Ziyi Han

zyhan24@cse.cuhk.edu.hk +86-15927206410

Research Interests

Online learning, Edge intelligence, Optimization Algorithm Design

Education

The Chinese University of Hong Kong 08/2024 – present
Doctor of Philosophy Hong Kong, China
- Department of Computer Science and Engineering
- Supervisor: John C.S. Lui

Wuhan University 09/2021 – 06/2024
Master of Engineering Wuhan, China
- School of Cyber Science and Engineering
- Supervisor: Ruiting Zhou
- GPA: 3.57/4.0

Wuhan University 09/2017 – 06/2021
Bachelor of Engineering Wuhan, China
- School of Cyber Science and Engineering
- GPA: 3.81/4.0
- rank: 9/106

Publications

Published Papers

- **Ziyi Han**, Ruiting Zhou*, Chengzhong Xu, Yifan Zeng, and Renli Zhang InSS: An Intelligent Scheduling Orchestrator for Multi-GPU Inference with Spatio-Temporal Sharing, IEEE Transactions on Parallel and Distributed Systems (**TPDS**), 2024, to appear. (**CCF A**)
- Ruiting Zhou, **Ziyi Han***, Yifan Zeng, Zhi Zhou, Libing Wu, Wei Wang. SAFE: Intelligent Online Scheduling for Collaborative DNN Inference in Vehicular Network, in Proc. of **CSCWD** 2024, Tianjin, China, May 8-10, 2024. (**CCF C**)
- **Ziyi Han**, Ruiting Zhou*, Jinlong Pang, Haisheng Tan*, Yue Cao. Online Scheduling Unbiased Distributed Learning over Wireless Edge Networks, in Proc.of IEEE **ICPADS**, Beijing, China, Dec.13-15, 2021. (**CCF C**, **Outstanding Paper Award**)
- Yifan Zeng, Ruiting Zhou*, Lei Jiao, **Ziyi Han**, Jieling Yu. Efficient Online DNN Inference with Continuous Learning in Edge Computing, in Proc. of IEEE/ACM **IWQoS** 2024, Guangzhou, China, Jun. 19-21, 2024. (**CCF B**)
- Jinlong Pan, **Ziyi Han**, Ruiting Zhou*, Renli Zhang, John C.S. Lui and Hao Chen. Eris: An Online Auction for Scheduling Unbiased Distributed Learning over Edge Networks. IEEE Transactions on Mobile Computing (**TMC**), 23(6):7196-7209, June 2024. (**CCF A**)
- Jinlong Pang, **Ziyi Han**, Ruting Zhou*, Haisheng Tan, Yue Cao. Online Scheduling Algorithms for Unbiased Distributed Learning over Wireless Edge Networks. Journal of Systems Architecture (**JSA**), 131,1-14, Oct. 2022. (**CCF B**)

Papers in Submission/Revision

- **Ziyi Han**, Ruiting Zhou, Haisheng Tan, "Online Scheduling with Trajectory Prediction for Collaborative DNN Inference in Vehicular Networks", submitted to **INFOCOM**, 2025. (**CCFA**)

Honors and Scholarships

First Prize of Cybersecurity Excellence Scholarship	2021-2023
First Prize Scholarship of Academic Excellence	2022
Huawei Scholarship	2022
Excellent Bachelor Thesis	2021
Third Prize of University Scholarship	2020

Research Experiences

GPU Cluster Scheduling for DNN Inference Services

TPDS 2024 10/2022 - 05/2023

InSS is an intelligent scheduler of inference tasks on a GPU cluster with spatio-temporal sharing. It build a latency analytical model for each DNN batch, and make adaptive spatio-temporal and batch processing adjustments for online dynamic inference tasks in a very short time, to maximize cluster throughput while ensuring the SLO requirement.

- Design an interference-aware latency analytical model, which considers the heterogeneity in DNN models and GPU devices, along with the potential interference caused by co-located models.
- Propose a two-layer intelligent method for dynamic system configuration (including batch size, model placement, resource allocation) to maximize cluster throughput while ensuring the SLO requirement of inference tasks.
- Evaluation on a GPU cluster shows that InSS can improve the throughput by up to 189% while satisfying SLOs.

Collaborative DNN Inference in Vehicular Network

CSCWD 2024 04/2022 - 08/2022

SAFE is an intelligent scheduler for collaborative DNN inference acceleration in vehicular network. It employs DNN partitioning method to distribute partial workload between vehicles and RSUs, and carefully selects RSUs to minimize inference latency and ensure stable and reliable data transmission.

- Capture the characteristics of DNN models and the mobility of vehicles, we formulate the collaborative inference process in vehicular networks as an average latency minimization problem.
- Propose a two-layer SAC-D based collaborative inference framework, which combines the advantages of both deep reinforcement learning and optimization algorithm.
- Evaluation based on PyTorch shows that SAFE improves the success rate by up to 65% while reducing the average latency by up to 80%, compared to three benchmarks.

Resource Scheduling in Distributed Machine Learning

TMC 2023, JSA 2022, ICPADS 2021 10/2020 - 08/2021

This work focuses on resource scheduling for distributed ML jobs at wireless edge networks to minimize total cost without any future information.

- Model and formulate the online cost minimization (the sum of latency cost and bandwidth cost) into a mixed integer nonlinear program (MINLP).
- Design a heuristic algorithm for dynamically allocating resources for jobs to minimize total cost.
- Evaluation on a Kubernetes cluster shows the proposed algorithm achieves high resource efficiency, and outperforms baselines more than 55%.