Yitao Hu

Contact Information

Cell: (+1) 213-300-1057
E-mail: yitaoh@usc.edu
Homepage: sugartom.com

941 Bloom Walk, SAL 227
University of Southern California
Los Angeles, CA 90089

Research Interests

Cloud/Edge Computing, GPU Cluster Management, Networking, Localization, Video Analytics Systems, Distributed Machine Learning Systems, Real-time AR/VR Systems, Crowdsourcing Systems.

Education

University of Southern California (USC) — **Aug. 2014 - Dec. 2020 (expected)**
Ph.D. in Dept. of Computer Science — *Advised by Prof. Ramesh Govindan*

Shanghai Jiao Tong University (SJTU) — **Sept. 2010 - June 2014**
B.S. in Dept. of Electrical Engineering — *Advised by Prof. Xinbing Wang*

Publication

- Yitao Hu, *et al.*, one paper on GPU cluster management for edge computing, under review.
- Yitao Hu, *et al.*, one paper on scaling for data dependent workload, under preparation.
- Yitao Hu, Swati Rallapalli, Bongjun Ko, Ramesh Govindan. "Olympian: Scheduling GPU Usage in a Deep Neural Network Model Serving System," in *Proceedings of ACM/USENIX Middleware 2018*.
- Yitao Hu, Xiaochen Liu, Suman Nath, Ramesh Govindan. "ALPS: Accurate Landmark Positioning at City Scales," in *Proceedings of ACM UbiComp 2016*.
- Yitao Hu, Xinbing Wang, Xiaoying Gan. "Critical Sensing Range for Heterogeneous Mobile Camera Sensor Networks," in *Proceedings of IEEE INFOCOM 2014*.

Professional Experience

Networked Systems Laboratory, Research Assistant — Aug. 2014 - Present

- Built a GPU cluster management system, Rim, which can satisfy throughput and latency requirements of video and audio streaming applications, while enabling high cluster utilization.
- Designed novel algorithms to manage placement of multi-DNN pipelines, forward streaming data among distributed GPU machines, and dynamically adapt to load and failures.

IBM Research Watson, Research Intern — May 2016 - Aug. 2016

- Designed a serving middleware system, Olympian, which can schedule multiple concurrent DNNs on a single GPU to achieve fairness or service differentiation objectives.
- Developed novel techniques that can accurately estimate GPU usage, support a variety of scheduling policies, and switch between concurrent DNNs at timescales of 1-2 ms with low overhead.

Samsung Research America, Research Intern — May 2015 - Aug. 2015

- Analyzed Android's background denial log to classify the policy rule for automatic policy refinement.
- Built an internal pipeline to understand the relationship between user operation and denial log entries, as well as to extract domain knowledge among terabytes of denial logs.

Microsoft Research, Collaborator — Aug. 2014 - Apr. 2016

- Built a landmark localization system, ALPS, which can discover and localize common landmarks (e.g., stop signs and fire hydrants) at the scale of a city accurately and with high coverage.
- Developed several novel techniques that help improve the accuracy, coverage, and scalability of localization.

Selected Awards

Chun-Tsung Scholars (Top 1%, Funded by Nobel Prize winner Tsung-Dao Lee)	*2013*
National Scholarship, Minister of Education (Top 1%, Highest Scholarship in China)	*2011*
SCSK Scholarship, SCSK Corporation (Top 1%)	*2013*
Academic Excellence Scholarship, Shanghai Jiao Tong University (Top 5%)	*2011, 2013*
Travel Grant: UbiComp'16, OSDI'16, ATC'17, GTC'17, Middleware'18	