Presentation Abstract

A Compiler Overview for A Deep Neural Network Accelerator with Software-controlled Memory Hierarchies

Kazuaki Ishizaki[1,a)] Eri Ogawa[1] Hiroshi Inoue[1] Swagath Venkataramani[2]
Jintao Zhang[2] Wei Wang[2] Vijayalakshmi Srinivasan[2]
Moriyoshi Ohara[1] Kailash Gopalakrishnan[2]

Presented: March 13, 2020

This presentation gives the design and implementation of a compiler for a deep neural network accelerator that provides high performance and power efficiency. The accelerator consists of multiple heterogeneous units and each unit has a limited set of instructions. Also it does not have hardware-controlled caches for power efficiency. It is very hard for programmers to directly write a program for such an accelerator. Thus, we develop a compiler that automatically generates native code for each unit from a program in deep learning frameworks such as TensorFlow. This compiler generates outer loops around highly-tuned hand-crafted inner kernel loops for each unit with a wide range of neural network parameters. It effectively supports various types of parameters such as zero padding sizes around an image. This compiler also generates data transfer code between hierarchical memories, and applies code optimizations to reduce the code size.

[1] IBM Research - Tokyo, Chuo, Tokyo 103–8501, Japan
[2] IBM Research - Watson Research Center, Yorktown Heights, NY 10598
[a)] ishizaki@jp.ibm.com