Qingqing Cao

im@awk.ai awk.ai linkedin.com/in/qqcao

Highlights

I have 8+ 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 (visual) question answering, faster Transformer models, and accurate energy estimation of NLP models.

Work Experience

AI Research Scientist @ Apple, US	Jan 2024 - Present
Postdoctoral Scholar @ University of Washington, US	Oct 2021 - Jan 2024
Research Intern @ Microsoft Research Redmond, US	Jun 2018 - Aug 2018
Research Intern @ Bell Labs Cambridge, UK	Jul 2017 - Sept 2017

Education

Stony Brook University **Ph.D.** in Computer Science	Aug. 2015 - Sept. 2021
Wuhan University **B.Eng.** in Computer Science & Tech	Sept. 2011 - June 2015

Awards

Postdoc Research Award, University of Washington	2022, 2023
Catacosinos Fellowship (**2** out of 232 PhD students), Stony Brook University	2021
CDAC Rising Stars in Data Science, University of Chicago	2021
MobiSys Student Travel Grant, ACM SIGMOBILE	2017
Special CS Department Chair Fellowship, Stony Brook University	2015
Meritorious Winner in the Mathematical Contest in Modeling, COMAP	2014
National Scholarship (top **0.2%**), China Ministry of Education	2013

Recent Publications

1. **[arxiv 2024]** Sachin Mehta, Mohammad Hossein Sekhavat, **Qingqing Cao**, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. *OpenELM: An Efficient Language Model Family with Open Training and Inference Framework*. Apple Machine Learning Research, Apple WWDC 2024
2. **[ICML 2024]** Bowen Zhao, Hannaneh Hajishirzi, and **Qingqing Cao**. *APT: Adaptive Pruning and Tuning Pretrained Language Models for Efficient Training and Inference*. Oral (1.5%)
3. **[ICLR 2024] Qingqing Cao**, Sewon Min, Yizhong Wang, and Hannaneh Hajishirzi. *BTR: Binary Token Representations for Efficient Retrieval Augmented Language Models*. Spotlight (5%)
4. [**ACL 2023**] **Qingqing Cao**, Bhargavi Paranjape, Hannaneh Hajishirzi, *PuMer: Pruning and Merging Tokens for Efficient Vision Language Models*.

Service

Area Chair: ACL 2024, ACL Rolling Review. Student Area Chair for Computer Science PhD Admissions 2023, the University of Washington

Program Committee: COLM 2024, ACL Rolling Review, NeurIPS 2023, EMNLP 2021-2023, ACL 2021-2023, NAACL 2021, Eurosys 2021 (shadow), ACL 2020 (demo), MobiSys 2018 (PhD forum), IEEE Transactions on Mobile Computing reviewer (2018, 2023). Student Committee Member for Computer Science PhD Admissions 2022, University of Washington

Teaching and Volunteering Service: Student volunteer for MobiSys 2017 and ACL 2020, mentor for Stony Brook CS Grad Buddies Program. Instructor for Women in Science & Engineering (WISE) 380.

Previous Publications

1. [**ACL 2023**] Qing Zhang et al. including **Qingqing Cao**, *A Survey for Efficient Open Domain Question Answering*.
2. [**TACL 2023**] Marcos Treviso et al. including **Qingqing Cao**, *Efficient Methods for Natural Language Processing: A Survey*.
3. [**IMWUT/UbiComp 2022**] **Qingqing Cao**, Prerna Khanna, Nicholas D. Lane, Aruna Balasubramanian, *MobiVQA: Efficient On-Device Visual Question Answering*.
4. [**ACL 2021**] **Qingqing Cao**, Yash Lal, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, *IrEne: Interpretable Energy Prediction for Transformers*.
5. [**EMNLP 2021** Demo] Yash Kumar Lal, Reetu Singh, Harsh Trivedi, **Qingqing Cao**, Aruna Balasubramanian, Niranjan Balasubramanian, *IrEne-viz: Visualizing Energy Consumption of Transformer Models*.
6. [**ACL 2020**] **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, *DeFormer: Decomposing Pretrained Transformers for Faster Question Answering*.
7. [**SustaiNLP** workshop@EMNLP 2020] Qingqing Cao, Aruna Balasubramanian, Niranjan Balasubramanian, *Towards Accurate and Reliable Energy Measurement of NLP Models*.
8. [**MobiSys 2019**] **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian. *DeQA: On-device Question Answering*.
9. [**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*.
10. [**EMDL** workshop@MobiSys 2021] **Qingqing Cao**, Alexandru Eugen Irimiea, Mohamed Abdelfattah, Aruna Balasubramanian and Nicholas D. Lane, *Are Mobile DNN Accelerators Accelerating DNNs?*
11. [**EMDL** workshop@MobiSys 2017] **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian, *MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU*.
12. [**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*.