Lei Wang

LeiWang1999 • leiwang1999@outlook.com • leiblog.wang

Education

University of Chinese Academy of Science • Beijing, China — August 2021 – Present
Masters • *Computer Science*

Nanjing Tech University • Nanjing, China — August 2017 – June 2021
Bachelor • *Electronic Engineering* • **Overall GPA: 3.95/4.0**

Work Experience

Microsoft Research Asia – System Research Intern — April 2022 – Present
Beijing, China

- Advised by Dr. Lingxiao Ma and Dr. Jilong Xue
- Maintaining Microsoft DNN compiler NNFusion .
- Research focus: Sparse Tensor Compilation, Auto Tensorize, LLM inference foundation.

Netease – NPU Development Intern — Sep. 2021 – Oct. 2021
Beijing, China

- NVDLA FPGA deployment and Software stack remapping.

Projects

Ladder **waiting for publication** — 2023

- Tensor Code Generation for Accelerator. obtaining comparable performance with cuBLAS/Cutlass and outperforms tensorrt of tensor core program. Supporting diverse tensor format remapping.

AutoGPTQ.tvm **code** — 2023

- tvm inference kernel for GPTQ.

Full Stack FPGA Implementation of NVDLA — 2021
Code Archive • post:DLA Deploy • post:Compiler Design

- Full-stack FPGA implementation of NVDLA. To enhance the utility of this accelerator, we designed a new compiler and runtime to allow networks auto fallback between CPU and hardware accelerators.

FPGA Accelerator for Beam Forming — 2020
Demo

- FPGA acceleration to enhance sounds from specific points with tetragonal microphone array.

FPGA Accelerator for Digital Recognition — 2020
Demo

- This project aims to provide accelerated digital analysis with lenet5.

Opensource Contributions **LeiWang1999** — -

- gptq-integration to mlc-llm, matrix core support for tvm, general n:m training for apex, etc.

Publications

- Lin Bin*; Zheng Ningxin*; **Wang Lei***; Cao Shijie; Ma Lingxiao; Zhang Quanlu; Zhu Yi; Cao Ting; Xue Jilong; Yang Yuqing; et al. **Efficient GPU Kernels for N: M-Sparse Weights in Deep Learning**. *Proceedings of Machine Learning and Systems*, Vol. 5, 2023. (* represents co-first author) Read Paper
- Sun Xiaotian; Wang Xinyu; Li Wanqian; **Wang Lei**; Han Yinhe; Chen Xiaoming. **PIMCOMP: A Universal Compilation Framework for Crossbar-based PIM DNN Accelerators**. *60th. Design Automation Conference*, 2023. Read Paper
- **Lei Wang**; Lingxiao Ma; Shijie Cao; Ningxin Zheng; Quanlu Zhang. **Ladder: Efficient Tensor Compilation on Customized Data Format**. *17th USENIX Symposium on Operating Systems Design and Implementation (Poster)*, 2023.

Awards

- **2018 Chinese National Scholarship (Top 0.3%)**
- 2021 Excellent New Student Award of Chinese Academy of Science
- **Njtech Person of Year 2020**
- **First Price of 2019 NUEDC (Top 0.5%)**
- Third Price of Integrated Circuit Innovation Competition (FPGA hardware Accelerator for digital recognition)
- Third prize of National FPGA Competition (FPGA based FOSDA Alogrithom Implementation)

Last updated: September, 2023.