Ran Xu

martin.xuran@gmail.com | https://starsthu2016.github.io/

RESEARCH INTERESTS

Computer Vision, Machine Learning, System Optimizations with Approximation Techniques, and Mobile/Embedded Systems.

EDUCATION

Purdue University — West Lafayette, IN
Ph.D., Electrical and Computer Engineering — 8/2016 – 12/2021
Advisor: Prof. Saurabh Bagchi
Dissertation: Approximation for streaming video analytics on mobile devices
GPA: 3.8/4.0. Graduate Courses: Computer Vision, Deep Learning, Operating System, Parallel Computer Architecture, etc.

Tsinghua University — Beijing, China
B.E., Electrical Engineering — 8/2012 – 7/2016
Advisor: Prof. Yong Li
GPA: 3.8/4.0. Ranking: Top 10%

EMPLOYMENT

NVIDIA Corporation — Santa Clara, CA
Senior Deep Learning Software Engineer — 12/2021 – Present

Meta Platform Inc. (formerly Facebook, Inc) — Bellevue, WA
PhD Intern — 6/2020 – 8/2020

Adobe Research — San Jose, CA
Video Research Intern — 6/2019 – 9/2019

AT&T Labs Research — Bedminster, NJ
Research Intern — 5/2018 – 7/2018

AT&T Labs Research — Bedminster, NJ
Research Intern — 5/2017 – 7/2017

PUBLICATIONS

[C] stands for conference publications. [J] stands for journal publications. [P] stands for patents.

[C1] VideoChef: Efficient Approximation for Streaming Video Processing Pipelines
Ran Xu, Jinkyu Koo, Rakesh Kumar, Peter Bai, Subrata Mitra, Sasa Misailovic, and Saurabh Bagchi
USENIX ATC 18: The 2018 USENIX Annual Technical Conference, pp. 43-56, Jul 11-13, 2018, Boston, MA.

[C2] Pythia: Improving Datacenter Utilization via Precise Contention Prediction for Multiple Co-Located Workloads
Ran Xu, Subrata Mitra, Jason Rahman, Peter Bai, Bowen Zhou, Greg Bronevetsky, and Saurabh Bagchi
Middleware 18: In Proceedings of the 19th International Middleware Conference, pp. 146-160, Dec 10-14, 2018, Rennes, Fr.

[C3] JANUS: Benchmarking Commercial and Open-Source Cloud and Edge Platforms for Object and Anomaly Detection Workloads
Karthick Shankar, Pengcheng Wang, **Ran Xu**, Ashraf Mahgoub, and Somali Chaterji
CLOUD 20: The IEEE International Conference on Cloud Computing, pp. 590-599, Oct 18-24, 2020, virtual.

[C4] ApproxDet: Content and Contention-Aware Approximate Object Detection for Mobiles
Ran Xu, Chen-lin Zhang, Pengcheng Wang, Jayoung Lee, Subrata Mitra, Somali Chaterji, Yin Li, and Saurabh Bagchi
SenSys 20: The 18th ACM Conference on Embedded Networked Sensor Systems, pp. 449-462, Nov 16-19, 2020, virtual.

[C5] Closing-the-Loop: A Data-Driven Framework for Effective Video Summarization
Ran Xu, Haoliang Wang, Stefano Petrangeli, Viswanathan Swaminathan, and Saurabh Bagchi
ISM 20: The 22nd IEEE International Symposium on Multimedia, pp. 201-205, Dec 2-4, 2020, virtual.

[C6] Benchmarking Video Object Detection Systems on Embedded Devices under Resource Contention
Jayoung Lee, Pengcheng Wang, **Ran Xu**, Noah Weston, Venkat Dasari, Yin Li, Saurabh Bagchi, and Somali Chaterji

EMDL 21: The 5th International Workshop on Embedded and Mobile Deep Learning, pp. 19-24, Jun 25, 2021, virtual.

[C7] LiteReconfig: Cost and Content Aware Reconfiguration of Video Object Detection Systems for Mobile GPUs
Ran Xu, Jayoung Lee, Pengcheng Wang, Saurabh Bagchi, Yin Li, and Somali Chaterji
EuroSys 22: The European Conference on Computer Systems 2022, pp. 334-351, Apr 5-8, 2022, Rennes, France.

[C8] SmartAdapt: Multi-Branch Object Detection Framework for Videos on Mobiles
Ran Xu, Fangzhou Mu, Jayoung Lee, Preeti Mukherjee, Somali Chaterji, Saurabh Bagchi, and Yin Li
CVPR 22: Computer Vision and Pattern Recognition 2022, pp. 2528-2538, Jun 19-24, 2022, New Orleans, LA.

[J1] On the Opportunistic Topology of Taxi Networks in Urban Mobility Environment
Ran Xu, Yong Li, and Sheng Chen
IEEE Transactions on Big Data, vol. 6, no. 1, pp. 171-188, Mar 2020.

[J2] New Frontiers in IoT: Networking, Systems, Reliability, and Security Challenges
Saurabh Bagchi, Tarek F. Abdelzaher, Ramesh Govindan, Prashant Shenoy, Akanksha Atrey, Pradipta Ghosh, and **Ran Xu**
IEEE Internet of Things Journal, vol. 7, no. 12, pp. 11330-11346, Jul 2020.

[J3] ApproxNet: Content and Contention-Aware Video Object Classification System for Embedded Clients
Ran Xu, Rakesh Kumar, Pengcheng Wang, Peter Bai, Ganga Meghanath, Somali Chaterji, Subrata Mitra, and Saurabh Bagchi
ACM Transactions on Sensor Networks, vol. 18, no. 1, pp. 1-27, Feb 2021.

