My research vision is to bring together systems and natural language processing (NLP) research to make language processing systems and applications more **energy-efficient**, **privacy-preserving**, and **run faster** and more **widely applicable** to heterogeneous hardware. To this end, my current research focuses on question answering (QA) systems.

QA systems essentially power many real-world applications ranging from intelligent personal assistants (like Alexa, Siri, and Google Assistant) to commercial search engines such as Google and Bing. However, QA systems use complex deep learning models that run in the cloud, requiring expensive energy and compute resources. Even worse, they cannot run on mobile devices, making on-device, privacy-preserving QA impractical. Much existing work has centered on smaller NLP models by designing compact architectures or compressing large models; both require huge retraining costs. My research adopts a different paradigm to make NLP systems like QA run efficiently: digging deeper into which components in the NLP models are effective and how they interact with hardware resources like CPU, memory, and GPU.

My work combines systems principles with a deep understanding of NLP models. For example, I show how to run complex NLP models on mobile devices using fine-grained bottleneck and critical path analysis and exploring data caching and reuse opportunities [**MobiSys'19, MobiCom'17, EMDL'17**]. I have also made contributions in efficient NLP models by developing efficient model architecture variants that identifies and removes the representation dependency in the attention blocks of Transformers [**ACL'20**]. Recently, I have focused on modeling the energy consumption of large NLP models, preliminary results [**SustaiNLP'20**] show existing software energy measurements without calibration are problematic and using hardware power meters provide more accurate energy measurements. Earlier in my research, I worked on the UIWear [**MobiCom'17**] project that made mobile applications more efficient and more accessible. Additionally, I researched how to make web QA systems more practical for dynamic business-related queries [**WWW'21** under review]. In what follows, I describe concrete examples that paint a picture of my research career so far.

Efficient Privacy-Preserving On-Device NLP Systems. On the systems side, my work focuses on developing on-device NLP systems that preserve data privacy and provide device-wide capabilities for mobile users. Existing deep learning QA systems are designed for the cloud and cannot run efficiently on mobile phones. To address this problem, I developed DeQA [**MobiSys'19**], a suite of latency- and memory- optimizations that adapt state-of-the-art QA systems to run locally on mobile devices. DeQA effectively reduces QA latency on mobile phones *from over a minute to under 5s.* DeQA moves the neural encoding computation off the critical path by pre-computing them, and then loads this representation on-demand at runtime. DeQA also uses a dynamic early stopping algorithm that predicts when further processing will not yield better accuracy, so paragraph processing can be stopped early. DeQA reduces memory requirements by loading paragraph-level partial index into memory and replacing in-memory embeddings lookups with a key-value database. Earlier in my research, I developed the MobiRNN [**EMDL'17**@MobiSys] framework to run RNNs on the mobile phone GPUs efficiently. The core idea is to group the cells in RNNs in a coarse-grained manner so that the underlying low-level compute library can automatically decide offloading units to avoid scheduling overheads.

Efficient Models and Practical Systems for Sustainable NLP. On the NLP side, my research focuses on developing efficient NLP algorithms that run faster and are more energy efficient for both the mobile and cloud. State-of-the-art NLP models like large pre-trained Transformers are effective in many NLP tasks but consume enormous computing resources. There is an increasing need to deploy these models in large-volume web-scale services like Google and Bing search engines for search and QA. To solve large Transformers' resource efficiency problems for QA tasks, I designed DeFormer ACL

’20, which decomposes Transformer-based NLP models such as BERT that removes the dependencies between question and paragraph processing in the lower layers of the model. This decomposition allows DeFormer to precompute paragraph processing, improving QA inference latency by *over 4 times* without sacrificing accuracy. A key design decision in DeFormer is to decompose Transformer models without requiring pre-training, so that this expensive process does not have to be repeated.

Ongoing Work: Energy Consumption of NLP Models. The energy consumption problem has spired a growing interest in the NLP community because accurate energy measurement is critical for reducing the costs in training large NLP models and deploying them to battery-powered mobile devices. However, existing energy work in NLP often underestimates the challenges and uses uncalibrated software measurement approaches. I recently studied the hardware-based approach to accurately measure energy and quantify the error (>20%) of existing software measurements [**SustaiNLP’20**]. I found that existing utilization-based software methods are inaccurate and cause misleading design choices because they do not address issues like power lag, tail energy, or non-utilization behaviors (e.g. data movement in GPUs). My ongoing work addresses these challenges by abstracting meaningful estimation features of the NLP models and profiling runtime resources usage besides utilization.

Future Work. My long-term research goal is to build efficient, intelligent algorithms and systems that are more privacy-focused, faster to learn, and more energy-efficient to run on heterogeneous hardware. In the future, I strongly believe it is promising to work on the following research problems:

(1) **Efficient On-Device Hardware-Aware Multi-Modality Systems.** NLP research advances rapidly, while the need for deployments to heterogeneous devices is ever increasing. It is insufficient to focus on specific models that work only for text. I plan to broaden the research scenarios to include multi-modal applications (such as visual question answering) that help millions of visually-impaired people. However, on-device multi-modal applications present unique challenges; for example, processing images or videos consume much energy on battery-powered devices. Research multi-modal applications will make systems and NLP optimizations scale to a broader range of future NLP applications that can run faster with fewer resources and less energy footprint on more diverse hardware.

(2) **Data-Efficient Learning Algorithms for NLP.** Existing successful NLP models are pre-trained on large amounts of data. However, this trend is unsustainable as the computing resources cannot support exponentially growing models and data volumes (e.g., GPT3 has 175 billion parameters and was trained on 400 billion tokens). My prior work in DeFormer focuses on reducing the model computations. I believe it is crucial to develop more sustainable NLP algorithms that can learn faster using less data.

References

[**WWW’21**(under review)] **Qingqing Cao**, Oriana Riva, Aruna Balasubramanian, and Niranjan Balasubramanian. Bew: Towards Answering Business-entity-related Web Questions.

[**SustaiNLP’20** workshop@EMNLP] **Qingqing Cao**, Aruna Balasubramanian, and Niranjan Balasubramanian. Towards Accurate and Reliable Energy Measurement of NLP Models.

[**ACL’20** (long paper)] **Qingqing Cao**, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering.

[**MobiSys’19**] **Qingqing Cao**, Noah Weber, Niranjan Balasubramanian, and Aruna Balasubramanian. DeQA: On-Device Question Answering.

[**EMDL’17** workshop@MobiSys] **Qingqing Cao**, Niranjan Balasubramanian, Aruna Balasubramanian. MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU.

[**MobiCom’17**] Jian Xu (co-primary), **Qingqing Cao** (co-primary), Aditya Prakash, Aruna Balasubramanian, and Don Porter. UIWear: Easily Adapting User Interfaces for Wearable Devices.