Ruobing Han

Email : hanruobing@gatech.edu Cell phone : 404-988-3989

EDUCATION

Georgia Institute of Technology(GT) 2021.5 – Now, USA
PhD , Computer Science (CS)
Research Area: compiler, architecture, ML system
Advisor: Prof. Hyesoon Kim
Peking University(PKU) 2014.9 – 2018.7, China
Bachelor of Science , Computer Science (CS)

PROJECT

Porting CUDA to the x86 architecture (2021.05 – 2021.09)

- Build a framework to execute CUDA source code with the latest features on CPU devices
- Implement a transformer to translate CUDA's SPMD kernels into MPMD+SIMD format based on LLVM
- Improve the coverage from 68% (previous projects) into 90% on CUDA SDK 10.0 samples
- Generate close or even higher performance programs compared with the current SOTA solution.

Porting CUDA to an extended RISC-V GPU architecture (2020.11 – 2021.05)

- Build a framework to execute CUDA source code on a RISC-V GPU
- Implement a translator to translate NVVM IR into SPIR-V
- Support executing unmodified CUDA source code on Vortex (a RISC-V GPU architecture)
- Related paper is shown in the Fifth Workshop on RISC-V for Computer Architecture Research (https://arxiv.org/pdf/2109.00673)

RISC-V GPU development (2021.05-2021.08)

- Write the tutorial for Vortex project
- Implement Cache prefetching for Vortex GPU
- Use SystemVerilog in this project

Low-precision distributed training Neural Network(2019.05 – 2020.06)

- Propose an algorithm to avoid overflow while using low-precision floating-point for gradients
- Build a simulator for Pytorch to use customized floating-point formats
- Use 8 bits floating-point for distributed training ResNet50 and FCN on large scale distributed system
- Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming

Deployment of Pytorch Models (2020.04 – 2020.09)

- Contributor of OpenMMlab project (https://github.com/open-mmlab), which has 20k stars.

- Provide code support convert Detection/Segmentation/Editing models from Pytorch into ONNX
- Support comparing the numerical precision between Pytorch and ONNX automatically.
- Provide TensorRT plugins supports the converted ONNX model

EXPERIENCE

SenseTime Research 2017.11 – 2020.10, Beijing

- **High speed distributed training Neural Network**
  - Propose a sparse communication algorithm and developed LARS algorithm for special cases.
  - Implement Hierarchical communication and smooth-label algorithm.
  - Achieve the baseline accuracy within 1.5 minutes, which was the world record on 2018.11.
  - Proceedings of the IEEE Transactions on Big Data 2020.
- **Deep Learning Operator Compiler**
  - Build Deep Learning Operator Compiler based on TVM and Halide
  - Generate convolution codes 10 times faster than original TVM.

PUBLICATION

Conferences and Workshops

- **Ruobing Han**, Jaewon Lee, Jaewoong Sim, Hyesoon Kim. “COX: CUDA on X86 by Exposing Warp-Level Functions to CPUs“ under review 2021
- **Ruobing Han**, Blaise Tine, Jaewon Lee, Jaewoong Sim, Hyesoon Kim. “Supporting CUDA for an extended RISC-V GPU architecture“ the Fifth Workshop on RISC-V for ComputerArchitecture Research 2021
- **Ruobing Han**, Min Si, James Demmel, Yang You. “Dynamic scaling for low-precision learning“ Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming 2021
- **Ruobing Han**, Yang You and James Demmel. “Auto-Precision Scaling for Distributed Deep Learning“ International Conference on High Performance Computing 2021

Journals

- Peng Sun, Wansen Feng, **Ruobing Han**, Shengen Yan and Yonggang Wen. “Optimizing Network Performance for Distributed Deep Neural Network Training on GPU Clusters: ImageNet/AlexNet Training in 1.5 Minutes.” In IEEE Transactions on Big Data 2020.