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, On-Device AI, AI for Systems

EDUCATION

University of Pittsburgh, Pittsburgh, PA 2019/9-2024/5
Ph.D., Electrical and Computer Engineering
Dissertation: Bring Agile and Self-Evolvable Intelligence to Weak Embeded Devices
Advisor: Prof. Wei Gao
Intelligent Systems Lab: https://pittisl.github.io

University of Science and Technology of China (USTC), Hefei, Anhui, 2015/9-2019/7
B.E., Electronic Information Engineering

RESEARCH EXPERIENCE

Research Assistant 2019/9-2024/5
Dept. of Electrical and Computer Engineering, University of Pittsburgh

- **Mitigating Illegal Adaptation of Diffusion Models**: Developed a new scheme (FreezeAsGuard) for mitigating illegal adaptation of diffusion models (e.g., generating fake public figures' portraits) by freezing model tensors that are adaptation-critical only for illegal domain, while having the minimal impact on the adaptation in other legal domains.
- **Efficient Modality Adaptation of Large Multimodal Models**: Developed the first work (mPnP-LLM) that allows multimodal LLMs to elastically switch between input data modalities at runtime, for embodied AI applications (e.g., autonomous navigation). On Nvidia Jetson AGX Orin, it shows 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.
- **Efficient Fine-tuning of Large Language Models**: Developed an adaptive backpropagation scheme (GreenTrainer) for fine-tuning large language models, which can reduce up to 64% training FLOPs and 61% wall-clock time without noticeable accuracy loss.
- **Efficient Fine-tuning of Vision Models**: Developed a parameter-selective training scheme (ElasticTrainer) that can speed up fine-tuning vision models (e.g., CNN and ViT) by up to 3.5x on weak embedded devices (e.g., Nvidia Jetson TX2 and Raspberry Pi 3B).
- **Real-time Inference via Agile Offloading**: Developed an offloading scheme (AgileNN) that allows extremely weak devices (e.g., STM32F746 MCUs with <1MB memory) to achieve real-time (<20ms) neural network inference. It is the first work that leverages Explainable AI to optimize model partitioning and speed up neural network inference on weak devices.
- **Knowledge-Informed Modularization of AI Models**: Developed a wireless backscatter system (RAScatter) 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 with a limited amount of data.

PUBLICATIONS

Conference Papers

* indicates equal contributions

1. [**arXiv**] Kai Huang, Wei Gao. "FreezeAsGuard: Mitigating Illegal Adaptation of Diffusion Models via Selective Tensor Freezing." *arXiv preprint* arXiv:2405.17472.
2. [**arXiv**] Jifeng Song, Kai Huang, Xiangyu Yin, Boyuan Yang, Wei Gao. "Achieving Sparse Activation in Small Language Models." *arXiv preprint* arXiv:2406.06562.
3. [**arXiv**] Kai Huang, Boyuan Yang, Wei Gao. "Modality Plug-and-Play: Elastic Modality Adaptation in Multimodal LLMs for Embodied AI." *arXiv preprint* arXiv:2312.07886.

4. [**MobiCom'24**] Kai Huang, Xiangyu Yin, Tao Gu, Wei Gao. "Perceptual-Centric Image Super-Resolution using Heterogeneous Processors on Mobile Devices." *In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2024. *Acceptance Rate*: 20.9%
5. [**ICLR'24**] Kai Huang, Hanyun Yin, Heng Huang, Wei Gao. "Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation." *The 12th International Conference on Learning Representations (ICLR)*, 2024. *Acceptance Rate*: 30.8%
6. [**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 (ACM MobiSys)*, pp. 110-123. 2023. *Acceptance Rate*: 20.7%
7. [**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 (ACM MobiSys)*, pp. 56-69. 2023. *Acceptance Rate*: 20.7%

   Received ACM Artifact Available, Functional, Reusable, Results Replicated Badges (23.5%)
8. [**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 (ACM SenSys)*, pp. 134-148. 2022. *Acceptance Rate*: 24.8%
9. [**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 (CML-IOT)* (**Best Paper Award**)
10. [**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 (ACM MobiCom)*, pp. 200-213. 2022. *Acceptance Rate*: 17.8%
11. [**IoTDI'22**] Kai Huang, Ruirong Chen, Wei Gao. "RAScatter: Achieving Energy-Efficient Backscatter Readers via AI-Assisted Power Adaptation." *The 7th ACM/IEEE International Conference on Internet-of-Things Design and Implementation (ACM/IEEE IoTDI)*, pp. 1-13. IEEE, 2022. *Acceptance Rate*: 33.3%
12. [**IPSN'22**] Xingzhe Song, Kai Huang, Wei Gao. "FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones." *The 21st ACM/IEEE Conference on Information Processing in Sensor Networks (ACM/IEEE IPSN)*, pp. 145-157. IEEE, 2022. *Acceptance Rate*: 30.2%
13. [**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 (ACM ASPLOS)*, pp. 285-299. 2022. *Acceptance Rate*: 20.2%
14. [**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 (ACM MobiSys)*, pp. 148-160. 2020. *Acceptance Rate*: 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 (ACM 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 (ACM MobiCom), InterContinental Sydney, Australia, Oct 2022
4. "RAScatter: Achieving Energy-Efficient Backscatter Readers via AI-Assisted Power Adaptation." The 7th ACM/IEEE International Conference on Internet-of-Things Design and Implementation (ACM/IEEE 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

Miscellaneous

- Created a subjective learning guide for Generative AI research:
  `https://github.com/pittisl/Generative-AI-Tutorial`