Hongyu Zhu

Male, Chinese
Date of Birth: Nov. 21, 1990
E-mail: `serailhydra@gmail.com`

Research Interests

System for Machine Learning, Distributed Systems, Profiling

Education

- **Ph.D** — ***Now***
  University of Toronto, Toronto, ON, Canada
  Supervised by Prof. Gennady Pekhimenko
- **Master** — ***April 2016***
  McGill University, Montreal, QC, Canada
  GPA: 3.94/4.0
- **Bachelor (prestigious permission)** — ***June 2013***
  Shanghai Jiao Tong University
  ACM class
  Advisor: Prof. Yong Yu, Prof. Minyi Guo
  GPA: 86.6/100

Work Experience

Jun. 2018 — Sept. 2018	Research Intern in Microsoft Research, supervised by Amar Phanishayee. Working on performance debugging for DNN computation
Jul. 2014 — Dec. 2014	Research Intern in New York University in Shanghai (NYU-Shanghai), supervised by Prof. Zheng Zhang. Working on Deep Learning Training Platform (Minerva Project)
Jul. 2012 — Feb. 2013	Research Intern in Microsoft Research Asia (MSRA), supervised by Dr. Zheng Zhang, the vice Dean of MSRA, and Dr. Zhengping Qian. Working on Streaming DAG Distributed System (TimeStream project)
Jul. 2011 — Jul. 2013	Research Intern in Embedded Pervasive Computing Center (EPCC), Shanghai Jiao Tong University (SJTU), supervised by Prof. Minyi Guo. Working on GPU communication framework

Projects

- **Daydream: Estimating Efficacy of Performance Optimizations for DNN Training.** The efficacy of software-level optimizations for DNN training can vary significantly when used in different deployment configurations. In this project, we aim to aim to answer predictive questions such as ”How will optimization X affect the performance of my model?”. We achieve this goal by using three key ideas: (i) constructing a kernel-level dependency graph by utilizing vendor-provided profiling tools, while tracking dependencies among concurrently executing tasks; (ii) mapping low-level traces to DNN layers in a synchronization-free manner; (iii) introducing a set of rules for programmers to effectively describe and model different optimizations. This project is a sub-project of Project Fiddle of Microsoft Research. A paper was accepted by USENIX ATC’20.

- **Benchmarking and Analyzing DNN Training:** The goal of this project is to understand the performance bottlecks of training modern DNNs. The project consists of several parts: maintaining a diverse benchmark suite with state-of-the-art DNN models, defining performance metrics, and building tools to extract these metrics. We apply our toolchains to our benchmark suite, gaining some insights for optimizations. A paper has been published in 2018 IEEE International Symposium on Workload Characterization (IISWC18), and the project is still active.

Professional Skills

- Master in Python, familiar with C, C++, Java and C#
- Familiar with GPU profiling tools (e.g. Nvprof, CUPTI, Nsight)
- Familiar with system design & implementations of main-stream distributed deep learning systems
  - Instrumenting MXNet framework for memory profiling
  - Instrumenting PyTorch framework for performance profiling
- Understanding common machine learning and deep learning algorithms
- Adept at algorithms and data structures
  - Competition in Informatics Olympiad about algorithms for 2 years
  - Experience in solving online problem archives (Ural, UVa, etc)

Publications

- Zhu H., Phanishayee A., Pekhimenko G. Daydream: Accurately Estimating the Efficacy of Optimizations for DNN Training. In 2020 USENIX Annual Technical Conference (ATC'20).
- Zhu H., Akrout M., Zheng B., Pelegris A., Phanishayee A., Schroeder B., Pekhimenko G. (2018). Benchmarking and Analyzing Deep Neural Network Training. In IEEE International Symposium on Workload Characterization 2018 (IISWC18).
- El-Sayed N, Zhu H, Schroeder B. Learning from Failure Across Multiple Clusters: A Trace-Driven Approach to Understanding, Predicting, and Mitigating Job Terminations[C]//Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on. IEEE, 2017: 1333-1344.
- Qian, Z., He, Y., Su, C., Wu, Z., Zhu, H., Zhang, T., ... , Zhang, Z. Timestream: Reliable stream computation in the cloud. In Proceedings of the 8th ACM European Conference on Computer Systems, ACM, 2013.

Invited Talks

- Tools and Methodologies for Evaluating Platforms
  A Talk in MLBench Tutorial in ISCA 2019 Jun 2019
- Holistic Approach to DNN Training Efficiency: Analysis and Optimizations
  Guest Lecture for Advanced Computer Architecture course Mar 2019
- Benchmarking and Analyzing DNN Training
  Huawei, Markham, ON, Canada Apr 2018

Service

- Member of the working group of the speech benchmark in MLPerf
- Member of Artifact Evaluation Committee of MLSys 2019