# Yang You

Office: 583, Soda Hall, UC Berkeley, CA, USA youyang@cs.berkeley.edu
Phone: (+001) 510-508-4506 http://www.cs.berkeley.edu/~youyang/

## Research Interest: Parallel & Distributed Machine Learning Algorithm

- **High Performance Computing**: Scalable Algorithms, Parallel Computing, Distributed Systems
- **Machine Learning**: Deep Learning, Optimization Algorithm, Matrix Computations

## Current Education (08/2015 — present)

**PhD candidate at UC Berkeley** Computer Science Division
**Advised by Prof. James Demmel** *Focus: Parallel & Distributed Machine Learning Algorithm*

## Previous Education (09/2009 — 07/2015)

**Tsinghua University** Computer Science Department
**Ranking: 1st out of 134 students** *Master Degree in Computer Science*
**China Agricultural University** Honors Program (most selective program)
**Ranking: 1st out of 52 students** *Bachelor Degree in Computer Science*
- Project-985 Universities Freshmen acceptance rate in Henan province (year 2009): 0.5% (5060/959,000).

## Selected Awards

**Best Paper Award of ICPP 2018 (1 out of 313 submissions: 0.3%, plenary presentation)** [Link]
**ACM/IEEE-CS George Michael Memorial HPC Fellowship**: the only PhD fellowship on ACM website.
Media Coverage: [ACM] [Berkeley] [China] [EurekAlert] [IEEE] [insideHPC]
NeurIPS 2016 student travel award from Google (1000 USD) [Link]
**Best Paper Award of IPDPS 2015 (4 out of 496 submissions: 0.8%, plenary presentation)** [Link]
Outstanding Graduate of Tsinghua University (ranked 1st among 134 students, top 3 got the awards) [Link]
Outstanding Graduate of Beijing (ranked 1st among 134 students, top 4 got the awards) [Link]
Outstanding Graduate of Tsinghua CS Department (ranked 1st among 134 students, top 20 got the awards) [Link]
2015 Best Thesis Award of Tsinghua University (10 out of 134 students: 7%) [Link]
**Siebel Scholar (35,000 USD for one year), 85 top students from the world's leading universities** [link]
IEEE TCPP Student Travel Grants to IPDPS [Link]
Outstanding Graduate of Beijing (157 of 3,255: 5%, no ranking) [Link]
Outstanding Graduate of CAU (505 of 3,255: 15%) [Link]
2012 Outstanding Youth Nomination of CAU(**30 of over 30,000: 0.1%**) [Link]
**First Prize, 2011 National Programming Contest (20 of over 10,000: 0.2%)** [Link]
2011 National Scholarships of China (ranked 1 among 52 students, top 2 got the award) [Link]
2011 President Scholarship (ranked 1 among 52 students, top 1 got the award) [Link]
2010 National Scholarships of China (ranked 1 among 52 students, top 2 got the award) [Link]
2010/2011 Merit Student of CAU [Link]
Third Prize, 27th Undergraduate Physics Competition in China [Link]
Third Prize, Undergraduate Mathematical Competition in China [Link]
2009-2012 Merit Student of CAU [Link]

## First-Author Publications (Peer-Reviewed)

- **[TPDS'19] Y. You**, Z. Zhang, C. Hsieh, J. Demmel, K. Keutzer. Fast Deep Neural Network Training on Distributed Systems and Cloud TPUs, IEEE Transactions on Parallel and Distributed Systems, h5-index=76, **accepted**
- **[ICPP'18] Y. You**, Z. Zhang, C. Hsieh, J. Demmel, K. Keutzer. ImageNet Training in Minutes, 47th International Conference on Parallel Processing. August 13th - 16th, Eugene, USA. **Best Paper Award (1 out of 313 submissions: 0.3%)**. [pdf] [code]