[P1] Enhancing Media Content Effectiveness using Feedback between Evaluation and Content Editing
Haoliang Wang, Viswanathan Swaminathan, Stefano Petrangeli, and **Ran Xu**
US Patent 11,170,389, 11/9/2021.

[P2] System for Content Aware Reconfiguration of Video Object Detection
Somali Chaterji, Saurabh Bagchi, and **Ran Xu**
US Patent 63/318,433 filed, 3/10/2022.

[P3] System and Methods for Content and Contention-Aware Approximate Object Detection
Somali Chaterji, Saurabh Bagchi, and **Ran Xu**
US Patent 17/710,233 filed, 3/31/2022.

[P4] SmartAdapt: Multi-Branch Object Detection Framework for Videos on Mobiles
Somali Chaterji, Saurabh Bagchi, **Ran Xu**, and Yin Li
US Patent 63/351,674 filed, 6/13/2022.

RESEARCH AND WORK EXPERIENCES

Senior Deep Learning Software Engineer 12/2021 - Present
NVIDIA Corp. Manager: Kevin Vincent

cuDNN Heuristics

· Improved the performance of the cuDNN library.

Work area: Deep Learning, Performance Modeling Programming languages: Python, C/C++, CUDA C, and SQL

Graduate Research Assistant 8/2016 – 12/2021
Dependable Computing Systems Laboratory, Purdue University Advisor: Prof. Saurabh Bagchi

Approximation-Enabled Video Analytics and Processing Systems (Publications: [C1, C4, C6, C7, C8, J3, P2, and P3])

· Designed novel deep neural networks (DNN) for video object detection running on mobile or embedded devices.
· (cont'd) Achieved 52% lower latency and 11.1% higher accuracy over YOLOv3.
· Designed a system for approximate optimization of video processing pipelines.
· (cont'd) Found the optimal approximate configuration given a quality threshold and minimized the overhead.

Research area: Computer Vision, Machine Learning, System Optimizations Programming languages: Python, C/C++, Matlab, and Latex

Workload Profiling and Scheduling in the Data-Center (Publication: [C2])

· Built tools to profile the memory contention between multiple collocated workloads in the data-center.
· Modeled the memory contention between workloads and came up with prediction models for incoming workloads.
· Designed the scheduling algorithm by putting multiple workloads on one host and guaranteeing the QOS of them.

Research area: Cloud Systems, System Optimizations, Scheduling Programming languages: Python, Shell, and Latex

PhD Intern 6/2020 - 8/2020
Meta Platforms, Inc. (formerly Facebook Inc.) Mentor: Jay Ye and Sha Meng

Graph Embedding Learning on Search Data

· Developed the training and evaluation pipelines for learning the entity embeddings from the Search data.

· Improved the recall of predicting the target group when searching from 31.9% to 40.3% with the learned embedding.

Research area: Graph Learning Programming languages: Python, Presto SQL, and Hive SQL

Video Research Intern 6/2019 - 9/2019
Adobe Research Mentor: Haoliang Wang, Stefano Petrangeli, and Vishy Swaminathan

Improving Creators' Experience on Content Creation (Publications: [C5 and P1])

· Developed closed-loop feedback on video/image ad to the creators based on the performance data.

· Developed general framework for any creation app and evaluation app to reduce the cost of searching an optimal variant.

Research area: Computer Vision, Machine Learning Programming languages: Python (Keras and Tensorflow)

Research Intern 5/2018 - 7/2018
AT&T Labs Research Advisor: Brian Amento, Hal Purdy, Kaustubh Joshi

Utility Pole Detection with Drone Cameras

· Built image classification system for detecting utility poles using Single Shot MultiBox Detector (SSD).

· Developed software simulation to verify the robustness of image classification models.

· Sped up the inference of the deep neural networks with the edge clouds.

Research area: Computer Vision, Machine Learning, Edge cloud systems Programming languages: Python (Keras), C/C++, and Latex

Research Intern 5/2017 - 7/2017
AT&T Labs Research Advisor: Rajesh Panta, Yih-Farn Chen

Erasure-Code Repair Performance in the Distributed Object Storage System (ceph)

· Designed and implemented a novel repair protocol to enable parallel repair of the data chunks.

· Debugged the performance bottleneck and reduced the repair latency.

Research area: Storage systems, Distributed Systems Programming languages: C++ and Matlab

Undergraduate Research Assistant 9/2013 - 7/2016
Future Communications & Internet LAB, Tsinghua University Advisor: Prof. Yong Li

Topology Analysis in Vehicular Networks (Publication: [J1])

· Analyzed the temporal link topology of urban large-scale vehicular networks.

· Proposed a novel opportunistic reachability graph to characterize the topology and an efficient algorithm to compute.

Research area: Vehicular Networks, Delay Tolerant Networks, Mobile Computing Programming languages: C++, Matlab, and Latex

Coverage Analysis in Cellular Networks (Undergraduate Thesis)

· Analyzed the spatial and temporal coverage pattern of urban large-scale cellular networks.

· Analyzed the signal strength especially inside buildings.

Research area: Cellular Networks Programming languages: C++ and Matlab

AWARDS

USENIX ATC Student Grant	6/2018
Outstanding Bachelors Graduates	7/2016
Scholarship, for Academic Merit, Tsinghua University	12/2013, 12/2014
Scholarship, for Sport Merit, Tsinghua University	12/2013

COMPUTER SKILLS

Programming Languages: Python, Shell, C/C++, CUDA C, SQL (Presto, Hive), Java, JavaScript, HTML, and Latex.

Machine Learning Libraries: TensorFlow (Keras), and PyTorch.

Operating Systems: Ubuntu, macOS, Windows, Virtual Machine (AWS and OpenStack).