NEIWEN LING

Department of Computer Science, Yale University
Address: Arthur K. Watson Hall, 51 Prospect Street, New Haven, CT 06511, U.S.
Email:neiwen.ling@yale.edu ⋄ Phone: (+1)(203)9965536 ⋄ Homepage:https://neawhen.github.io/neiwen.github.io/

RESEARCH INTERESTS

My primary research interests lie at the intersection of Edge Computing, Machine Learning, Cyber-Physical Systems (CPS), and Real-time Systems. I am committed to advancing **time-sensitive AI systems**, with a particular focus on designing AI (i.e, DL, FM, LLM) systems for **embodied agents** such as robots, assistive devices, and autonomous vehicles. These systems have wide-ranging applications, including **autonomous driving, embodied AI, and smart cities**. Additionally, I have extensive experience in developing real-world testbeds, including a campus-scale smart lamppost testbed.

- Systems for Large Language Model (LLM), Systems for Embodied AI, Time-sensitive LLMs in CPS/IoT
- On-device Deep Learning (DL), Time-sensitive DL for Autonomous Driving
- Distributed DL Systems, Cooperative Edge Computing

RESEARCH EXPERIENCE

Yale University 10/2023-Present
Postdoctoral Associate, Efficient Computing Lab, Department of Computer Science
Supervisor: Prof. Lin Zhong
Research Directions: Time-sensitive LLM Serving System, System for Embodied AI

The Chinese University of Hong Kong 08/2022-10/2023
Postdoctoral Fellow, AIoT lab, Department of Information Engineering
Research Directions: Distributed DL System, Foundation Model for IoT

Shenzhen Institute of Artificial Intelligence and Robotics for Society 11/2019-01/2020
Visiting Ph.D. student

EDUCATION

The Chinese University of Hong Kong 08/2018-07/2022
Ph.D., Information Engineering
Supervisor: Prof. Guoliang Xing
Research Directions: DL on CPU-GPU heterogeneous platforms, Edge Computing

Northwestern Polytechnical University 09/2014-07/2018
Bachelor's Degree, Electronics and Information Engineering

HONORS & AWARDS

- Best Artifact Award Runner-Up, ACM MobiCom 2024
- Best Paper Award Finalist, ACM SenSys 2022
- Best Poster Award, ACM SenSys 2022
- N2Women Young Researcher Fellowship, ACM SenSys 2021
- Postgraduate Scholarship, The Chinese University of Hong Kong 2018-2022
- Undergraduate Excellent Graduation Project, Northwestern Polytechnical University 2018

PUBLICATIONS

Conference Papers

- Neiwen Ling, Xuan Huang, Zhihe Zhao, Nan Guan, Zhenyu Yan, and Guoliang Xing, "BlastNet: Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference", The 20th Conference on Embedded Networked Sensor Systems, 15 pages double column
  ACM SenSys 2022, Best Paper Award Finalist
- Neiwen Ling, Kai Wang, Yuze He, Guoliang Xing, and Daqi Xie, "RT-mDL: Supporting real-time mixed deep learning tasks on edge platforms", The 19th Conference on Embedded Networked Sensor Systems, 14 pages double column
  ACM SenSys 2021

- Shuyao Shi*, Neiwen Ling*(*co-first authors), Zhehao Jiang*, Xuan Huang*, Yuze He, Xiaoguang Zhao, Bufang Yang, Chen Bian, Jingfei Xia, Zhenyu Yan, Raymond Yeung, and Guoliang Xing, "Soar: Design and Deployment of A Smart Roadside Infrastructure System for Autonomous Driving", The 30th Annual International Conference On Mobile Computing And Networking, 16 pages double column
  ACM MobiCom 2024, Best Artifact Award Runner-Up

- Zhehao Jiang*, Neiwen Ling*(*co-first authors), Xuan Huang, Shuyao Shi, Chenhao Wu, Xiaoguang Zhao, Zhenyu Yan, and Guoliang Xing, "CoEdge: A Cooperative Edge System for Distributed Real-Time Deep Learning Tasks", The 22nd ACM/IEEE Conference on Information Processing in Sensor Networks, 14 pages double column
  ACM/IEEE IPSN 2023

- Zhihe Zhao, Neiwen Ling, Nan Guan, and Guoliang Xing, "Miriam: Exploiting Elastic Kernels for Real-time Multi-DNN Inference on Edge GPU", the 21th Conference on Embedded Networked Sensor Systems, 14 pages double column
  ACM SenSys 2023

- Bufang Yang, Lixing He, Neiwen Ling, Zhenyu Yan, Guoliang Xing, Xian Shuai, Xiaozhe Ren and Xin Jiang, "EdgeFM: Leveraging Foundation Model for Open-set Learning on the Edge", the 21th Conference on Embedded Networked Sensor Systems, 14 pages double column
  ACM SenSys 2023

- Zhihe Zhao, Kai Wang, Neiwen Ling, and Guoliang Xing, "Edgeml: An automl framework for real-time deep learning on the edge", The 6th ACM/IEEE Conference on Internet of Things Design and Implementation, 12 pages double column
  ACM/IEEE IoTDI 2021

- Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang, "Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training", The 21st ACM International Conference on Mobile Systems, Applications, and Services, 14 pages double column
  ACM MobiSys 2023

