Jian Zhang

CONTACT
E-mail: zjian@stanford.edu
Personal Homepage: http://cs.stanford.edu/people/zjian

PRESENT ADDRESS
Computer Science Department
353 Serra Mall, Stanford University
Stanford, CA 94305-9025, USA

Research Interests

Training and inference for hardware-efficient ML systems:
Word embedding compression. Low precision kernel methods. Optimization algorithms for ML accelerators.

Large-scale ML systems: Distributed asynchronous training systems.

Applications of ML: Machine reading comprehension on natural language. Visual scene understanding.

EDUCATION

Stanford University Sep. 2015 - Present
Ph.D. Candidate in Computer Science

ETH Zürich Sep. 2013 - Jun. 2015
Master in Computer Science
GPA: 5.50/6

Tsinghua University Jul. 2010 - Jul. 2013
Bachelor in Electronic Information Science and Technology
Major GPA: 90.4/100

Tsinghua University Aug. 2009 - Jul. 2010
Undergraduate in Electrical Engineering and Automation Major Changed
Major GPA: 92.9/100

PREPRINTS

On the Downstream Performance of Compressed Word Embeddings *To be on arXiv soon.* A. May, **J. Zhang**, Tri Dao, C. Ré

High-accuracy Low-precision Training. *arXiv preprint arXiv:1803.03383 2018.* C. De Sa, M. Leszczynski, **J. Zhang**, A. Marzoev, C. Aberger, K. Olukotun, C. Ré

PUBLICATION

Low-precision Random Fourier Features for Memory-constrained kernel approximation. *International Conference on Artificial Intelligence and Statistics (**AISTATS**) 2019.* **J. Zhang***, A. May*, T. Dao, C. Ré.

YellowFin and the Art of Momentum Tuning. *SysML Conference (**SysML**) 2019.* **J. Zhang**, I. Mitliagkas.

Training with Low-precision Embedding Tables. *Workshop on Systems for ML and Open Source Software at NeurIPS 2018.* **J. Zhang**, J. Yang, H. Yuen.

Analysis of the Time-to-accuracy Metric and Entries in the DAWNBench Deep Learning Benchmark. *Workshop on Systems for ML and Open Source Software at NeurIPS 2018.* C. Coleman, D. Kang, D. Narayanan, L. Nardi, T. Zhao, **J. Zhang**, P. Bailis, K. Olukotun, C. Ré, M. Zaharia

Exploring the Utility of Developer Exhaust. *Workshop on Data Management for End-to-End Machine Learning at SIGMOD 2018.* **J Zhang**, M. Lam, S. Wang, P. Varma, L. Nardi, K. Olukotun, C. Ré

DAWNBench: An End-to-end Deep Learning Benchmark and Competition. *SysML Conference (**SysML**) 2018.* C. Coleman, D. Narayanan, D. Kang, T. Zhao, **J. Zhang**, L. Nardi, P. Bailis, K. Olukotun, C. Ré, M. Zaharia

YellowFin: Adaptive optimization for (A)synchronous systems. *SysML Conference (**SysML**) 2018.* **J. Zhang**, I. Mitliagkas.

Peta-scale Deep Learning: Supervised and Semi-supervised classification for scientific data. *Supercomputing (**SC**) 2017.* T. Kurth, **J. Zhang**, N. Satish, I. Mitliagkas, E. Racah, M. Patwary, T. Malas, N. Sundaram, W. Bhimji, M. Smorkalov, J. Deslippe, M. Shiryaev, S. Shridharan, Prabhat, P. Dubey.

SQuAD: 100,000+ questions for machine comprehension of text. *Conference on Empirical Methods in Natural Language Processing (**EMNLP**) 2016.* P. Rajpurkar, **J. Zhang**, K. Lopyrev, P. Liang. **Best resource paper award**

Parallel SGD: When does Averaging Help? *Optimization in Machine Learning Workshop (**OptML Workshop ICML**) 2016.* **J. Zhang**, C. De Sa, I. Mitiliagkas, and C. Ré.

Higher-order Inference for Multi-class Log-supermodular Models. *International Conference in Computer Vision (**ICCV**) 2015.* **J. Zhang**, J. Djolonga, A. Krause.

Message Passing Inference for Large Scale Graphical Models with High Order Potentials. *Neural Information Processing Systems (**NIPS**) 2014.* **J. Zhang**, A.Schwing, R. Urtasun.

Estimating Indoor Layout with Its Clutter from Depth Sensors. *International Conference in Computer Vision (**ICCV**) 2013.* **J. Zhang**, K. Chen, A.Schwing, R. Urtasun.

Non-iterative Normalized Feature Extraction in Large Viewpoint Variances Based on PCA of Gradient. *IS&T/SPIE Electronic Imaging (**EI**) 2013.* **J. Zhang**, S. Cao, D. Wen.

RESEARCH EXPERIENCE

Research Assistant Sep. 2016 - President
Advisor: **Prof. Christopher Ré** Statistical Machine Learning Group, DAWN Group, Stanford University

Project: Compressed training and inference under memory constraints

- Stochastic optimization for ML-accelerators with low precision arithmetic and limited memory.
- Investigated the performance of compressed word embeddings, low precision kernel approx. features.

