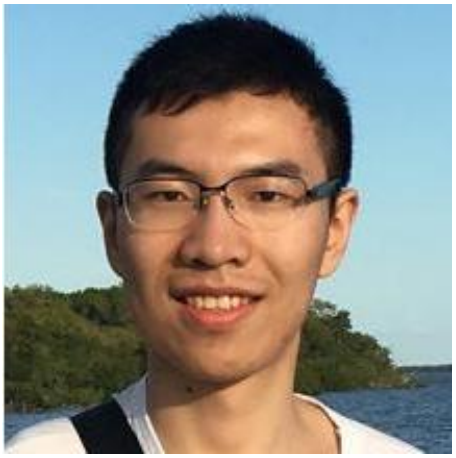
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



Xiaowei Wang

Efficient, Reconfigurable, and QoS-Aware Systems for Deep Neural Networks

Thursday, September 8, 2022
2:00 – 4:00pm
Virtual – Zoom

ABSTRACT: Deep Neural Networks (DNN) are an important machine learning application which has high compute and memory bandwidth requirements to the underlying computer systems. Prior works have proposed domain specific accelerators for DNNs, and in-memory computing architectures are especially promising as they can provide high memory bandwidth and computing throughput at the same time.

The first part of the dissertation targets improving the efficiency of a recent SRAM-based in-memory DNN accelerator on sparse and low-bitwidth DNN models. In this system, the last level cache of a CPU is repurposed as a highly parallel bit-serial vector processor, achieved by the circuit modification to the SRAM array peripherals and a transposed data mapping. We propose two hardware/software co-design methods with customized model pruning algorithms to fully utilize the sparsity. A specialized bit-serial algorithm is developed for the operations on low bitwidth data.

While the above in-cache computing system is highly efficient, its hardware architecture lacks the flexibility to be optimally reconfigured for different DNN models. The second part of the dissertation proposes a reconfigurable in-SRAM computing DNN accelerator based on block RAMs (BRAM) on FPGAs. We propose circuit changes to the BRAM to enable bit-serial in-memory computing, which turns BRAMs as both bit-serial vector units and data storage. Building on the compute-capable BRAMs, we further propose customized accelerator instances for different DNN models, which outperforms a state-of-the-art DNN accelerator on FPGA.

DNN workloads can also be run on general purpose CPUs in datacenters. Cache compression is a technique to reduce the cache miss rate on CPU, which benefits DNNs as well as many other applications. In the third part of the dissertation, we present a novel method to compress cache data with efficient in-SRAM data comparison. Further, as datacenters frequently collocate multiple workloads to increase server utilization, the Quality-of-Service (QoS) of DNN workloads, such as latency, can be affected. The final part of the dissertation proposes a systematic approach to achieve the QoS of DNNs under collocation, with resource partition to reduce interference, and a proposed latency prediction model to choose the partition that satisfy the QoS requirement.

CHAIR: Prof. Reetuparna Das