Shang Wang

CONTACT INFORMATION

Phone: +1 (647) 767-0418
E-mail: wangsh46@cs.toronto.edu

EDUCATION

University of Toronto
PhD in Computer Science Sept. 2021 — Present
Computer Systems and Networks Research Group; Vector Institute
- Advisor: Gennady Pekhimenko

MSc in Computer Science Sept. 2018 — Jan. 2020
Computer Systems and Networks Research Group; Vector Institute
- Thesis: *BPPSA: Scaling Back-propagation by Parallel Scan Algorithm*
- Advisor: Gennady Pekhimenko
- Cumulative GPA 4.0/4.0

BASc in Computer Engineering Sept. 2013 — June 2018
- Cumulative GPA 3.98/4.0, Dean's Honours List for all semesters, graduated with High Honours.
- Highest cumulative academic standing in the program (Adel S. Sedra Gold Medal).
- Capstone project: *A Machine Learning Model for Toxic Language Classification*.

RESEARCH INTERESTS

Systems for Machine Learning, Hardware-efficient and Distributed Training Systems and Algorithms

PUBLICATIONS

Shang Wang, Peiming Yang*, Yuxuan Zheng*, Xin Li*, Gennady Pekhimenko. *Horizontally Fused Training Array: An Effective Hardware Utilization Squeezer for Training Novel Deep Learning Models*. Fourth Conference on Machine Learning and Systems (*MLSys'21*). April 2021.

Shang Wang, Yifan Bai, Gennady Pekhimenko. *BPPSA: Scaling Back-propagation by Parallel Scan Algorithm*. Third Conference on Machine Learning and Systems (*MLSys'20*). March 2020.

PROFESSIONAL EXPERIENCE

CentML
Co-founder & Chief Technology Officer July 2022 — Present
- Building CentML from scratch. CentML democratizes machine learning by optimizing the associated hardware efficiency and reducing the consequential cost.
- Leading the research and engineering efforts.

NVIDIA
Senior Software Engineer, Deep Learning July 2022
Software Engineer, Deep Learning Mar. 2020 — June 2022
GPU Compute Architecture Group:
- Analyzed and conducted system optimizations of many machine learning workloads from NVIDIA researchers and engineers, resulting in significant GPU hour savings (>7K per week at one point).
- Contributed to the PyTorch implementation of the BERT model in NVIDIA Deep Learning Examples.
- Created and then led the development of the Language Datasets and Data Loaders (LDDL) library. LDDL reduces the dataset preprocessing latency for BERT pre-training from 24 hours to 3 minutes via multi-node scaling, and accelerates $\text{BERT}_{\text{LARGE}}$ Phase 2 pre-training by $1.7\times$ via sequence binning.
- Contributed to the MLCommons (a.k.a., MLPerf) Benchmark Infra Working Group (representing NVIDIA); assisted in the logging and validation of NVIDIA's MLPerf submissions.

Huawei Technologies Canada
Research Intern May 2019 — Aug. 2019
Interned in Toronto Heterogeneous Compiler Lab:
- Built the support for scalar and tensor intrinsic functions in TVM hybrid script.

Google
Software Developer Intern, Tools and Infrastructure May 2017 — Aug. 2017

Interned in Display Ads Engineering Productivity:
- Built a new, faster and more scalable, multitier debugging tool with a data pipeline.
- The new tool reduced ~3 minutes of run-time and ~3 GB of storage for the integration testing infrastructure per run (launching every 5 minutes).
- Optimized the infrastructure and reduced ~10 GB of peak memory usage per run.

Intel
Intellectual Property Core Engineering Intern Aug. 2016 — May 2017
Interned in PCIe Intellectual Property (HIP) Core team, Programmable Solutions Group:
- Developed the hardware farm testing framework and infrastructure.
- Implemented the dynamic example design generation of PCIe HIP cores on Stratix 10 FPGA device.
- Debugged and fixed issues in the IP cores.

Google
Software Engineering Intern May 2016 — Aug. 2016
Interned in Borglet, Borg and participated in the development of Task Performance Analysis (TPA):
- Implemented a system that enables TPA to correlate task (service) performance anomalies with hardware/platform anomalies.
- Improved the performance and usability of TPA by parallelization and architectural enhancement.

Altera
Software Engineering Intern May 2015 — Aug. 2015
Interned in Platform Software Team, Altera Virtualization Lab:
- Built a cross-quadrant communication system (Tunnel) in the platform software which operates among user and kernel spaces in host machines and the kernel space in virtual machines with routing capability.

University of Toronto
Summer Research Student May 2014 — Aug. 2014
Supervised by Prof. Jonathan Rose:
- Developed an Android app that helps smokers to quit smoking.
- The prototype was presented on 2014 Undergraduate Engineering Research Day.

