Yuchen Jin

https://homes.cs.washington.edu/~yuchenj
yuchenj@cs.washington.edu

Research Interests

Machine Learning Systems, AutoML, Computer Networking, Distributed Systems

Education

Univerisity of Washington September 2015 - Present
Ph.D. candidate in Computer Science & Engineering
Advisor: Arvind Krishnamurthy

Huazhong University of Science & Technology September 2011 - June 2015
Bachelor of Computer Science & Engineering
(China National Scholarship, MediaTek Scholarship)

Industrial Internships

Research Intern, Microsoft Research, 06/2020 - 09/2020
Mentor: Sharad Agarwal

- Designed an SDWAN-based edge and wide-area network (WAN) routing optimizer and the underlying measurement infrastructure, initially for Microsoft Azure enterprise customers.

Research Intern, ByteDance AI Lab, 10/2018 - 03/2019
Mentors: Yibo Zhu, Chang Lan, Chuanxiong Guo

- Proposed and developed an automatic, practical, and effective learning rate scheduler. This learning rate scheduler has been integrated into the toolbox of the GPU training cluster at ByteDance and being used by ML researchers at ByteDance to automatically tune the learning rate of Deep Neural Networks (DNNs) over the course of training without human involvement.
- Accelerated large-scale distributed training of several DNN models developed in PyTorch, TensorFlow, and MXNet at ByteDance.

Research Intern, Microsoft Research, 06/2017 - 09/2017
Mentors: Ganesh Ananthanarayanan, Venkat Padmanabhan

- Designed and developed BlameIt, a tool that automatically and quickly localizes the faulty autonomous system (AS) when there is performance degradation between clients and the cloud. It has been in production deployment at Azure for 3 years.

Research Projects

Automatic Learning Rate Scheduler

- Tuning the learning rate schedule is one of the overheads that ML practitioners and researchers have to pay when training deep learning models. A good learning rate schedule makes DNN training faster and yields better generalization performance. I proposed and implemented an automatic and practical learning rate schedule method that helps researchers to automatically and quickly find an effective learning rate schedule for the DNNs they design, greatly reducing experimentation time. It can be applied to a wide range of DNNs trained with different optimizers due to its generality. The LR schedules auto-generated by our scheduler leads to a speedup of $1.22\times$, $1.43\times$, and $1.5\times$ when training ResNet-50, Transformer, and BERT, respectively, compared to the LR schedules in their original papers, and an average speedup of $1.31\times$ over state-of-the-art highly tuned LR schedules.

Deep Neural Networks Serving System

- I participated in the research and development of Nexus, a scalable and efficient serving system for DNN applications on a cluster of accelerators (e.g., GPUs). Nexus schedules multiple DNN jobs to meet their latency Service Level Objectives (SLOs) while efficiently utilizing GPUs. It uses techniques such as batching-aware resource allocation, complex query scheduling, and prefix batching that allows batching of common layers across DNN models. When required to stay within latency constraints at least 99% of the time, Nexus can process requests at rates 1.8-12.7× higher than the state-of-the-art systems (Tensorflow Serving and Clipper).

Probabilistic AS Relationship Inference

- Knowledge of the business relationships between Autonomous Systems (ASes) is essential to understanding the behavior of the Internet routing system, and a wide range of real-world applications rely on it. I developed a probabilistic AS relationship inference algorithm called ProbLink that overcomes the challenges in inferring hard-to-infer links, such as routing violating the valley-free assumption, limited visibility, and non-conventional peering practices. Compared to the state-of-the-art algorithm before ProbLink, ProbLink reduces the error rate for all links by $1.6\times$ and, importantly, by up to $6.1\times$ for various types of hard links. ProbLink increases the precision and recall of route leak detection by $4.1\times$ and $3.4\times$ respectively, reveals 27% more complex relationships, and increases the precision of predicting the impact of selective advertisements by 34%.

Internet Fault Localization

- I designed and implemented a tool, BlameIt, that automatically localizes the faulty AS when there is latency degradation between clients and cloud locations, using a combination of analysis of passive measurements (TCP handshake RTTs) and selective active measurements (traceroutes). Such a tool is highly valuable for global cloud providers such as Microsoft, Google, Amazon, and Facebook. BlameIt smartly leverages the passively collected measurements and a small amount of active probes for its fault localization. In doing so, BlameIt avoids the problems of intractability that stifled prior tomography-based solutions and prohibitively high overhead that plagued probing solutions based on global vantage-points. BlameIt has been in production deployment for 3 years at Azure and produces results with high accuracy at low overheads.

Routing Optimization Service for Azure

- Optimizing routing performance for Azure enterprise customers is hard because the ISPs in-between customers and Azure do not always get customers to the “best” Azure peering point due to the nature of Internet routing, and enterprise customers often have very complex networking topologies. I designed a measurement and routing optimization system called WayPoint to do optimal route selection for customers based on performance measured. WayPoint does performance-based ingress traffic engineering by cooperating with the SD-WAN devices sitting at the enterprise customers’ branch sites, which was very hard to do with BGP tricks. This project is still ongoing.

Publications

AutoLRS: Automatic Learning-Rate Schedule by Bayesian Optimization on the Fly
Yuchen Jin, Tianyi Zhou, Liangyu Zhao, Yibo Zhu, Chuanxiong Guo, Marco Canini, Arvind Krishnamurthy
(ICLR 2021)

Nexus: A GPU Cluster Engine for Accelerating DNN-Based Video Analysis
Haichen Shen, Lequn Chen, **Yuchen Jin**, Liangyu Zhao, Bingyu Kong, Matthai Philipose, Arvind Krishnamurthy, Ravi Sundaram
(SOSP 2019)

Zooming in on Wide-area Latencies to a Global Cloud Provider
Yuchen Jin, Sundararajan R, Ganesh Ananthanarayanan, Junchen Jiang, Venkat Padmanabhan, Manuel Schroder, Matt Calder, Arvind Krishnamurthy
(SIGCOMM 2019)

Stable and Practical AS Relationship Inference with ProbLink
Yuchen Jin, Colin Scott, Amogh Dhamdhere, Vasileios Giotsas, Arvind Krishnamurthy, Scott Shenker
(NSDI 2019)

UniDrive: Synergize Multiple Consumer Cloud Storage Services
Haowen Tang, Fangming Liu, Guobin Shen, **Yuchen Jin**, Chuanxiong Guo
(Middleware 2015)