Qingqing Cao

Room 330, Computer Science Department, Stony Brook University, NY 11794
✉: qicao@cs.stonybrook.edu ☎: (+1) 631-880-2948 🔗: awk.ai

EDUCATION

Stony Brook University Stony Brook, New York, United States
Ph.D. Candidate, Department of Computer Science Aug. 2015 - Present
Advisor: Prof. Aruna Balasubramanian & Prof. Niranjan Balasubramanian

Wuhan University Wuhan, Hubei, China
B.Eng. in Computer Science & Tech, Computer School Sept. 2011 - June 2015

RESEARCH INTERESTS

NLP Applications, **Mobile Systems**, **Edge Computing**

HONORS AND AWARDS

MobiSys 2017 Student Travel Grant Award	2017
Special CS Department Chair Fellowship	2015
Meritorious Winner in the Mathematical Contest in Modeling (MCM)	2014

PUBLICATIONS

1. **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, "DeQA: On-device Question Answering", The 17th Annual International Conference on Mobile Systems, Applications, and Services, **MobiSys 2019**.
2. **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, "MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU", 1st International Workshop on Embedded and Mobile Deep Learning, **EMDL 2017**(colocated with MobiSys).
3. Jian Xu (co-primary), **Qingqing Cao (co-primary)**, Aditya Prakash, Aruna Balasubramanian, and Don Porter. "UIWear: Easily Adapting User Interfaces for Wearable Devices", Proceedings of the 23nd ACM Annual International Conference on Mobile Computing and Networking, **MobiCom 2017**.
4. Jian Xu (co-primary), **Qingqing Cao (co-primary)**, Aditya Prakash, Aruna Balasubramanian, and Don Porter. "UIWear: Easily Adapting User Interfaces for Wearable Devices", Proceedings of the 23nd ACM Annual International Conference on Mobile Computing and Networking, **MobiCom 2017 Demo**. Video: `https://youtu.be/YEQ3HNeQnts`

PROJECTS

Optimizing Transformer Models for Faster Inference Current project
Large Transformer models have gained great success in many NLP tasks, however, running these models comes with huge cost w.r.t memory and computation, this project aims to design novel optimization techniques to reduce the inference overhead. Preliminary optimizations have shown >**2x** speedup with minimal accuracy drop.
Work in progress.

DeQA: On-device Question Answering Sept. 2018 - Mar. 2019
DeQA is a local question answering system for mobile devices that uses the state-of-the-art machine reading comprehension techniques and greatly improve end user privacy.

* Ported end to end question answering systems to mobile devices using state-of-the-art machine comprehension models with GPU support.
* Optimized question answering system pipeline, and improved performance for mobile platforms by **7** $\sim$ **13x**.

Dynamic Web QA, Microsoft Research, Jun. 2018 - Present
Work in progress. Mentor: Oriana Riva
(Paper under preparation)

Mobile Deep Learning Accelerator, Bell Labs Cambridge, Jul. 2017 - Sept. 2017
During this summer intern, I studied the performance of running deep learning models on the Movidius Neural Compute Stick accelerator. Mentor: Nic Lane
(Paper under preparation)

MobiRNN: Efficient RNN Execution on Mobile Mar. 2017 - Jun. 2017
MobiRNN is a mobile specific optimization library for RNNs that focusses on offloading deep learning tasks to the mobile GPU.

UIWear: Easily Adapting User Interfaces for Wearable Devices Jan. 2016 - Dec. 2016
UIWear is a "write once and extend to many" programming framework for wearable devices enabling users to use smartphone applications from any of their wearable devices.

* Developed I/O multiplexing mechanism to enable multi-device user interaction. Created UI metaprogram to automatically build companion apps for wearables like smartwatch with minimal developer effort.
* Optimized UIWear protocol (for UI data cross-device communication and rendering) and improved latency by **27%** compared to existing systems.
* Implemented UIWear system on Android Phone and Watch.

SERVICE

Technical Committee Member of MobiSys PhD Forum	2018
Reviewer for IEEE Transactions on Mobile Computing	2018

SKILLS

Python, Java, Android, C.