Zirui Wang

eeb9sd@virginia.edu | Github | Homepage

Research Interests

Computer Systems, Storage Systems, Systems for ML

Education

University of Virginia — Charlottesville, VA
PhD in Computer Science — *Sept. 2024 – Present*

Boston University — Boston, MA
Master of Science in Computer Science — *Sept. 2022 – Jan. 2024*

Hangzhou Dianzi University — Hangzhou, China
Bachelor of Engineering in Computer Science and Technology — *Sept. 2018 – Jun. 2022*

Publication

- ***NSDI' 26*** *Towards Efficient LLM Storage Reduction via Tensor Deduplication and Delta Compression.*
  23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI'26)
  Zirui Wang, Tingfeng Lan, Zhaoyuan Su, Juncheng Yang, Yue Cheng.

- ***Under Review*** *Efficient and Workload-Aware LLM Serving via Runtime Layer Swapping and KV Cache Resizing.*
  Zhaoyuan Su, Tingfeng Lan, **Zirui Wang**, Juncheng Yang, Yue Cheng.

- ***VLDB'24*** *Everything You Always Wanted to Know About Storage Compressibility of Pre-Trained ML Models but Were Afraid to Ask.*
  50^{th} International Conference on Very Large Data Bases (**VLDB'24**).
  Zhaoyuan Su, Ammar Ahmed, **Zirui Wang**, Ali Anwar, Yue Cheng.

- ***ICCV'21*** *Temporal Cue Guided Video Highlight Detection with Low-Rank Audio-Visual Fusion.*
  International Conference on Computer Vision (**ICCV'21**).
  Qinghao Ye*, Xiyue Shen*, Yuan Gao*, **Zirui Wang***, Qi Bi, Ping Li, Guang Yang.

Research Projects

ZipLLM: Efficient Storage Optimization for LLM Repositories — Sep. 2024 – Apr. 2025
University of Virginia — *Charlottesville, VA*

- Conducted the first large-scale measurement study on Hugging Face model repositories, identifying structural redundancy in fine-tuned models.
- Designed BitX, a fast, lossless XOR-based delta compressor for LLM weights, achieving nearly 50% storage savings and over 2GB/s throughput.
- Built a full pipeline ZipLLM combining tensor-level deduplication and BitX compression, significantly outperforming chunk-based and model-oblivious baselines.

Novel Cache Policy Design for Optimizing LLM Download Traffic — Jan. 2024 – Dec. 2024
University of Virginia — *Charlottesville, VA*

- Collected and organized traces of LLM download behavior, conducting analysis and visualization.
- Implemented initial optimizations based on the **AdaptSize** algorithm and modified the **libCacheSim** library to support AdaptSize, achieving at least a **4%** improvement in hit ratio for LLM download traces.
- Conducted cache simulation experiments on different traces, analyzing and visualizing the results.

ELF Compression Algorithm Acceleration — Jul. 2023 – Dec. 2023
Remote work with Prof.Cheng — *Boston, MA*

- Optimized ELF compression algorithm to increase the compression rate of 32-bit float floating-point numbers from around 1.2x to 1.25x.

- Achieved parallel acceleration of ELF algorithm on SmartSSD, and used P2P transfer to significantly improve the I/O throughput, with the compression throughput reaching 1.3 GB/s in a single compute unit.
- The paper was submitted to **VLDB'24**.

Video Highlight Detection Based on Deep Learning Method Sept. 2020 – Jul. 2021

Hangzhou Dianzi University *Hangzhou, China*

- Used a hierarchical temporal context coding structure and proposed a low-rank decomposition-based video and audio fusion method to improve the detecting accuracy and speed. Successfully **exceeding the SOTA level** and improving the mAP value from 0.584 to 0.629. Paper accepted by **ICCV'21**.

Projects

Stream Processing System with State Disaggregation Feb. 2023 – May 2023

- Designed and implemented a streaming data processing system capable of automated task allocation, load balancing, and state storage disaggregation.
- Implemented operators that handle the computation of stream data, including **stateless** operators such as Filter, KeyBy, Map, Union, and **stateful** operators such as Reduce, Count, and Sliding Window.
- Wrote test scripts in Java to test the latency of the system using local storage as well as remote state storage. Used **Prometheus** for real-time status monitoring of system latency.

Alibaba Tianchi Global Video Cloud Innovation Challenge Mar. 2021 – Jul. 2021

- According to the competition problem, the Fast Instance Segmentation + Mask Refinement method is proposed to solve the problems of motion blur, frequent scene switching, and character edge refinement, making it possible to perform segmentation quickly and accurately.
- The competition ended up with a **bronze prize**  (ranking 5/2904).

Technical Skills

Programming Language: C/C++, Rust, Python, Java, Go

Framework: PyTorch, Flink

Tools&Platforms: Git, Docker, Redis, Linux, SQL, Github, AMD Vitis