December 2020

CES Chair for Embedded Systems
Prof. Dr. J. Henkel

Master Thesis

Analytical Modeling of a Machine Learning Algorithm

Convolutional neural networks (**CNNs**) have achieved remarkable success in applications such as image classification and object detection. These networks have specifications such as massive parallelism and being MAC dominated that make them **suitable for custom acceleration in hardware**.
There are several interdependent aspects to consider in order to design a CNN accelerator. In this project, we want to investigate **modeling some of these interdependent aspects which affect the off-chip memory access, energy, and performance of the system**. These modeling will help a designer to know how changing each of the aspects will affect the specifications of the accelerator.

A convolution layer within CNNs consists of the convolution of the input feature map (IFM) with filter weights (FW) to compute the output feature map (OFM). Since these layers have a large amount of data, the data needs to be shuttled between on-chip memory and off-chip memory in the form of tiles of data chunks. The tile size for each kind of data and the order of the computation will affect the off-chip memory access volume. Thus the tile size and order of the computation are two of the aspects that need to be considered.
Another aspect to consider is the **shape of the computing array**. The CNN accelerators mainly have a 2D compute array which is responsible for computations. This computing array consists of several compute elements. Different array sizes may lead to different resource utilization for different layers and CNNs. Therefore, the size of the array is another aspect that will affect performance and off-chip memory access.

In this project, an analytical model of these interdependent aspects and their effects on the performance, energy, and off-chip memory access shall be investigated.

Skills Required with the thesis

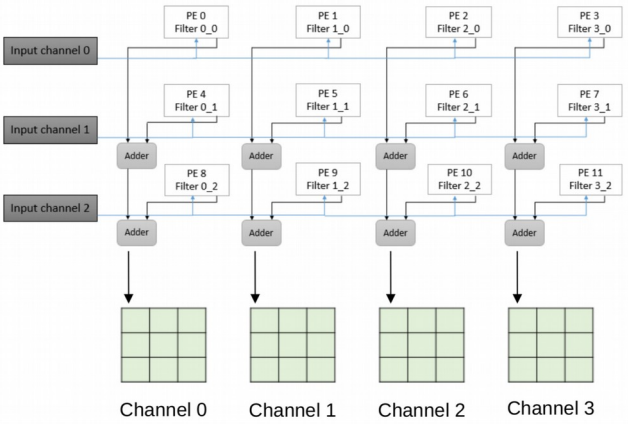
- Good knowledge of programming

Helpful Knowledge

- Experience with hardware design

Skills Required with the thesis

- Good understanding of architecture of the convolutional neural networks
- Understanding of analytical Modeling



Contact:
Faeze Faghih, faeze.faghih@kit.edu
Dr.-Ing. Lars Bauer, lars.bauer@kit.edu

Prof. Dr. Jörg Henkel – Chair for Embedded Systems
Department of Computer Science
KIT Karlsruhe
Haid-und-Neu-Str. 7, Gebäude 07.21

Karlsruhe Institute of Technology