



Qingqing Cao

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

HIGHLIGHTS

I have 5+ 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.

RECENT PUBLICATIONS

1. **[under review] Qingqing Cao**, Sewon Min, Yizhong Wang, Hannaneh Hajishirzi. BTR: Binary Token Representations for Efficient Retrieval Augmented Language Models.
2. **[under review] Hao Peng, Qingqing Cao**, 10 others, Noah A. Smith, Hannaneh Hajishirzi. Efficiency Pentathlon: A Standardized Arena for Efficiency Evaluation.
3. **[NeurIPS 2023] Adaptive Representations for Semantic Search**. Aniket Rege*, Aditya Kusupati*, Sharan Ranjit, Alan Fan, **Qingqing Cao**, Sham Kakade, Prateek Jain, and Ali Farhadi.
4. **[ACL 2023] Qingqing Cao**, Bhargavi Paranjape, Hannaneh Hajishirzi, PuMer: Pruning and Merging Tokens for Efficient Vision Language Models.
5. **[IMWUT/UbiComp 2022] Qingqing Cao**, Perna Khanna, Nicholas D. Lane, Aruna Balasubramanian, MobiVQA: Efficient On-Device Visual Question Answering.
6. **[ACL 2021] Qingqing Cao**, Yash Lal, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, IrEne: Interpretable Energy Prediction for Transformers.
7. **[ACL 2020] Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, Niranjan Balasubramanian, DeFormer: Decomposing Pretrained Transformers for Faster Question Answering.

EXPERIENCE

Postdoctoral Scholar @ University of Washington, US Mentor: Prof. Hannaneh Hajishirzi	Oct. 2021 - Present
Research Intern @ Microsoft Research Redmond, US Mentor: Oriana Riva Topic: dynamic business web queries	Jun. 2018 - Aug. 2018
Research Intern @ Bell Labs Cambridge, UK Mentor: Nicholas Lane Topic: mobile deep learning accelerators	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

SERVICE

Program Committee: 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).

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

- [**ACL 2023**] Qing Zhang et al. including **Qingqing Cao**, A Survey for Efficient Open Domain Question Answering.
- [**TACL 2023**] Marcos Treviso et al. including **Qingqing Cao**, Efficient Methods for Natural Language Processing: A Survey.
- [**EMNLP 2021 Demo**] Yash Kumar Lal, Reetu Singh, Harsh Trivedi, **Qingqing Cao**, Aruna Balasubramanian, Niranjana Balasubramanian, IrEne-viz: Visualizing Energy Consumption of Transformer Models.
- [**SustaiNLP workshop@EMNLP 2020**] Qingqing Cao, Aruna Balasubramanian, Niranjana Balasubramanian, Towards Accurate and Reliable Energy Measurement of NLP Models.
- [**MobiSys 2019**] DeQA: On-device Question Answering. Qingqing Cao, Niranjana Balasubramanian, Aruna Balasubramanian.
- [**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.
- [**EMDL workshop@MobiSys 2021**] Qingqing Cao, Alexandru Eugen Irimiea, Mohamed Abdelfattah, Aruna Balasubramanian and Nicholas D. Lane, Are Mobile DNN Accelerators Accelerating DNNs?
- [**EMDL workshop@MobiSys 2017**] Qingqing Cao, Niranjana Balasubramanian, Aruna Balasubramanian, MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU.
- [**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.