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), Time-sensitive LLMs in CPS
- On-device Deep Learning (DL), Time-sensitive DL
- 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: On-device DL, Time-sensitive DL, Edge Computing, Autonomous Driving

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

SELECTED HONORS

- 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

SELECTED PUBLICATIONS

[1] **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

[2] **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

[3] Shuyao Shi*, **Neiwen Ling***, 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

[4] Zhehao Jiang*, **Neiwen Ling***, 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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, 7 pages double column
ACM HotMobile 2023

[11] Zhihe Zhao, **Neiwen Ling**, Nan Guan, and Guoliang Xing, "Poster Abstract: Aaron: Compile-time Kernel Adaptation for Multi-DNN Inference Acceleration on Edge GPU", The 20th Conference on Embedded Networked Sensor Systems, 2 pages double column
ACM SenSys 2022 Poster, Best Poster Award

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

PROFESSIONAL SKILLS

- **Programming Language**
  Python, C/C++ Language, Verilog HDL, Assembly Language
- **Development Experience on Hardware**
  NVIDIA Jetson TX2, NVIDIA AGX Xavier, NVIDIA Jetson Nano, FPGA, Arduino, 51 SCM, STM32F1, STM32F4
- **Development Experience on Software/Toolkit**
  Hugging Face Transformers, vLLM, PyTorch, LibTorch, Microsoft NNI, TensorRT, KubeEdge, ROS2

PROFESSIONAL SERVICES

- Technical Program Committee Chair, FMSys workshop in CPS-IoT Week 2024
- Technical Program Committee Member, SenSys 2025, ICPADS 2024, CHASE 2023
- Invited Reviewer, TMC 22/23/24, TOSN 23/24, TIOT 23/24,IMWUT/UbiComp22, INFOCOM24

[1] * Equal contribution