Dayou Du

+1 (646) 206-7968| dayoudu@nyu.edu |30 River Court, Jersey City, NJ 07310

EDUCATION

New York University – New York, NY *Sep. 2017 – May 2019 (Expected)*

Master of Science in Computer Science, Courant Institute of Mathematical Sciences

- GPA: **4.0/4.0**, Rank: **#1**
- Courses: Advanced Database System, Distributed System, Deep Learning, Computer Vision, Real-time and Big Data Analytics
- Teaching Assistant in: Operating Systems, Computer Systems Organization, Fundamental Algorithms

Peking University – Beijing, China *Sep.2013 – July 2017*

Bachelor of Science in Computer Science (Summa Cum Laude), Double Major in Economics

- Major GPA: **3.71/4.0**
- Core courses: *Algorithm Design and Analysis, Data Structure and Algorithm, Operating Systems, Computer Organization, Computer Architectures, *Computer Networks, *Software Engineering, *Database Systems, Parallel Computing, Linear Algebra, Set Theory and Graph Theory, Probability Theory and Mathematical Statistics; (*: Honor Track)

SKILLS

Programming: C++, C, CUDA, OpenMP, MPI, Scala, Golang, Python, Spark, Hadoop, Hive, Impala, SQL, Html
Tools and Skills: Vim, Visual Studio, IntelliJ, Eclipse; Git, Gprof, Valgrind, Perf; MySQL, MongoDB; Caffe; Working on Linux

INTERNSHIP / RESEARCH

NVIDIA – Santa Clara, CA *May 2018 – Aug. 2018*

AI Developer Technology Engineer Intern, Top Contributor

- Implemented and optimized reduced precision computing techniques (aka. quantization) on a big customer's RNN models. Achieved 3.84x speedup on core operations with neglectable precision lost comparing to the SOTA FP32 implementation.
- Obtained 1.8x speedup on optimizing Softmax & TopK in cuDNN and TensorRT, which benefits DL developers around the world.
- Optimized multi-node & multi-card k-means algorithm with varies of parallel computing techniques. Achieved over 90x speedup.

Megvii Technology Limited. (aka. Face ++) – Beijing, China *Feb. 2017 – May 2017*

Research Development Intern, Engine Group

- Developed and maintained an internal deep learning framework on various platforms, including x86_64/ARM/CUDA. The framework is the core infrastructure of company-wide products and researches.
- Obtained over 3x speedup for convolution and matrix multiplication operations comparing to OpenBLAS library.

School of Informatics, CArD group, University of Edinburgh – Edinburgh, UK *June 2016 – Sep. 2016*

Research Intern

- Developed Lift-lang, a novel approach to achieve performance portability on parallel accelerators. Lift combine a high-level data parallel language with a system of rewrite rules which encode algorithmic and hardware-specific optimization choices.

Institute of High Performance Computing (IHPC), A*STAR – Singapore *June 2015 – Sep. 2015*

Research Assistant

- Deployed Caffe and object detection application on mobile platforms equipping mobile GPU cores in C++/CUDA.
- Discovered and investigated the sparsity in the applications. Designed a sparse-dense matrix-mult algorithm based on novel compression format, obtained 1.82x speedup and 46% energy savings compared to the baseline cuBLAS implementation.
- Concluded our work and submitted a paper to ACM TECS, resulted in publication.

Center for Energy-efficient Computing and Applications, Peking University – Beijing, China *Sep. 2014 – May 2017*

Undergraduate Research Intern

- Proposed and implemented an algorithm to improve parallel scalability by eliminating data-transmission latency inter/intra different nodes using C/MPI/Pthreads, which achieved nearly linear scalability on a 12-node TK-1 cluster.
- Modified and Deployed Caffe library on FPGA platform. Extracted and rebuilt CNN-based stereo matching algorithm on FPGA platform, reduced 90% of the memory required by tiling convolution layers.

SELECTED COURSE PROJECTS

Implement Raft and Fault-tolerant & Sharded Key/Value Service – Distributed Systems *Fall 2018*

- Implement raft protocol in Golang and a fault-tolerant key/value service on the top of it.
- Shard the key/value service over multiple replicated state machine clusters for a better performance.

Regional Happiness Index Assessment in NYC – Real-time and Big Data Analytics *Fall 2017*

- Implemented a scalable sentiment analysis/opinion mining system with Hadoop on Twitter data-stream in a real-time fashion.
- Developed a neighborhood quality analysis model based on the collected data using Hadoop Hive, Impala and Spark MLlib.

SELECTED PUBLICATIONS

Dayou Du, Xinfeng Xie, Qian Li, etc., Exploiting Sparsity to Accelerate Fully Connected Layers of CNN-based Applications on Mobile SoCs*, ACM Transactions on Embedded Computing Systems(TECS), Volume 17 Issue 2, January 2018, Article No.37*

SELECTED GRANTS/HONORS/AWARDS

- Ranked 8th in the 3rd ProgNova Programming Contest (ICPC division) 2017
- Guanghua Scholarship (5%) ; PKU Research Excellent Award (5%) 2016
- The 3rd Prize, "The Way to Silicon Valley" Innovation and Entrepreneurship Competition (top 5 out of 600 teams) 2016
- Honorable Mention, Mathematical Contest in Modeling 2015
- Grants of National Students Innovation and Entrepreneurship Training Program 2015
- Recognition Award, PAC National Parallel Application Challenge 2014
- The 3rd Prize, 13th "Schlumberger Cup" ACM Programming Contest 2014