Ziyang Zhang

· E-mail：zhangzy@stu.hit.edu.cn

EDUCATION

Ph.D., Computer Science and Technology Sep. 2020 - Sep. 2024
Harbin Institute of Technology Harbin, China
Advisor: Prof. Jie Liu (IEEE Fellow, ACM Distinguished Scientist)

M.Eng., Electronics and Communications Engineering Sep. 2018 - July. 2020
Nankai University (GPA 4.09/5.0) Tianjin, China
Advisor: Prof. Guiling Sun (Assistant dean of school of electronic information)

B.Eng., Electronic Information Science and Technology Sep. 2014 - July. 2018
Shandong University of Science and Technology (GPA 4.32/5.0) Qingdao, China

RESEARCH INTEREST

My main research interest is **edge computing**, **embedded intelligence**, with a focus on **high-performance and energy-efficent edge DNN inference**.

- **Edge Computing**
- **Machine Learning System**
- **Deep Reinforcement Learning**

Publications

Conference Papers

- POS: An Operator Scheduling Framework for Multi-model Inference on Edge Intelligent Computing
  Ziyang Zhang, Huan Li, Yang Zhao, Changyao Lin, Jie Liu
  ACM/IEEE IPSN, 2023 (*Acceptance Rate: 22/83, 26.5%*)
- Octopus: SLO-Aware Progressive Inference Serving via Deep Reinforcement Learning in Multi-Tenant Edge Cluster
  Ziyang Zhang, Yang Zhao, Jie Liu
  Springer ICSOC, 2023 (*Acceptance Rate: 35/208, 16.8%*)
- Choosing Appropriate AI-enabled Edge Devices, Not the Costly Ones
  Ziyang Zhang, Feng Li, Changyao Lin, Shihui Wen, Xiangyu Liu, Jie Liu
  IEEE ICPADS, 2021

Journal Papers

- DVFO: Learning-Based DVFS for Energy-Efficient Edge-Cloud Collaborative Inference
  Ziyang Zhang, Yang Zhao, Huan Li, Changyao Lin, Jie Liu
  IEEE Transactions on Mobile Computing, 9042-9059, 2024
- BCEdge: SLO-Aware DNN Inference Services with Adaptive Batch-Concurrent Scheduling on Edge Platforms
  Ziyang Zhang, Yang Zhao, Huan Li, Jie Liu
  IEEE Transactions on Network and Service Management, 4131-4145, 2024
- Multi-Sensor Data Fusion Algorithm Based on Trust Degree and ImprovedGenetics
  Ziyang Zhang, Guiling Sun, Bowen Zheng, Yangyang Li
  Sensors, 19(9), 1-18, 2019
- TOP: Task-Based Operator Parallelism for Asynchronous Deep Learning Inference on GPU
  Changyao Lin, Zhenmin Zhang, Ziyang Zhang, Jie Liu
  IEEE Transactions on Parallel and Distributed Systems, 1-16, 2024

Poster Papers

- DVFO: Dynamic Voltage, Frequency and Offloading for Efficient AI on Edge Devices
  Ziyang Zhang, Yang Zhao, Jie Liu
  ACM/IEEE IPSN, 2023

- E4: Energy-Efficient Early-Exit DNN Inference Framework for Edge Video Analytics
  Ziyang Zhang, Yang Zhao, Jie Liu
  ACM SenSys, 2023
- ECSRL: A Learning-Based Scheduling Framework for AI Workloads in Heterogeneous Edge-Cloud Systems
  Changyao Lin, Huan Li, Ziyang Zhang, Jie Liu
  ACM SenSys, 2021

Under Review

- E4: Energy-Efficient DNN Inference for Edge Video Analytics Via DVFS and Early Exiting
  Ziyang Zhang, Yang Zhao, Jie Liu
  AAAI, 2025
- E3: Early Exiting with Explainable AI for Real-Time and Accurate DNN Inference in Edge-Cloud Systems
  Changyao Lin, Zhenmin Zhang, Ziyang Zhang, Jie Liu
  ACM SenSys, 2025
- E3A: Energy-efficient Edge Analytics with Early Exit and Frequency Domain Distillation
  Shaowei He, Ziyang Zhang, Shusheng Li, Yang Zhao
  IEEE Transactions on Geoscience and Remote Sensing, 2024

Professional Service

- IEEE Transactions on Computers, reviewer
- IEEE Transactions on Parallel and Distributed Systems, reviewer
- IEEE Transactions on Vehicular Technology, reviewer
- IEEE Internet of Things Journal, reviewer
- Elsevier Internet of Things, reviewer
- ACM SenSys'24, AE Committee
- ACM SenSys/BuildSys Workshop on DATA’ 23, TPC member

Skill

- Programming Language: C, C++, Python, CUDA
- Embedded Platforms: Arm-Linux, GPU, FPGA
- Deep Learning Frameworks: TensorFlow, PyTorch, TensorRT, TVM, ONNX

Selected Awards and Honors

- **The First Prize**, MCM/ICM, 2017
- **National Scholarship**, Highest honor in China, 2019
- **Tencent Scholarship**, Tencent, 2024