Kai Huang

CONTACT INFORMATION

3700 O'Hara St, Pittsburgh, PA 15213
Dept. of Electrical and Computer Engineering — Telephone: (412)277-5047
University of Pittsburgh — Email: k.huang@pitt.edu
Pittsburgh, PA, 15213 — Homepage: https://hellokevin07.github.io

RESEARCH INTERESTS

Efficient AI, AI for Systems

EDUCATION

University of Pittsburgh, Pittsburgh, PA — On-going
Ph.D. Candidate, Electrical and Computer Engineering, Intelligent Systems Lab @ Pitt
Advisor: Prof. Wei Gao

University of Science and Technology of China (USTC), Hefei, Anhui, — July 2019
B.E., Electronic Information Engineering

RESEARCH EXPERIENCE

Research Assistant — 2019-present
Dept. of Electrical and Computer Engineering, University of Pittsburgh

- Developed the first work that allows multimodal LLMs to elastically switch between input data modalities at runtime, for embodied AI applications such as autonomous navigation. Our implementations on NVidia Jetson AGX Orin demonstrate short modality adaptation delays of a few minutes with mainstream LLMs, 3.7x fine-tuning FLOPs reduction, and 4% accuracy improvements on multimodal QA tasks.
- Developed a green and sustainable training scheme for fine-tuning large language models, which can reduce 64% training FLOPs without noticeable accuracy loss.
- Developed a parameter-selective training scheme that can accelerate neural network training by up to 3.5x on weak embedded devices (e.g., Nvidia Jetson and Raspberry Pi).
- Developed and implemented an offloading scheme that allows extremely weak devices (e.g., MCUs with <1MB memory) to achieve real-time (<20ms) neural network inference. It is the first work that leverages Explainable AI to speed up neural network inference on weak devices.
- Developed and implemented a wireless backscatter system that leverages neural network inference to save its RF energy by up to 80%. The neural network is tailored based on the domain knowledge of backscatter communication, and hence is very lightweight and can be effectively trained even with a limited amount of data.

PUBLICATIONS

Conference Papers

* indicates equal contributions

1. [**arXiv**] Kai Huang, Boyuan Yang, Wei Gao. "Modality Plug-and-Play: Elastic Modality Adaptation in Multimodal LLMs for Embodied AI." arXiv preprint 2023.
2. [**ICLR'24**] Kai Huang, Hanyun Yin, Heng Huang, Wei Gao. "Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation." Twelfth International Conference on Learning Representations, 2024. *Acceptance Ratio*: 31%.
3. [**MobiSys'23**] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, Wei Gao. "PTEase: Objective Airway Examination for Pulmonary Telemedicine using Commodity Smartphones." In Proceedings of the 21st International Conference on Mobile Systems, Applications, and Services, pp. 110-123. 2023. *Acceptance Ratio*: 20.7%
4. [**MobiSys'23**] Kai Huang, Boyuan Yang, Wei Gao. "ElasticTrainer: Speeding Up On-Device Training with Runtime Elastic Tensor Selection." In Proceedings of the 21st International Conference on Mobile Systems, Applications, and Services, pp. 56-69. 2023. *Acceptance Ratio*: 20.7%

   Awarded ACM Artifact Available, Functional, Reusable, Results Replicated Badges (23.5%)

5. [**SenSys'22**] Chen Ruirong, Kai Huang, Wei Gao. "AiFi: AI-Enabled Interference Cancellation in WiFi Networks with Commodity PHY-Layer Information." In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, pp. 134-148. 2022. *Acceptance Ratio*: 24.8%
6. [**CML-IOT'22**] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, Wei Gao. "Out-Clinic Pulmonary Disease Evaluation via Acoustic Sensing and Multi-Task Learning on Commodity Smartphones." In The Fourth Workshop on Continual and Multimodal Learning for Internet of Things (**Best Paper Award**)
7. [**MobiCom'22**] Kai Huang, Wei Gao. "Real-time Neural Network Inference on Extremely Weak Devices: Agile Offloading with Explainable AI." In Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, pp. 200-213. 2022. *Acceptance Ratio*: 17.8%
8. [**IoTDI'22**] Kai Huang, Ruirong Chen, Wei Gao. "RAScatter: Achieving Energy-Efficient Backscatter Readers via AI-Assisted Power Adaptation." In 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation, pp. 1-13. IEEE, 2022. *Acceptance Ratio*: 33.3%
9. [**IPSN'22**] Xingzhe Song, Kai Huang, Wei Gao. "FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones." In 2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 145-157. IEEE, 2022. *Acceptance Ratio*: 30.2%
10. [**ASPLOS'22**] Boyuan Yang, Ruirong Chen, Kai Huang, Jun Yang, Wei Gao. "Eavesdropping user credentials via GPU side channels on smartphones." In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, pp. 285-299. 2022. *Acceptance Ratio*: 20.2%
11. [**MobiSys'20**] Yihao Liu*, Kai Huang*, Xingzhe Song, Boyuan Yang, Wei Gao. "MagHacker: eavesdropping on stylus pen writing via magnetic sensing from commodity mobile devices." In Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, pp. 148-160. 2020. *Acceptance Ratio*: 19.4%

Public Speaking

Presentations

1. "ElasticTrainer: Speeding Up On-Device Training with Runtime Elastic Tensor Selection." Elijah Group Meeting, Dept. of Computer Science, Carnegie Mellon University, October 2023
2. "AiFi: AI-Enabled WiFi Interference Cancellation with Commodity PHY-Layer Information." In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys), Boston, USA, Nov 2022.
3. "Real-time neural network inference on extremely weak devices: agile offloading with explainable AI." In Proceedings of the 28th Annual International Conference on Mobile Computing And Networking (MobiCom), InterContinental Sydney, Australia, Oct 2022
4. "RAScatter: Achieving Energy-Efficient Backscatter Readers via AI-Assisted Power Adaptation." In 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI), Virtual, May 2022
5. "Towards Real-time Neural Network Inference on Extremely Weak Devices", Elijah Group Meeting, Dept. of Computer Science, Carnegie Mellon University, November 2021
6. "Tailoring Neural Network Designs to Computing System Domains", Elijah Group Meeting, Dept. of Computer Science, Carnegie Mellon University, March 2021

Teaching and Mentoring Experience

Teaching:

- **Teaching Assistant**, ECE1175 - Embedded Systems Design — Spring 2021
  Dept. of Electrical and Computer Engineering, University of Pittsburgh
- **Teaching Assistant**, ECE1175 - Embedded Systems Design — Fall 2020
  Dept. of Electrical and Computer Engineering, University of Pittsburgh

- **Teaching Assistant**, ECE0202 - Embedded Processors and Interfacing — Spring 2020
  Dept. of Electrical and Computer Engineering, University of Pittsburgh

Professional Activities

Journal Reviewer

- IEEE Transactions on Mobile Computing
- Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)

Conference Reviewer

- IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS) 2022,
- IEEE Conference on Computer Communications (INFOCOM), 2022, 2023, 2024