- Wenjing Xie, Tao Hu, Neiwen Ling, Guoliang Xing, Chun Jason Xue, and Nan Guan, "Timely Fusion of Surround Radar/Lidar for Object Detection in Autonomous Driving Systems", the 30th IEEE International Conference on Embedded and Real-Time Computing Systems and Application, 6 pages double column
  IEEE RTCSA 2024

Selected Preprint

- Neiwen Ling, Guojun Chen, and Lin Zhong, "TimelyLLM: Segmented LLM Serving System for Time-Sensitive Robotic Applications", under review, 16 pages double column

- Guojun Chen, Xiaojing Yu, Neiwen Ling, and Lin Zhong, "ChatFly: Low-Latency Drone Control with Large Language Models", under review, 14 pages double column

Workshop and Poster/Demo Papers

- Neiwen Ling*, Yuze He*, Nan Guan, Heming Fu, and Guoliang Xing, "Dataset: An Indoor Smart Traffic Dataset and Data Collection System", The 5th International SenSys/BuildSys Workshop on Data
  ACM DATA 2022, SenSys/BuildSys 2022 Workshop

- Zhihe Zhao, Xian Shuai, Neiwen Ling, Nan Guan, Zhenyu Yan, and Guoliang Xing, "Moses: Exploiting Cross-device Transferable Features for On-device Tensor Program Optimization", The 24th International Workshop on Mobile Computing Systems and Applications 2023
  ACM HotMobile 2023

- Zhihe Zhao, Neiwen Ling, Nan Guan, and Guoliang Xing, "Aaron: Compile-time Kernel Adaptation for Multi-DNN Inference Acceleration on Edge GPU", The 20th Conference on Embedded Networked Sensor Systems
  ACM SenSys 2022 Poster, Best Poster Award

- Zhihe Zhao, Neiwen Ling, Kaiwei Liu, Nan Guan, and Guoliang Xing, "Unifying On-device Tensor Program Optimization through Large Foundation Model", The 21th Conference on Embedded Networked Sensor Systems, 2 pages double column
  ACM SenSys 2023 Poster

[1] * Co-first authors, Equal contribution

PROFESSIONAL SERVICES

- **Organizing Committee Member**
  - General Co-Chair, International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys'25), in Cyber-Physical Systems and Internet-of-Things Week 2025 (CPS-IoT Week 2025)
  - Technical Program Committee Co-Chair, International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys'24), in Cyber-Physical Systems and Internet-of-Things Week 2024 (CPS-IoT Week 2024)
  - Organizer, N2Women (Networking Networking Women) Meeting in the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys 2021)
- **Technical Program Committee Member**
  - The 23rd ACM Conference on Embedded Networked Sensor Systems (ACM SenSys 2025)
  - The 30th International Conference on Parallel and Distributed Systems (IEEE ICPADS 2024)
  - Artifact Evaluation, The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2024)
  - The 8th IEEE/ACM Conference on Connected Health: Applications, Systems, and Engineering Technologies (IEEE/ACM CHASE 2023)
  - Poster & Demo, The 8th ACM/IEEE International Conference on Internet of Things Design and Implementation 2023 (IEEE/ACM IoTDI 2023)
- **Invited Reviewer**
  - IEEE Transactions on Mobile Computing (TMC 2022, 2023, 2024)
  - ACM Transactions on Sensor Networks (TOSN 2023, 2024)
  - ACM Transactions on Internet of Things (TIOT 2023, 2024)
  - IEEE Transactions on Computers 2024
  - SCIENCE CHINA Information Sciences (SCIS 2023)
  - IEEE Network Magazine 2023
  - ACM Transactions on Computing for Healthcare (HEALTH 2023)
  - ACM International Conference on Pervasive and Ubiquitous Computing (IMWUT/UbiComp 2022)
  - IEEE International Conference on Computer Communications (INFOCOM 2024)
  - The First Workshop on DL-Hardware Co-Design for AI Acceleration in the 37th AAAI Conference on Artificial Intelligence (DCAA 2023)

TEACHING AND MENTORING

Mentorship

- Sebastian Orozco, UG Student at Yale (intern project, LLM-powered Virtual Robot Control) 05/2024-08/2024
- Claire Qu, UG Student at Yale (intern project, Robot Arm Control with LLM) 06/2024-08/2024
- Alondra Martinez Damazo, Yale Pathway (intern project, Time Model for Robot) 07/2024-08/2024
- Zhihe Zhao, Ph.D. Student at CUHK (major research projects, Compiling-level Real-time DL) 08/2021-10/2023
- Bufang Yang, Ph.D. Student at CUHK (research project, Foundation Model at Edge) 11/2022-07/2023

Teaching Assistant

- ENGG1100: Introduction to Engineering Design 2018 Fall
- IERG4230: Introduction to Internet of Things 2019 Spring
- ENGG1110: Problem Solving by Programming 2019/2020/2021 Fall
- IERG2602: Engineering Practicum 2020/2021 Spring

REFERENCES

Prof. Lin Zhong
Professor, ACM Fellow, IEEE Fellow, Yale University, United States
Prof. Guoliang Xing
Professor, IEEE Fellow, The Chinese University of Hong Kong, Hong Kong SAR
Prof. Xiaofan (Fred) Jiang
Associate Professor, Columbia University, United States
Prof. Rui Tan
Associate Professor, Nanyang Technological University, Singapore