Project: Optimization with momentum adaptivity and large scale deep learning on HPC

- Designed *YellowFin*, an SGD based optimizer with both momentum and learning rate adaptivity.
- ***YellowFin* is adopted in projects at Facebook.**
- Collaborated with Intel/NERSC in an asynchronous deep learning system with a cluster of 9600 nodes.
- **The asynchronous system design is adopted in production code at Intel.**

Research Assistant March. 2016 - July. 2016
Advisor: **Prof. Percy Liang** Natural Language Processing Group, Stanford University

Project: SQUAD The Stanford Question Answering Dataset

- Collaborated in collecting 100,000+ question-answer pairs on 500+ wikipedia articles.
- Conducted thorough analysis and developed baseline models on the collected dataset.
- **Currently the standard testbench for question answering system based on deep learning.**

Research Assistant Sep. 2014 - Apr. 2015
Advisor: **Prof. Andreas Krause** Learning & Adaptive Systems Group, ETH Zurich

Project: Scalable Parallel Inference for Multi-class Log-supermodular Models

- Incorporated partition matroids for multi-class modeling with log-supermodular models.
- Parallelized optimization for marginal and smoothed MAP inference over additive submodular energies.
- Presented theoretical analysis on the trade-off between accuracy and time-efficiency for smoothed MAP.

Research Intern & Visiting Student Jul. 2012 - Jun. 2014
Advisor: **Prof. Raquel Urtasun** Toyota Technological Institution at Chicago (TTIC)

Project: Efficient Inference for Densely Connected High-order Graphical Models

- Proposed a distributed formulation and a partition strategy of region graph based inference.
- Incorporated an efficient dual BCD approach as a parallel rescheduled message passing algorithm.
- The algorithm is magnitudes faster than state-of-art for densely connected high-order vision models.

Project: Joint Indoor Layout Estimation and Scene Parsing with RGB-D Data

- Proposed a novel joint model on layout estimation and superpixel-wise scene segmentation.
- Designed a fast Integral Geometry accumulation algorithm.
- Investigated an efficient alternating inference framework for the high order joint model.

INDUSTRIAL EXPERIENCE

Research Intern June. 2018 - Now
Facebook AML, Menlo Park, USA

Project: Algorithm and system for sparse low precision optimization

- Designed low precision training algorithm for sparse models.
- Achieved 2x memory saving and 1.3x training throughput for large-scale recommender system.
- **Algorithm and system design adopted in Facebook production recommender systems.**

Software Engineering Intern June. 2016 - Sep. 2016
NovuMind, Santa Clara, USA

Project: Efficient Deep Learning System in High Performance GPU Clusters

- Redesigned and implemented ring-based GPU communication for multiple GPU training.
- Achieved 27.9x speedup using 32 TITAN X GPU for 101-layer Resnet.

Data Science Intern May. 2015 - Aug. 2015
Teralytics AG, Zürich, Switzerland

Project: Transfer Learning of Gaussian Process in Train Positioning for Rider Counting

- Proposed a spatial-temporal model to predict train positions based on cellphone connection data.
- Trained models of new train routes without GPS data using limited GPS data from other routes.
- Presented 3.5 times smaller average positioning error compared with the model in the latest product.
- Demonstrated practical running time for positioning in offline applications.

TEACHING EXPERIENCE

Teaching Assistant Fall 2018
Machine Learning CS 229, Stanford University Prof. Ron Dror, Prof. Andrew Ng

Teaching Assistant Fall 2014
Linear Algebra 401-0131-00L, ETH Zürich Prof. Roman Glebov, Prof. Marc Pollefeys

Teaching Assistant Spring 2014, Spring 2015
Design of Digital Circuits 252-0014-00L, ETH Zürich Prof. Srdjan Capkun, Prof. Frank K. Gürkaynak

Teaching Assistant Fall 2013
Informatiks 252-0847-00L, ETH Zürich Prof. Bernd Gärtner

TALKS

YellowFin and the Art of Momentum Tuning
SysML Conference 2019, USA — Apr. 2019

Training with Low-precision Embedding Tables
Facebook AI System Co-design Group, USA — Apr. 2019

Efficient Parallel Inference for Densely Connected High-order Graphical Models
Machine Learning Group, University of Toronto, Canada — Dec. 2014

PROFESSIONAL ACTIVITIES

Reviewer for

- SysML Conference 2018
- IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)

SELECTED HONORS

Best Resource Paper	21st Conference on Empirical Methods in Natural Language Processing (EMNLP)	Nov. 2016
Distinguished Graduate	Tsinghua University (Cum Laude)	Jul. 2013
Member	Talents Program for Technological Innovations, Tsinghua University (36 out of 3000 undergraduates)	Dec. 2012
1st Class (Top 3%)	Scholarship for Research and Innovation	Oct. 2012
2nd Class (Top 10%)	Zheng Geru Scholarship for Academic Excellence	Oct. 2011
Travel Grant	28th Neural Information Processing Systems (NIPS)	Dec. 2014
Travel Grant	13th International Conference on Computer Vision (ICCV)	Dec. 2013

GRADUATE COURSES

Math & **Statistics**	Linear Optimization, Convex Optimization, Statistical Inference, Probabilistic Graphical Models.
Computer Science	Big Data, Principle of Distributed Computing, Seminar of Distributed Computing, Machine Learning, Statistical Learning Theory, Algorithm Lab.

PROGRAMMING SKILLS

Programming Language	C/C++, Python, Matlab, R, Verilog HDL, MIPS Assembly.
Computing Tools	OpenMP, MPI, Map-Reduce
Operating System	Linux, Windows.

LANGUAGE SKILLS

TOEFL	Total	110	Reading	29	Listening	29	Speaking	26	Writing	26
GRE	Verbal	157		Quantitative		170		Analytical Writing		3.5