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

Assistant (Student)

Compressed Representations of Convolutional Neural Networks for Data Movement Optimization

Convolutional Neural Networks (CNNs) have become the state of the art in image classification and other computer vision tasks. This has led to a substantial effort from industry and academia, to bring such neural networks to edge devices. With tight area, power and latency constraints, this challenge presents many optimization opportunities.

Data movement optimization has been of high interest in the field of CNN accelerator design. As it accounts for a significant portion of the total power consumption of the system, researchers have used loop blocking methods and tailored dataflows to maximize the reuse of every piece of data read from a memory higher up in the hierarchy.

Goals

Smaller chunks of an entire neural network can be compressed to lower dimensionality. This representation can serve the purpose of reducing the size of transactions from off-chip memory to on-chip memory. Accessing off-chip memory can cost orders of magnitude more energy than on-chip memory requests.

The goal is to find the sweet spot between the loss in accuracy due to dimensionality reduction and the power consumption improvement brought about by the reduced data movement between memory hierarchies.

Prerequisites

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

- Very good programming skills in Python and Tensorflow
- Good knowledge of neural networks, particularly convolutional neural networks

The student is expected to be highly motivated and independent.

Learning Objectives

By completing this project, you will be able to:

- Find accurate, compressed representations of neural networks
- Analyze the effects of data movement on energy efficiency
- Test and evaluate compression and expansion methods
- Present your work in the form of a scientific report

Contact

Nael Fasfous
Department of Electrical and Computer Engineering
Chair of Integrated Systems
Arcisstr. 21
80333 Munich
Germany

Phone: +49.89.289.23858
Building: N1 (Theresienstr. 90)
Room: N2116
Email: nael.fasfous@tum.de

This project is in cooperation with BMW AG.

Advisors

Nael Al-Fasfous