- [**ICS'18**] **Y. You**, J. Demmel, C. Hsieh, R. Vuduc. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems, ACM International Conference on Supercomputing (ICS), June 12-15, Beijing, China. **18.7% (36/193) acceptance rate** [pdf]
- [**SysML'18**] **Y. You**, Z. Zhang, C. Hsieh, J. Demmel, K. Keutzer. Speeding up ImageNet Training on Supercomputers, System Machine Learning Conference, Feb 15, Stanford, USA. **The first year of this conference only accepts 2-page paper** [pdf]
- [**NeurIPS-W'17**] **Y. You**, I. Gitman, B. Ginsburg. Scaling SGD Batch Size to 32K for ImageNet Training. NIPS workshop. Widely used in industry. **Available in Intel Caffe, NVIDIA Caffe, Facebook Caffe2 (PyTorch), and Google's distributed TensorFlow**. [pdf]
- [**SC'17**] **Y. You**, A. Buluc, J. Demmel. Scaling Deep Learning on GPU and Knights Landing Clusters, International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing), November 12-17, Denver, USA. **18.7% (61/327) acceptance rate** [pdf]
- [**ICPP'17**] **Y. You**, J. Demmel. Runtime Data Layout Scheduling for Machine Learning Dataset, 46th International Conference on Parallel Processing. 28.4% (60/211) acceptance rate. [pdf]
- [**TPDS'16**] **Y. You**, J. Demmel, K. Czechowski, L. Song, R. Vuduc. Design and Implementation of a Communication-Optimal Classifier for Distributed Kernel Support Vector Machines, IEEE Transactions on Parallel and Distributed Systems, h5-index=76, DOI: 10.1109/TPDS.2016.2608823 [pdf]
- [**NeurIPS'16**] **Y. You**, X. Lian, J. Liu, H. Yu, I. Dhillon, J. Demmel, C. Hsieh. Asynchronous Parallel Greedy Coordinate Descent, Conference on Neural Information Processing Systems, Dec 05-10, Barcelona, Spain. **22.7% (568/2500) acceptance rate** [pdf] [link]
- [**JPDC'16**] **Y. You**, H. Fu, D. Bader, G. Yang. Designing and Implementing a Heuristic Cross-Architecture Combination for Graph Traversal, Journal of Parallel and Distributed Computing, h5-index=36, DOI: 10.1016/j.jpdc.2016.05.007 [pdf]
- [**IPDPS'15**] **Y. You**, J. Demmel, K. Czechowski, L. Song, R. Vuduc. CA-SVM: Communication-Avoiding Support Vector Machines on Distributed Systems. **Best Paper Award (4 out of 496 submissions: 0.8%)** of IEEE International Parallel and Distributed Processing Symposium, May 25-29, Hyderabad, INDIA. DOI: 10.1109/IPDPS.2015.117 [pdf] [code]
- [**IPDPS'14**] **Y. You**, S. Song, H. Fu, A. Marquez, M. Dehnavi, K. Barker, K. Cameron, A. Randles, G. Yang. MIC-SVM: Designing A Highly Efficient Support Vector Machine For Advanced Modern Multi-Core and Many-Core Architectures. IEEE Parallel and Distributed Processing Symposium, May 19-23, Phoenix, USA. **21% (114/541) overall acceptance rate; 17.5% acceptance rate for software track.** DOI: 10.1109/IPDPS.2014.88 [pdf] [code]
- [**JPDC'14**] **Y. You**, H. Fu, S. Song, A. Randles, D. Kerbyson, A. Marquez, G. Yang, A. Hoisie. Scaling Support Vector Machines on the Modern HPC Platforms, Journal of Parallel and Distributed Computing, h5-index=36, DOI: 10.1016/j.jpdc.2014.09.005 [pdf]
- [**ICPP'14**] **Y. You**, D. Bader, M. Dehnavi. Designing a Heuristic Cross-Architecture Combination for Breadth-First Search, 43rd International Conference on Parallel Processing, Sep 9-12, Minneapolis, USA. 36% (54/150) acceptance rate. DOI: 10.1109/ICPP.2014.16 [pdf]
- [**IJHPCA'14**] **Y. You**, H. Fu, S. Song, M. Dehnavi, L. Gan, X. Huang, G. Yang. Evaluating the Many-core and Multi-core architectures through accelerating LWC stencil on Multi-core and Many-core architectures. International Journal of High Performance Computing Application (2013 SCI IF=1.625), **21% (5/24) acceptance rate**. DOI: 10.1177/1094342014524807 [pdf]
- [**ICS'14**] **Y. You**, S. Song, D. Kerbyson. An adaptive cross-architecture combination method for graph traversal, **one-page short paper**, ACM International Conference on Supercomputing, June 10-13, Munich, Germany. DOI: 10.1145/2597652.2600110 [pdf]
- [**IPDPS-W'13**] **Y. You**, H. Fu, X. Huang, G. Song, L. Gan, W. Yu, G. Yang. Accelerating the 3D Elastic Wave Forward Modeling on GPU and MIC. IEEE Parallel and Distributed Processing Symposium **Wrokshops**, May 20-24, Boston, USA. One of the **best papers** of AsHES workshop. DOI: 10.1109/IPDPSW.2013.216 [pdf]

## Co-Author Publications (Peer-Reviewed)

- [**BMC Genomics'18**] Y. Zhao, C. Sun, D. Zhao, **Y. You**, et al. PGAP-X: extension on pan-genome analysis pipeline, BMC Genomics, DOI: 10.1186/s12864-017-4337-7 [pdf]
- [**J-STARS'17**] W. Li, H. Fu, **Y. You**, L. Yu, J. Fang. Parallel Multiclass Support Vector Machine for Remote Sensing Data Classification on Multicore and Many-Core Architectures. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017 h5-index=45. DOI: 10.1109/JSTARS.2017.2713126 [pdf]

