Kun Wang

Kowloon Tong, Hong Kong
☎ 852-63527911 ✉ wk840103821@gmail.com https://moveisthebest.github.io/

Education

City University of Hong Kong **Sep. 2020 – Aug. 2024**
Ph.D. in Computer Science *Hong Kong SAR, China*

Xidian University **Sep. 2016 – Jul. 2020**
B.E. in Software Engineering *Xi'an, China*

Research Publications

- **ϕ-Cache: Significantly reducing the memory occupation of KV Cache during LLMs decoding.**
  Kun Wang, Zimu Zhou, and Zhenjiang Li.
  [On-going Work]
- **LATTE: Layer Algorithm-aware Training Time Estimation for Heterogeneous Federated Learning.**
  Kun Wang, Zimu Zhou, and Zhenjiang Li.
  [CCF-A] **ACM MobiCom, 2024, Conditionally Accept**
- **SwapNet: Efficient Swapping for DNN Inference on Edge AI Devices Beyond the Memory Budget.**
  Kun Wang, Jiani Cao, Zimu Zhou, and Zhenjiang Li.
  [CCF-A] **IEEE Transactions on Mobile Computing, 2024**
- **A Workload-Awar DVFS Robust to Concurrent Tasks for Mobile Devices.**
  Chengdong Lin, Kun Wang, Zhenjiang Li and Yu Pu.
  [CCF-A] **ACM MobiCom, 2023**
- **Enhancing Human Motion Sensing with synthesized Millimeter-Waves.**
  Xiaotong Zhang, Kun Wang, Zhenjiang Li and Jin Zhang.
  [CCF-B] **IEEE/ACM IPSN Poster 2024**

Research Projects

Optimizing the memory occupation of KV-Cache during LLMs decoding. *May. 2024 – Present*

- Observing huge memory occupation from KV-Cache in long-context LLM Serving.
- Proposing ϕ-Cache, a new component replaces traditional KV-Cache, significantly reducing the KV-Cache memory usage from large-and-keep increasing to tiny-and-fixed size.
- Developing an efficient ϕ-Cache update mechanisms, significantly reducing attention computation costs.

Accurate Training Time Estimation for Heterogeneous Federated Learning. *June. 2023 – Mar. 2024*

- Observing ML frameworks (e.g., PyTorch, TensorFlow, MindSpore) will choose different algorithms (e.g., GEMM, FFT, Winograd) for layer implementation due to varying runtime optimizations, which affects the training time estimation.
- Designing LATTE, a lightweight and accurate training time estimator for DNNs in heterogeneous devices.
- Utilizing LATTE in heterogeneous FL to optimize sub-model allocation, enabling similar training times in heterogeneous devices and accelerating FL convergence.

Swapping-based DNN Inference on Mobile Edge Devices Beyond the Memory Budget. *Sept. 2021 – April. 2023*

- Executing DNNs on mobile edge devices face memory constraints. Proposing splitting DNNs into blocks and swapping them between storage and memory for block-by-block execution.
- Observing redundant memory copy existing in 1) when ML frameworks invoke GPU and 2) when DNNs model weights loading into memory, causing significant latency overhead during block swapping.
- Optimizing PyTorch's CPU and GPU memory allocation source code to implement a Unified Memory Allocator and design an efficient weight loading method, fully eliminating redundant copies and significantly reducing latency overhead.

Programming Skills

Programming Languages: CUDA, Python, C/C++, Java
Technologies/Frameworks: PyTorch, TensorFlow, Linux, Git, Docker

Awards and Honors

Postgraduate Scholarship, CityU, Hong Kong SAR	2020-2024
National Scholarship, Ministry of Education, China	2019
National Scholarship, Ministry of Education, China	2018
Second Prize Scholarship, Xidian University, China	2017