Rongxiang Wang (434) 257-9768 | rxpwang.github.io | waq9hw@virginia.edu

EDUCATION

University of Virginia, **PhD in CS** (GPA 4.0/4.0) Aug. 2022 – Current

Research theme: on-device ML; current focus: efficient serving systems for speech models. Advisor: Felix Lin

Tsinghua University, Bachelor of Engineering in Automation Aug. 2017 – Jun. 2022

PUBLICATIONS

[1] "Efficient Whisper on Streaming Speech," Rongxiang Wang, Zhiming Xu and Felix Xiaozhu Lin, in Arxiv, 2024

[2] "Turbocharge Deep Speech Understanding with Pilot inference," Rongxiang Wang and Felix Xiaozhu Lin, in MobiCom, 2024

[3] "AnA: An Attentive Autonomous Driving System (in camera-ready preparation)," Wonkyo Choe, Rongxiang Wang, and Felix Xiaozhu Lin, in ASPLOS, 2025

[4] "Accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data with scCASE," Songming Tang, Xuejian Cui, Rongxiang Wang, Sijie Li, Siyu Li, Xin Huang, and Shengquan Chen, in Nature Communication, 2024

[5] "ASTER: accurately estimating the number of cell types in single-cell chromatin accessibility data," Shengquan Chen, Rongxiang Wang, Wenxin Long, and Rui Jiang, in Bioinformatics, 2023

RESEARCH PROJECTS

A real time serving system for speech foundation model University of Virginia, Jan 2024 – Dec 2024

- Targets edge devices with heterogeneous processors, aiming to speed up speech foundation models for streaming speech processing and reduce the per word latency of the system.
- Proposes hush word, a novel adversarial attack to reduce the encoding redundancy. Proposes beam pruning, a reference guide method tailored for streaming tasks to reduce the decoding redundancy. Proposes CPU/GPU pipelinging to better utilize the hardware resources and reduce the overall latency.
- Achieves a 2x per word latency speedup, with as low as 0.5s per word latency and ~7 W power consumption.

On-device deep speech understanding University of Virginia, Mar 2023 – Dec 2023

- Targets mobile devices with limited hardware capabilities, aiming to speed up attention-based model local processing in streaming scenarios, as well as reduce offloading in collaboration with the cloud.
- Proposes pilot inference, which periodically processes partial data to attain tentative information that helps with local processing speed up and selective offloading. Proposes beam reduction, beam search termination prediction and CTC prefix scoring speedup for local execution, perplexing score / RNN / CNN based selective offloading for collaboration with the cloud.
- Achieves a 2x local processing speed up and reduces offloading by 50% in collaboration with the cloud.

Redesign of autonomous driving stack University of Virginia, Aug 2022 – Feb 2023

- Targets in-vehicle autonomous driving scenarios, aiming to reduce the computational cost of the autonomous driving system and make it more responsive.
- Proposes an attentive method that helps the perception module better focus on important regions based on the information from the downstream planning module.

- Reduces the computation cost of the autonomous driving system by 44% and avoids collisions by 2x.

Single cell genomic data enhancement (undergraduate thesis) Tsinghua University, Aug 2021 – May 2022

- Targets single-cell data with massive sampling loss, aiming to recover the data and enhance the data quality.
- Proposes an optimization-based low-rank matrix factorization method to assist with data recovery.
- Improves the data quality, as reflected in clustering metrics by 30%.

TEACHING

Teaching Assistant for the undergraduate Operating System (CS4414), Spring 2023, Spring 2024. Responsibility: help ~150 students to understand the course project, which builds an OS prototype on Raspberry Pi3 and covers OS bootup, process scheduling, memory management, multi thread programming and profiling, file system, and TEE.

SKILLS

Programming Languages Python, C, C++

Software Pytorch, Numpy, Pandas, Linux

Specializations Machine learning, Mobile computing, Speech processing, Transformer-based foundation models