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 modules [2].

Possible directions

1. In normal ANN models, only fully-connected layers exist. Comparing to ANN, a common CNN module 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. Extend it to support different kinds of CNN models.
2. Technology scaling could increase the susceptibility of NoCs components to faults. Redundancy is the main approach to provide fault tolerance. Different redundancy forms are used to tolerating different fault classes [2]. Applying some fault-tolerant mechanisms i.e. fault-tolerant routing algorithms or adding spatial redundancy (e.g. adding diagonal links to mesh) to bypass faulty routers. Thus, it is necessary to take into account faults and fault-tolerant mechanisms when designing a NoC-based ANN accelerator.
3. In the NoC-based neural network accelerator, processing elements perform the neuron operations. Mapping algorithms define how neurons are mapped to the processing elements that are connected to routers on a NoC. In general, computation of a single neuron is simple, a cost-efficient solution is to cluster multiple neurons to a PE. Thus, it is a challenging task to map and clustering different neurons from different layers to processing elements under the simulated ANN or CNN model, utilized NoC topology, NoC size, and consideration of router faults. Study and propose some mapping and clustering algorithms is another research direction.

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.