS, 2019
[3] Y. H. Chen, J. Emer and V. Sze, “Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks,” in ISCA, 2016

Prerequisites

To successfully complete this project, you should have the following skills and experiences:

- Very good programming skills in VHDL
- Good knowledge of neural networks, particularly convolutional neural networks

The student is expected to be highly motivated and independent.

By completing this project, you will be able to:

- Understand the execution sequence of CNNs on spatial architectures
- Test the impact of CNN dataflows on hardware accelerators
- Evaluate the complexity of communication hardware against efficiency and latency

Contact

Nael Fasfous
Department of Electrical and Computer Engineering
Chair of Integrated Systems
Arcisstr. 21, 80333 Munich, Germany
Phone: +49.89.289.23858
Email: nael.fasfous@tum.de

This project is in cooperation with BMW AG.