- [**ICPADS'14**] L. Gan, H. Fu, W. Xue, Y. Xu, C. Yang, X. Wang, Z. Lv, **Y. You**, G. Yang, and K. Ou. Scaling and Analyzing the Stencil Performance on Multi-Core and Many-Core Architectures. IEEE International Conference on Parallel and Distributed Systems (ICPADS). DOI: 10.1109/PADSW.2014.7097797 [pdf]

## Experience

**James Demmel & Kathy Yelick Group, UC Berkeley** — Berkeley, CA, USA
*Graduate Student Researcher (GSR)* — *08/2015 – present*
- Performance Benchmark and Optimization for Deep Neural Networks
- Communication Avoiding Machine Learning Algorithms on Distributed systems
- Communication-Efficient Solver for Kernel Ridge Regression (600× speedup without losing accuracy)
- Fast DNN Training for ImageNet on CPUs: AlexNet in 11 minutes and ResNet-50 in 15 minutes

**Google Brain** — Mountain View, CA, USA
*Student Researcher* — *01/2019 – 05/2019*
- Optimize TensorFlow for Large-Scale Deep Learning on TPU Pod

**Intel Labs** — Santa Clara, CA, USA
*Research Intern* — *08/2018 – 12/2018*
- Fast and Efficient LSTM Training

**Google Brain** — Mountain View, CA, USA
*Software Engineering Intern* — *05/2018 – 08/2018*
- Optimize TensorFlow for Large-Scale Deep Learning on TPU Pod

**Microsoft Research** — Redmond, WA, USA
*Research Intern* — *01/2018 – 05/2018*
- Fast LSTM Inference on Cloud System
- Design and Implement Approaches based on SVD and Tensor Decomposition
- Achieved up to 30× Speedup and 20× Parameter Reduction

**NVIDIA** — Santa Clara, CA, USA
*Deep Learning Intern* — *05/2017 – 08/2017*
- Scaling SGD Batch Size to 32K for ImageNet training by ResNet50 model
- Achieve 3× speedup over standard AlexNet-ImageNet Training on DGX station
- Enables multiple solvers on each GPU, which achieves 1.4× speedup over 1-solver-per-GPU

**IBM T. J. Watson Research Center** — Yorktown, NY, USA
*Research Intern* — *05/2016 – 08/2016*
- Design communication-optimized GPU-enabled learning algorithms
- Improve the communication efficiency of Elastic Averaging SGD
- Evaluate collective operations on GPUs (e.g., NCCL)

**High Performance Computing Lab, Georgia Institute of Technology** — Atlanta, GA, USA
*Research Assistant (Exchange Student)* — *05/2014 – 08/2014*
- Convert a communication-intensive algorithm (SMO) to a communication avoiding algorithm (CA-SVM)
- CA-SVM achieves 7× average speedup over the original algorithm with only 1.3% average losses in accuracy
- CA-SVM keeps 95.3% weak scaling efficiency when we increase the number of processors from 96 to 1536

**High Performance Computing Lab, Georgia Institute of Technology** — Atlanta, GA, USA
*Research Assistant (Exchange Student)* — *10/2013 – 11/2013*
- Adaptive method based on regression, which supports the runtime combination technique
- Cross-architecture combination, which achieves 8.5×, 2.6×, and 2.2× average speedup over MIC, CPU and GPU
- Pairwise comparison between CPU, GPU and MIC, which helps users select the best architectures

**Department of Computer Science, Tsinghua University** — Beijing, China
*Research Assistant* — *09/2012 – 07/2015*
- Design and implement MIC-SVM, a highly parallel support vector machines for x86 many-core architectures
- Adaptive support for input patterns and data parallelism to fully utilize the multi-level parallelism
- MIC-SVM achieves 4.4-84× and 18-47× speedups against LIBSVM on MIC and Ivy Bridge CPUs respectively

**Institute of High Performance Computing, Tsinghua University** — Beijing, China
*Research Assistant* — *06/2011 – 09/2011*
- Developed a distributed system for automated software deployment and user data storage

## Teaching

**UC Berkeley CS194-129 (funding from Google)** — Berkeley, CA, USA
*Designing, Visualizing and Understanding Deep Neural Networks* — *08/2016 – 12/2016*
- Algorithms, Applications, and Implementations of Deep Learning Techniques

- Head TA/GSI of Prof. John Canny

**UC Berkeley CS162** Berkeley, CA, USA
*Operating Systems and Systems Programming* *08/2018 – 12/2018*

- Theory, Algorithms, and Implementations of Operating Systems
- TA/GSI of Prof. Ion Stoica

