



Qingqing Cao

✉: qicao@cs.stonybrook.edu : [linkedin.com/in/qqcao](https://www.linkedin.com/in/qqcao) : [awk.ai](https://github.com/awk.ai)

EDUCATION

Stony Brook University Aug. 2015 - June 2021 (expected)
Ph.D. in Computer Science

Wuhan University Sept. 2011 - June 2015
B.Eng. in Computer Science & Tech

HIGHLIGHTS

I have 3+ years of research experience in *deep learning for NLP*, *mobile computing*, and *machine learning systems*. I have focused on building efficient and practical NLP systems for both edge devices and the cloud, such as on-device question answering (MobiSys 2019), faster Transformer models (ACL 2020), and accurate energy estimation of NLP models.

EXPERIENCE

Research Assistant @ Stony Brook University, US Jun. 2016 - Present
Advisors: Prof. Aruna Balasubramanian & Prof. Niranjan Balasubramanian

Research Intern @ Microsoft Research Redmond, US Jun. 2018 - Aug. 2018
Mentor: Oriana Riva Topic: dynamic business web queries

Research Intern @ Bell Labs Cambridge, UK Jul. 2017 - Sept. 2017
Mentor: Nicholas Lane Topic: mobile deep learning accelerators

PUBLICATIONS

1. **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, “Towards Accurate and Reliable Energy Measurement of NLP Models”, First Workshop on Simple and Efficient Natural Language Processing, **SustaiNLP@EMNLP 2020**. Paper: <https://awk.ai/assets/sustainlp.pdf>
Summary: Accurate and reliable energy measurements are critical when choosing and training large-scale NLP models. I use a hardware power meter to accurately measure energy and quantify the error (>20%) of existing software-based energy measurements. I find inaccurate measurements can cause misleading design choices and advocate for more accurate energy estimation methods that consider underlying hardware and resources utilizations.
2. **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, “DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering”, The 58th annual meeting of the Association for Computational Linguistics, **ACL 2020**. Paper: <https://awk.ai/assets/deformer.pdf>
Summary: Pre-training large Transformers is expensive and the inference in them is prohibitively slow. I design DeFormer that decomposes pre-trained Transformers to enable faster inference for QA without repeating the pre-training. DeFormer achieves >**3.1x** speedup inference speedup and >**65%** memory reduction with minimal ($\sim 1\%$) accuracy loss.
3. **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, “DeQA: On-device Question Answering”, The 17th Annual International Conference on Mobile Systems, Ap-

plications, and Services, **MobiSys 2019**. Paper: <https://awk.ai/assets/deqa.pdf>

Summary: DeQA is an on-device question answering system to help mobile users find information more efficiently without privacy issues. Deep learning-based QA models are slow and unusable on mobile. I design the latency- and memory- optimizations for the QA models to run locally on mobile devices. DeQA effectively reduces the memory footprint and improves the QA latency **6 ~ 13x** with minimal accuracy drop ($< 1\%$).

4. **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@MobiSys 2017**. Paper: <https://awk.ai/assets/mobirnn.pdf>

Summary: MobiRNN is a mobile specific optimization library for RNNs that focuses on offloading deep learning tasks to the mobile GPU.

5. 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 23rd ACM Annual International Conference on Mobile Computing and Networking, **MobiCom 2017**. Paper: <https://awk.ai/assets/uiwear.pdf>
6. 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 23rd ACM Annual International Conference on Mobile Computing and Networking, **MobiCom 2017 Demo**. Video: <https://youtu.be/YEQ3HNeQnts>

AWARDS

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

SERVICE

NAACL 2021 Program Committee

Eurosys 2021 Shadow TPC

Technical Committee Member of ACL 2020 (demo track)

Technical Committee Member of MobiSys 2018 PhD Forum

Reviewer for IEEE Transactions on Mobile Computing 2018

Secondary Reviewer for IMC 2017, MobiSys 2017-2020, MobiCom 2019-2021, EuroSys 2019, SIGCOMM 2019-2020, EMNLP 2020.

Student Volunteer for MobiSys 2017 and ACL 2020

Stony Brook CS Grad Buddies Program Mentor

SKILLS

Python, Java, C, TensorFlow, PyTorch.