Invited Talk II
13:40–14:30, Monday, March 29, 2021

Ingredients of Efficient Hardware Accelerators for Neural Networks

Prof. Juinn-Dar Huang
National Yang Ming Chiao Tung University, Taiwan

Abstract

Today, neural networks (NNs) are broadly used for numerous artificial intelligence (AI) applications including computer vision, image/video processing, speech recognition, and natural language processing (NLP). Though NN-based algorithms can provide better solutions on several AI application domains, those advantages come at the cost of extremely high computational complexity. Currently, GPU-based computing engines are most commonly used platforms for NN computation. Nevertheless, they are pricey, power-hungry and therefore inappropriate for certain application areas, such as edge computing. Therefore, specialized hardware accelerators optimized for a specific class of NNs are getting more attention these days. This short talk aims to introduce some common ingredients of efficient NN hardware accelerators, including dataflow-based optimization, weight pruning and compression, quantization, and numeric data formats. Algorithm-level design considerations facilitating efficient hardware implementations are also discussed. Accelerators for convolutional neural networks (CNNs) and multilayer perceptron (MLP) are used as examples for demonstrations.