HONORS AND AWARDS

- ***2023 Meta Research PhD Fellowship Finalist*** (Meta, Apr. 5th, 2023)
- ***2023-2026 Canada Graduate Scholarship-Doctoral (CGS-D)*** (Natural Sciences and Engineering Research Council of Canada (NSERC), Apr. 25th, 2023)
- ***2021-2022 Vector Research Grant*** (Vector Institute, July 28th, 2022)
- ***2022-2023 Ontario Graduate Scholarship*** (Government of Ontario, June 24th, 2022)
- Canadian Artificial Intelligence Association (CAIAC) ***AI Masters Thesis Award*** nominee (Department of Computer Science, University of Toronto, Feb. 20th, 2021)
- ***2019-2020 Vector Research Grant*** (Vector Institute, Mar. 31st, 2020)
- ***Adel S. Sedra Gold Medal*** (Faculty of Applied Science & Engineering, University of Toronto, June 4th, 2018)
- ***Certificate of Distinction*** for Electrical & Computer Engineering (ECE) final year capstone project (Department of ECE, University of Toronto, Apr. 6th, 2018)
- ***ECE Outstanding Student Award*** (Department of ECE, University of Toronto, Sept. 9th, 2015)
- ***University of Toronto Scholar*** (University of Toronto, Aug. 14th, 2015)
- ***First Place*** in Hardware Hackathon (IEEE University of Toronto, Feb. 7th, 2015)
- ***Third Place***, ***Innovation Award***, ***First Place Presentation*** in UTEK Junior Design (University of Toronto Engineering Kompetitions, Jan. 24th - 25th, 2015)
- ***University of Toronto Scholar*** (University of Toronto, Aug. 13th, 2014)
- ***John M. Empey Scholarship*** (Faculty of Applied Science & Engineering, University of Toronto, July 30th, 2014)
- ***Baptie Scholarship*** (Faculty of Applied Science & Engineering, University of Toronto, July 30th, 2014)
- ***University of Toronto Excellence Award*** (University of Toronto, Apr. 2014)
- ***First Place*** among more than 400 participants in the Connect-6 Game AI Competition (Department of ECE, University of Toronto, Nov. 29th, 2013)

TALKS

- *Horizontally Fused Training Array: An Effective Hardware Utilization Squeezer for Training Novel Deep*

Learning Models
  - *MLSys'21* (Apr. 8th, 2021)
- *BPPSA: Scaling Back-propagation by Parallel Scan Algorithm*
  - *MLSys'20* (Mar. 2nd, 2020)
  - NSERC COHESA (Aug. 10th, 2020)
  - Meta (i.e., Facebook) AI (Apr. 5th, 2022)
- *Think Your Models Run Efficiently? Think Again!*
  - Vector's annual AI Summit & Career Fair (Sept. 23rd, 2021)
- *Recent Trend in Machine Learning Compilers: A Survey*
  - 17th Workshop on Compiler Driven Performance, CASCON 2018 (Oct. 31st, 2018)

POSTER PRESENTATIONS

- *Horizontally Fused Training Array: An Effective Hardware Utilization Squeezer for Training Novel Deep Learning Models*
  - *MLSys'21* (Apr. 8th, 2021)
- *BPPSA: Scaling Back-propagation by Parallel Scan Algorithm*
  - *MLSys'20* (Mar. 2nd, 2020)
  - Vector Institute Workshop on Machine Learning Systems (Oct. 26th, 2019)

TEACHING

Teaching Assistant:
- CSC369: Operating Systems (Fall 2018, Winter 2019)
- ECE244: Programming Fundamentals (Fall 2019)

SERVICE

- *MLSys'23* External Review Committee (ERC) & shepherd
- *MLSys'20* Artifact Evaluation Committee
- SysNet Reading Group seminar organizer (2018 - 2019)
- Informal Reviewer: MICRO (2018 - 2021), HPCA (2019 - 2021), ASPLOS (2019, 2022), MLSys (2019 - 2023), ISCA (2019, 2020), ICS (2019), CGO (2020), EuroSys (2020, 2022, 2023), OSDI (2021)
- (MSc and PhD) Graduate Admission Triager (Department of Computer Science, University of Toronto, Dec. 3rd - 19th, 2021)

MENTORING

- Baorun Mu (2022 - present); during research collaboration; currently Research Software Development Engineer at CentML
- Qingyuan Qie (jointly with Bojian Zheng, 2021 - 2022); during research collaboration; currently Software Engineer at AWS
- Yuxuan Zheng (2020 - 2021); during research collaboration; currently MSCS student at CMU
- Peiming Yang (2020 - present); during research collaboration; currently MASc student at the University of Toronto
- Xin Li (2019 - present); during research collaboration; currently MASc student at the University of Toronto
- Joseph Jennings (jointly with Sharah Turuvekere Sreenivas, 2021); during his internship at NVIDIA; currently Deep Learning Software Engineer at NVIDIA
- Yifan Bai (2018 - 2019); during research collaboration; currently Software Engineer at Google

REFERENCES

References available upon request.