Master Thesis

NoC-based methodology for designing and simulating neural network accelerator

Motivation

In recent years, artificial neural networks (ANNs) have become the foundation for many artificial intelligence applications, such as pattern recognition, prediction, and control optimization. However, ANN requires longer computing time and larger computing power. Thus, it is necessary to design a hardware efficient ANN accelerator for computing such applications efficiently. Because of intensive computation and communication between different neurons, the interconnection between neurons could become complicated as the size of ANNs increases. Network-on-Chip (NoC) has been proposed as a suitable solution for providing efficient communication between neurons, as the NoC is a highly scalable, bandwidth-efficient, and a packet-switched network containing a certain number of routers and links [1].

The NoC-based neural network accelerator receives much attention recently because it can supply power efficiency, computational flexibility, and reconfigurability. Moreover, the NoC-based design methodology decouples the ANN operation into computation and data transmission. Regarding the computation part, different ANN computing models can be performed independently of the data flow. Regarding the data transmission part, the NoC interconnection can efficiently process various data flows for different ANN computing models [2].

In normal ANN models, only fully-connected layers exist. Comparing to ANN, a common convolutional neural network (CNN) model can have several convolution layers, pooling layers, and fully-connected layers. However, the existing NN-Noxim simulator (https://sites.google.com/site/cereslaben/links) only supports the simulation of ANN models. It cannot simulate the CNN models. Besides, technology scaling could increase the susceptibility of NoCs components to faults [3].

Goal

The goal of this thesis contains the following parts:

1. Study the existing NN-Noxim simulator with a focus on grouping neurons and mapping neurons to the processing elements.
2. Based on the achieved knowledge of grouping and mapping, propose how to perform the simulation of CNN models on NoC-based many-core on-chip systems. Extend the Noxim simulator to support the simulation of CNN models. In this thesis, LeNet and VGG-16 need to be evaluated.
3. Taking router faults into account, some packets may not arrive at their destinations. The dropped packets affect the accuracy of the simulated neural network. Improve the existing grouping and mapping algorithms concerning path connectivity between routers under consideration of faults. The suggested method is to use table-based routing, which computes the minimal route for a source and destination pair and stores the route information in a table.
4. Evaluate the performance of simulated models concerning classification precision and simulation latency under fault-free and various fault patterns.

Bibliography

[1] Kwon H, Samajdar A, Krishna T. Rethinking nocs for spatial neural network accelerators. In2017 Eleventh IEEE/ACM International Symposium on Networks-on-Chip (NOCS) 2017 Oct 19 (pp. 1-8). IEEE

[2] Chen KC, Wang TY. Nn-noxim: High-level cycle-accurate noc-based neural networks simulator. In2018 11th International Workshop on Network on Chip Architectures (NoCArc) 2018 Oct 20 (pp. 1-5). IEEE.

[3] Radetzki, M., Feng, C., Zhao, X. and Jantsch, A., 2013. Methods for fault tolerance in networks-on-chip. *ACM Computing Surveys (CSUR)*, *46*(1), pp.1-38.