

Yu (Emma) Wang

Email: yuemmawang@google.com
Homepage: yuemmawang.github.io

Work Experience

Software Engineer Nov 2019 – Present
Google LLC
1600 Amphitheatre Pkwy, Mountain View, CA 94043

Education

Ph.D. in Computer Science Jul 2013 – Nov 2019
M.S. in Computer Science Jul 2013 – Feb 2017
Advisors: Prof. David Brooks and Prof. Gu-Yeon Wei
Harvard University
1350 Massachusetts Ave, Cambridge, MA 02138

B.S. in Computer Science and Engineering Sep 2009 – Jul 2013
Advisors: Prof. Bo Yuan and Prof. Xiaoyao Liang
Shanghai Jiao Tong University
800 Dongchuan Rd, Shanghai, China 200240

Publications

Yu Wang, Carole-Jean Wu, Xiaodong Wang, Kim Hazelwood, David Brooks, “Exploiting Parallelism Opportunities with Deep Learning Frameworks.” arXiv preprint arXiv:1908.04705 (2019).

Yu Wang, Gu-Yeon Wei, and David Brooks. “A Systematic Methodology for Analysis of Deep Learning Hardware and Software Platforms.” MLSys (2020).

(The arXiv version of the above paper) **Yu Wang**, Gu-Yeon Wei, and David Brooks. “Benchmarking TPU, GPU and CPU for Deep Learning.” arXiv preprint arXiv:1907.10701 (2019).

Yu Wang, Yuhao Zhu, Glenn G. Ko, Brandon Reagen, Gu-Yeon Wei, and David Brooks. “Demystifying Bayesian Inference Workloads.” International Symposium on Performance Analysis of Systems and Software (ISPASS), 2019.

Yu Wang, Victor Lee, Gu-Yeon Wei, and David Brooks. “Predicting New Workload or CPU Performance by Analyzing Public Datasets.” ACM Transactions on Architecture and Code Optimization (TACO). vol. 15, no. 4 (2019): 53:1–53:21.

Yu Wang, Weikang Qian, Shuchang Zhang, Xiaoyao Liang, and Bo Yuan. “A Learning Algorithm for Bayesian Networks and Its Efficient Implementation on GPUs.” IEEE Transactions on Parallel and Distributed Systems. vol. 27, no. 1 (2016): 17–30.

Weichao Tang, **Yu Wang**, Haopeng Liu, Tao Zhang, Chao Li, and Xiaoyao Liang. “Exploring Hardware Profile-Guided Green Datacenter Scheduling.” International Conference on Parallel Processing (ICPP), pp. 11-20. 2015.

Research Experience

Research Assistant Jul 2013 – Sep 2019
Advisors: Prof. David Brooks and Prof. Gu-Yeon Wei
Harvard Architecture, Circuits, Compilers Group
Harvard University, Cambridge, MA

- Conducted full-stack benchmarking of machine learning workloads and extracted architectural insights to optimize those workloads.
- Proposed *BayesSuite*, a benchmark suite for Bayesian inference, revealed and resolved crucial computational bottlenecks of Bayesian inference workloads.
- Proposed ParaDnn, a deep learning benchmark suite that generates parameterized multi-layer models to allow systematic benchmarking across a wide range of model parameter sizes covering current and future deep learning models.

- Benchmarked deep learning models on TPU, GPU and CPU, highlighted insights for future deep learning accelerator design, and proposed potential directions for further optimization.
- Quantifying performance improvements of various deep learning framework design choices.

Research Intern (Mentors: Xiaodong Wang and Carole-Jean Wu) Nov 2018 – Feb 2019
Facebook AI Infrastructure Research Team
 Cambridge, MA

- Performed performance comparison across different deep learning frameworks and identified the source of performance difference in depth.
- Extracted insights to optimize Caffe2 from the analysis results.

