+1-778-847-8633
Vancouver, BC, Canada
dingqingy@ece.ubc.ca

Dingqing Yang

: dingqingy.github.io
: dingqingy
: dingqing-yang-105190137

Education

Doctor of Philosophy **Sept 2018 - May 2024 (expected)**
University of British Columbia - Electrical and Computer Engineering

Bachelor of Applied Science **Sept 2014 - May 2018**
University of British Columbia - Electrical and Computer Engineering

Research Projects and Publications

Algorithm-aware Sparse DNN Training Accelerator Design

- **Dingqing Yang**, Amin Ghasemazar, Xiaowei Ren, Maximilian Golub, Guy Lemieux, and Mieszko Lis. "Procrustes: a Dataflow and Accelerator for Sparse Deep Neural Network Training", **MICRO 2020**. (acceptance rate: 82/424 = 19.3%)
- an algorithm-hardware co-designed approach that first adapts the sparse training algorithm to meet the hardware constraints, and then develops suitable dataflow, data layout, and load balancing schemes to improve the latency and energy efficiency of the accelerator.
- up to 3.26× less energy and up to 4× speedup on a wide range of networks pruned up to an order of magnitude without loss of accuracy.

Side-Channel Attack to Steal Pruned DNNs on Sparse DNN Accelerator

- **Dingqing Yang**, Prashant J. Nair, and Mieszko Lis. "HuffDuff: Stealing Pruned DNNs from Sparse Accelerators", **ASPLOS 2023** (To Appear)
- the first (to the best of our knowledge) bus-snooping-based side-channel attacks to reverse-engineer **pruned** networks executed on **sparse** hardware.
- an automated probing framework that leverages (i) the boundary effect in CNN layers, and (ii) the timing side channel of on-the-fly activation compression to reverse-engineer CNN network architectures.
- significantly reduces the solution space up to 94 orders of magnitude.
- achieves on-par accuracy against the victim under iso-footprint constraints and improves the black-box targeted adversarial attack success rate.

Dynamic DNN Activation Compression through Channel-Clustering

- an activation compression method that dynamically compresses activations by clustering similar channels together.
- activation distributions are often similar across different channels, so compression can be shared across subsets of channels.
- achieves a top-1 accuracy improvement of up to 60% (from 5.99% to 66.11%) with 4-bit quantized activations over state-of-the-art data-free post-training compression methods, while reducing the energy by 35% compared to the 8-bit baselines.
- in preparation for submission.

Teaching Experience

I was a teaching assistant for the following courses:

- CPEN 211 - Computing Systems I
- CPEN 212 - Computing Systems II
- CPEN 291 - Computer Engineering Design Studio I
- CPEN 311 - Digital Systems Design
- CPEN 411 - Computer Architecture
- CPEN 511 - Advanced Computer Architecture

Awards and Honours

- Four Year Doctoral Fellowship (4YF)
- Graduate Support Initiative (GSI) Awards 2020W
- Faculty of Applied Science International Student Scholarship 2015W, 2016W, 2017W
- TREK Excellence Scholarship for Continuing Students 2015W, 2017W
- Dean's Honour List 2014W, 2015W, 2016W, 2017W