赵宪栋(实习)

北京市海淀区中关村 | 18811176157 | zhaoxiandong@ict.ac.cn | https://github.com/hustzxd

教育

直博研究生 计算机体系结构 中国科学院计算技术研究所 2017-2022

本科 计算机科学与技术学院 华中科技大学 2013-2017

研究方向与工作经历

中科院物栖 AI LAB 实习生

- ◊ 面向专用 NPU 的低比特量化与加速[1][p1] 2018-2019
- ◊ 针对 Bit-Serial 类 NPU 的 bit 稀疏加速方法研究[2][p1] 2018-2019
- ◊ 面向专用 NPU 的网络稀疏化方法设计与普适性扩展[p1] 2020

MIPS 指令集五级流水线实现 2016

项目列表

[p1] A PyTorch Framework for Efficient Pruning and Quantization for specialized NPU.
Algorithm: LLSQ, LSQ, BitPruner, ADMM NPU Pruner, ADMM Level Pruner, Sparsity and Quantization (SQ).
[p2] The PyTorch implementation of Learned Step size Quantization (LSQ) in ICLR2020.

论文列表

[1] **Xiandong Zhao**, Ying Wang, Xuyi Cai, Cheng Liu, Lei Zhang. "Linear Symmetric Quantization of Neural Networks for Low-precision Integer Hardware." In International Conference on Learning Representations (ICLR 2020).
[2] **Xiandong Zhao**, Ying Wang, Cheng Liu, Cong Shi, Kaijie Tu, Lei Zhang. "BitPruner: Network Pruning for Bit-serial Accelerators." In Design Automation Conference (DAC 2020)
[3] Yintao He, Ying Wang, **Xiandong Zhao**, Huawei Li and Xiaowei Li. "Towards State-Aware Computation in ReRAM Neural Networks." In Design Automation Conference (DAC 2020).
[4] Songyun Qu, Ying Wang, Bing Li, **Xiandong Zhao** and Lei Zhang. "RaQu: An automatic high-utilization CNN quantization and mapping framework for general-purpose RRAM Accelerator." In Design Automation Conference (DAC 2020).