Software Engineering Intern (Mentor: Hui Huang) May – Aug 2018
Google Platforms Team
 Sunnyvale, CA

- Benchmarked 3rd generation of Tensor Processing Units (TPU v3) with state-of-the-art deep learning workloads.
- Predicted potential bottlenecks of Cloud TPU v3.
- Quantified the impact of NUMA-aware allocation for Cloud TPU v3 (Silver Perfy Award in 2019 Q1 at Google).

Software Engineering Intern (Mentor: Cliff Young) Sep – Dec 2017
Google Brain Team
 Mountain View, CA

- Benchmarked 2nd generation of Tensor Processing Units (TPU v2) with state-of-the-art deep learning workloads and analyzed their bottlenecks.
- Quantified performance scalability and speedup of Cloud TPU v2.

Parallel Computing Intern (Mentor: Victor Lee) July 2015 – Jan 2016
Intel Parallel Computing Lab
 Santa Clara, CA

- Developed a set of tools to characterize CPU workloads, extracted platform independent features including memory locality, memory footprint, and branch entropy.

Research Assistant (Advisor: Prof. Xiaoyao Liang) June 2012 – Jul 2013
Advanced Computer Architecture Laboratory
 Shanghai Jiao Tong University, Shanghai, China

- Designed a hardware profile-guided scheduler for green datacenters.
- Reduce the datacenter energy cost up to 54% while maintaining fairly balanced processor utilization.

Research Assistant (Advisor: Prof. Bo Yuan) Sep 2011 – Jul 2013
Lab for Biocomputing and Bioinformatics
 Shanghai Jiao Tong University, Shanghai, China

- Optimized a Bayesian network learning algorithm and implemented on GPU.
- Achieved a 143× speedup on GPU over CPU.
- Applied this method to networks of up to 125 nodes.

Awards

Academic Excellence Scholarship (third-class) of Shanghai Jiao Tong University Nov 2012
 Third Prize of Chinese Physics Olympiad Oct 2008

Project Experience

Memory Trace Compression

Sep – Dec 2016

Algorithms at the End of the Wire by Prof. Michael Mitzenmacher
Harvard University

- Compared delta, k-means and lossy k-means compression schemes on memory traces.
- Lossy k-means compression provides a trade-off between compression rate and preserved information.

Die Photo Analyzer

Oct – Dec 2014

Computer Vision by Prof. Todd Zickler
Harvard University

- Detected the boundaries and SRAMs in CPUs, GPUs and accelerators.
- Computed the area fractions of SRAMs within the CPUs, GPUs and accelerators.
- Computed the area fractions of the CPUs, GPUs and accelerators within the chip.

Pre-training of Deep Neural Networks (DNNs) with Bayesian Network (BN) Structure Learning

Oct – Dec 2014

Research Topics in Operating Systems by Prof. Margo Seltzer
Harvard University

- Pretraining of DNNs with the results learnt by BN structure learning.
- Achieved more than 5X speedups compare with a trivial DNN to get a reasonable mean square error on measured benchmarks.

Cluster Benchmarks by Memory Features

Sep – Dec 2013

Advanced Machine Learning by Prof. Ryan Adams
Harvard University

- Collected feature vectors by Pin, including both load/store features and striding features, in both program level and instruction level.
- Clustered 20 benchmarks from Rodinia and Parsec using k-means algorithm.

Technical Skills

Languages C/C++, Python, Matlab, LaTeX

Frameworks Tensorflow, CUDA, RStan

OS Linux

Theoretical Expertise Deep Learning, Bayesian inference, Bayesian networks, Markov chain Monte Carlo, Gibbs Sampling, Linear/Logistic/Laplacian Regression, Stochastic Gradient Descent, Principle Component Analysis, Linear Discriminant Analysis

Hardware CPU, GPU, FPGA, TPU (Tensor Processing Unit)

Teaching Experience

CS246 Advanced Computer Architecture

Teaching Fellow for Prof. David Brooks

Spring 2017

Harvard University

CS141 Computing Hardware

Teaching Fellow for Prof. David Brooks

Fall 2014

Harvard University