## Contributions to Open-Source Software

[Asyn SVM]: the fastest implementation for Kernel Support Vector Machines on shared systems as of 2016
[CA-SVM]: a Communication-Avoiding approach for Kernel Support Vector Machines on distributed systems
[MIC-SVM]: an efficient design of Sequential Minimal Optimization approach for SVM on shared-memory systems
[NVIDIA-Caffe]: I enabled multiple solvers on each GPU, which achieves 1.4× speedup over 1-solver-per-GPU
[NVIDIA-Caffe]: I developed the LARS algorithm with B. Ginsburg and I. Gitman for large-batch training
[Intel-Caffe]: I helped Intel team implement large-batch DNN training algorithms
[Tensorflow]: I helped Sameer Kumar and Chris Ying implement large-batch DNN training algorithms

## Academic Services

- [**TOPC'19**] Reviewer of ACM Transactions on Parallel Computing [link].
- [**IBM'19**] Reviewer of IBM Journal of Research & Development [link].
- [**ICPP'18**] Reviewer of International Conference on Parallel Processing [link].
- [**TPDS**] 6 times reviewer of IEEE Transactions on Parallel and Distributed Systems, h5-index=76 [link].
- [**NCAA**] 2 times reviewer of Neural Computing and Applications [link].
- [**FCGS**] Reviewer of Future Generation Computer Systems, h5-index=63 [link].
- [**CCGRID'18**] Reviewer of IEEE International Symposium on Cluster Computing and the Grid [link].
- [**JMLR'17**] Reviewer of Journal of Machine Learning Research, h5-index=70 [link].
- [**JPDC**] Two times reviewer of Journal of Parallel and Distributed Computing, h5-index=36 [link].
- [**IJCAI'17**] **Senior Program Committee member** of International Joint Conference on Artificial Intelligence. Melbourne, Victoria, Australia, August 19 - 25, 2017 [link].
- [**IPDPS'17**] Sub-Reviewer in Algorithms Track of IEEE International Parallel and Distributed Processing Symposium. Orlando, Florida, USA, May 29 - June 2, 2017 [link].
- [**APDCM'16**] Reviewer of 18th Workshop on Advances in Parallel and Distributed Computational Models. Chicago, Illinois, USA, May 23 - 27, 2016 [link].

## Media Coverage on Research

- [**i-programmer**] ImageNet Training Record - 24 Minutes, Sep 21, 2017 [link], [copy].
- [**EurekAlert**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**ScienceDaily**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**NSF**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**Intel**] Solving Science and Engineering Problems with Supercomputers and AI, Nov 15, 2017 [link], [copy].
- [**Berkeley**] EECS-affiliated team break record for fastest deep learning training, Nov 15, 2017 [link], [copy].
- [**R&D Magazine**] Supercomputing Speeds Up Deep Learning Training, Nov 15, 2017 [link], [copy].
- [**fourthventricle**] Supercomputing Speeds Up Deep Learning Training, Nov 17, 2017 [link], [copy].
- [**techxplore**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**TACC**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**Science NewsLine**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**Topix**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**Technology News**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**Get Knows**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**Technology Networks**] Deep Learning Training Accelerated by Super Computing, Nov 14, 2017 [link], [copy].

- **[Primeur Magazine]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[World IT]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[Parallel State]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[CACM]** Supercomputing Speeds Up Deep Learning Training, Nov 17, 2017 [link], [copy].
- **[Intel Software]** Intel CPUs for Deep Learning Training, Nov 17, 2017 [link], [copy].
- **[The Next Web]** Facebooks nerds bested by Japans in the race to train AI, Nov 20, 2017 [link], [copy].

## Mentoring and Service

- **Since 2015 Fall:** Mentor for 1-2 UC Berkeley EECS undergraduate students in research/study per semester.
- **2018 Spring:** UC Berkeley Computer Science Division Student Core Committee for Faculty Hiring.
- **Since 2016 Spring:** Student volunteer and host for incoming UC Berkeley EECS PhD students each year.
- **2012 Fall - 2015 Spring:** Student leader in Tsinghua University youth league for organizing technology talks and exhibitions.
- **2010 Fall - 2011 Fall:** Chief student leader in CAU's EECS honors program.

## Skills

- General: **C/C++, Matlab**, **Python**, **Java**, **Scala**, **Lua** and **Shell script**
- Multi-Core GPUs, CPUs and MIC: **CUDA**, **OpenMP**, **Pthreads** and **Intel Cilk**
- Distributed Systems: **MPI**, **Hadoop**, and **Apache Spark**
- Tools: **Caffe**, **TensorFlow**