Yuwei An

Pittsburgh | ayw.sirius19@gmail.com | +1 4129805360

Education

Tsinghua University, BS in Computer Science and Technology Sept 2019 – June 2023

- GPA: 3.77/4.0 (Transcript)
- **Coursework:** Computer Architecture, Artificial Intelligence, Comparison of Learning Algorithms, Computational Theory

Carnegie Mellon University, MS in Electrical and Computer Engineering Sept 2023 – Now

- GPA: 4.0/4.0 (Transcript)
- **Coursework:** Deep Generative Models, Natural Language Processing,

Experience

Research Intern UChi **Advisor** Junchen Jiang, SeoJin Park Sept 2024 - Now

- Responsible for the FastRaG Project.
- The FastRaG Project aims to develop a high-throughput serving system for retrieval-augmented generation (RaG). The project focuses on optimizing the accuracy-latency tradeoff curve in RaG scenarios through innovations in attention mechanisms and KV-cache management. The team is targeting a paper submission to ATC 2025.

Research Intern CMU **Advisor** Beidi Chen June 2024 - Now

- Responsible for the MoE Inference Speedup System Project.
- By now we have implemented the system work for MoE Inference with **Huge** MoE model such as Deepseek and Qwen. The current work is algorithm work with expert scheduling for less communication and memory cost

Research Intern CMU **Advisor** Beidi Chen & Chenyan Xiong Feb 2024 - May 2024

- Responsible for the multi-batch dynamic pruning algorithm for LLM encoder and decoder
- Proposed Herd algorithm for batch inference with dynamic pruning methods and implemented corresponding paper

Research Intern Shanghai AI Lab **Advisor** Bo Dai Sept 2022 - Dec 2022

- Joined the team of 3D mesh generation team with Diffusion Model.
- Mainly Responsible for the benchmark part of mesh generation and deploy related qualifying code for multiple mush generation algorithm

Research Intern Tsinghua University **Advisor** Jie Tang June 2022 - June 2023

- Joined the team of OAG-Benchmark and mainly responsible for the benchmark and tasks of Concept Taxonomy Completion.
- Implemented GLM-based Concept Taxonomy Retrieval System with OAG Database

Publications

Controllable Mesh Generation Through Sparse Latent Point Diffusion Models Feb 2023

Zhaoyang Lv etc. CVPR 2023

https://arxiv.org/abs/2303.07938

OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining Feb 2024

Fanjin Zhang etc. KDD 2024

https://arxiv.org/abs/2402.15810

IFMoE: An Inference Framework Design for Fine-grained MoE August 2024

Yuwei An etc. NeurIPS ML For Systems workshop

Herd: Contextual Grouping for Multi-Batch Inference with Dynamic Pruning October 2024

Yuwei An etc. NAACL 2025 In submission

Projects

LMCache github:LMCache

- Main Developer for LMCache Project.
- LMCache lets LLMs prefill each text only once. By storing the KV caches of all reusable texts, LMCache can reuse the KV caches of any reused text

Design of Multi-Tenant Hierarchical Embedding Parameter Server(THU Thesis) N/A

- Designed a multi-layer multi-tenant parameter server to handle the largest parameter volume and slowest running speed problem of the Embedding layer in recommendation systems
- The first version codebase for paper MaxEmbed: Maximizing SSD Bandwidth Utilization for Huge Embedding Models Serving

Abase: SQL-like database implemented in Python github:abase

- Developed a database with python which includes basic database function with File, Record, Index, Parser and System Management.
- Introduced some advanced function such as ACID and Query Integrity

Awards

Outstanding Graduate Student of Computer Science and Technology Department, Tsinghua University 2023

Tsinghua University Excellent Academic Scholarship 2022

Tsinghua University Excellent Academic Scholarship 2021

Tsinghua University Freshman Scholarship 2019

Technologies

Languages: Cpp, C, Java, Objective-C, SQL, Matlab, Python, Cuda, Typescript

Software: .NET, Microsoft SQL Server, Spring Boot, React