



# Qingqing Cao

✉: qicao@cs.stonybrook.edu     linkedin.com/in/qqcao     awk.ai

## EDUCATION

---

Stony Brook University  
**Ph.D.** in Computer Science

Aug. 2015 - May 2021 (expected)

Wuhan University  
**B.Eng.** in Computer Science & Tech

Sept. 2011 - June 2015

## HIGHLIGHTS

---

I have 3+ years of research experience in **natural language processing**, **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. [SustaiNLP@EMNLP 2020] **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, “Towards Accurate and Reliable Energy Measurement of NLP Models”. Paper: <https://awk.ai/assets/sustainlp.pdf>  
*Summary:* Accurate energy measurement is critical for choosing and training large NLP models and deploying to battery-powered mobile devices. Existing utilization-based software methods do not address issues like power lag, tail energy issues. Non-utilization behaviors such as data movement in GPUs also cause energy. Resource profiling should avoid high overhead. I use a hardware power meter to measure energy accurately and quantify the error (>20%) of existing software measurements. I find current software measurements without calibration are inaccurate and cause misleading design choices.
2. [ACL 2020] **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, “DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering”. 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 (~ 1%) accuracy loss.
3. [MobiSys 2019] **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, “DeQA: On-device Question Answering”. 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. [EMDL@MobiSys 2017] **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, “MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU” 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. [MobiCom 2017] Jian Xu (co-primary), **Qingqing Cao (co-primary)**, Aditya Prakash, Aruna Balasubramanian, and Don Porter. “UIWear: Easily Adapting User Interfaces for Wearable Devices”. Paper: <https://awk.ai/assets/uiwear.pdf>
6. [MobiCom 2017 demo] Jian Xu (co-primary), **Qingqing Cao (co-primary)**, Aditya Prakash, Aruna Balasubramanian, and Don Porter. “UIWear: Easily Adapting User Interfaces for Wearable Devices”. Demo video: <https://youtu.be/YEQ3HNeQnts>

## AWARDS

---

CDAC Rising Stars in Data Science, University of Chicago	2020
MobiSys Student Travel Grant, ACM SIGMOBILE	2017
Special CS Department Chair Fellowship, Stony Brook University	2015
<b>Meritorious Winner</b> in the Mathematical Contest in Modeling, COMAP	2014
National Scholarship (top <b>0.2%</b> ), China Ministry of Education	2013
National Endeavor Fellowship (top <b>3%</b> ), China Ministry of Education	2012, 2014

## SERVICE

---

**Program Committee:** NAACL 2021, Eurosys 2021 (shadow), ACL 2020 (demo), MobiSys 2018 (PhD forum), IEEE Transactions on Mobile Computing (reviewer).

**Secondary Reviewer:** EMNLP 2020, IMC 2017, EuroSys 2019, MobiSys 2017~2020, MobiCom 2019~2021, SIGCOMM 2019~2020.

**Volunteering Service:** Student volunteer for MobiSys 2017 and ACL 2020, mentor for Stony Brook CS Grad Buddies Program.

## SKILLS

---

**Programming:** Python, C, Java.

**Machine Learning:** TensorFlow, PyTorch, Scikit-learn, Numpy, XGBoost.

## COURSES

---

**Computer Systems:** Analysis of Algorithms (CSE548), Operating Systems (CSE506), Fundamentals of Computer Networks (CSE534).

**Machine Learning:** Machine Learning (CSE512), Artificial Intelligence (CSE537).