XIJIE HUANG

Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR
E-mail: huangxijie1108@gmail.com | Homepage: https://huangowen.github.io/ | Google Scholar: nFW2mqwAAAAJ&hl

EDUCATION

Hong Kong University of Science and Technology — Hong Kong SAR
Ph.D. in Computer Science Engineering, HKUST Vision and System Design Lab — *Sept 2020 - Present*

- Advisor: Prof. Tim Kwang-Ting CHENG
- HKUST Postgraduate Studentship and RedBird Scholarship

Shanghai Jiao Tong University — Shanghai, China
B.E. in School of Electronics Information and Electrical Engineering — *Sept 2016 - June 2020*

- Overall GPA: 89.4/100 (91.3/100 for junior year) Ranking:2/55
- Advisor: Prof. Cewu Lu, Machine Vision and Intelligence Group, SJTU

University of California, Los Angeles — Los Angeles, USA
Visiting Research Student — *June 2019 - Sept 2019*

- Research intern to UCLA ECE department (Cross-disciplinary Scholars in Science & Technology Program)
- Advisors: Prof. Mani B. Srivastava, Department of Electrical Computer Engineering, UCLA

RESEARCH INTERESTS

My research interests lie in the general area of model compression, particularly in its applications in efficient human-centric computer vision and Large Language Models (LLMs). More concretely, My research interests focus on quantization, pruning, algorithm-hardware co-design, human-object interaction (HOI) recognition, and scene understanding.

RESEARCH/PROJECT EXPERIENCE

Hardware-Software Co-design of Model Compression, HKUST

· Proposed Stochastic Differentiable Quantization (SDQ), which is an efficient and effective mixed-precision quantization technique outperforming full-precision ResNet and MobileNet on ImageNet with an average bitwidth lower than 4.
· Proposed an efficient variation-aware vision transformer (ViT) quantization framework. It is the first work to analyze and locate the variation in ViT quantization. Our solution to variation in ViTs leads to state-of-the-art accuracy on the ImageNet-1K dataset across different ViT models (DeiT, Swin, SReT).
· Propose a new angle through the coreset selection to improve the training efficiency of quantization-aware training. Our method can achieve 8.39% of 4-bit quantized ResNet-18 on the ImageNet-1K dataset with only a 10% subset.
· As a member of AI Chip Center for Emerging Smart Systems (ACCESS), building and evaluating hardware-friendly model compression techniques on AI chips. Our tiny accelerator with a customized data fetch hardware architecture can achieve 1.40 to 2.98 greater DSP efficiency and offers 1.91 greater energy efficiency compared to the SOTA accelerators.

Reinforced Prompt Pruning for In-context Learning, Microsoft Research Asia (MSRA)

· Researching the efficiency of in-context learning of Large Language Models (LLMs).
· Proposed a coarse-to-fine context pruner based on reinforcement learning to select examples from prompt candidate sets and select informative tokens from each example. Our method achieve state-of-the-art performance on reasoning tasks (GSM8K, MultiArith, AddSub, SingleEq, SVAMP) across various LLMs (LLaMa-2-7B, LLaMa-2-13B,LLaMa-2-70B)

Automated Vision-Based Wellness Analysis for Elderly Care Centers, HKUST

· Building a vision-based elderly care system that can provide immediate assistance and useful insights for caretakers. Collaborating with the Heaven of Hope care center and has built a healthcare dataset based on video recording.
· Designing human-centered scene understanding model, achieving state-of-the-art accuracy on scene graph generation (SSG) task and human-object interaction (HOI) recognition task.

Machine Vision and Intelligence Group, Department of Computer Science, SJTU

· Proposed Transferable Interactiveness Network to tackle the imbalanced distribution in human action recognition problems, especially human-object interaction detection problems

- Built the state-of-the-art knowledge base and engine of human activity understanding **HAKE**. HAKE provides elaborate and abundant with 7 M+ fine-grained part level annotations in a large scale of images and videos. In supervised, few-shot and transfer learning, our approach achieves significant improvements on large-scale activity benchmarks

Networked & Embedded Systems Laboratory, Department of Electrical Computer Engineering, UCLA

- Proposed a Trojan backdoors detection framework called **NeuronInspect**, using visual interpretability technique to effectively detect Trojan backdoors in deep neural networks without restoring the trigger and any backdoor samples
- Evaluate **NeuronInspect** on different attack scenarios and prove better robustness and effectiveness over previous state-of-the-art trojan backdoor detection techniques by a great margin

PUBLICATIONS & PRE-PRINT

Efficient Quantization-aware Training with Adaptive Coreset Selection
Xijie Huang, Zechun Liu, Shih-Yang Liu, Kwang-Ting Cheng
Preprint

Efficient Variation-aware Vision Transformer Quantization
Xijie Huang, Zhiqiang Shen, Kwang-Ting Cheng
Preprint

LLM-FP4: 4-Bit Floating-Point Quantized Transformers
Shih-Yang Liu, Zechun Liu, **Xijie Huang**, Pingcheng Dong, Kwang-Ting Cheng
Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023 (Acceptance Rate: 21.3%)

SDQ: Stochastic Differentiable Quantization with Mixed Precision
Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Xianghong Hu, Jeffry Wicaksana, Eric Xing, Kwang-Ting Cheng
International Conference on Machine Learning (ICML) 2022 (Acceptance Rate: 21.9%)

Automated Vision-Based Wellness Analysis for Elderly Care Centers
Xijie Huang, Jeffry Wicaksana, Shichao Li, Kwang-Ting Cheng
AAAI Conference on Artificial Intelligence (AAAI) W3PHIAI 2022

FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation
Jeffry Wicaksana, Zengqiang Yan, Dong Zhang, **Xijie Huang**, Huimin Wu, Xing Yang, Kwang-Ting Cheng
IEEE Transactions on Medical Imaging (TMI) 2022

A Tiny Accelerator for Mixed-bit Sparse CNN based on Efficient Fetch Method of SIMO SPad
Xianghong Hu, Xuejiao Liu, Yu Liu, **Xijie Huang**, Xihao Guan, Luhong Liang, Chi-Ying Tsui, Kwang-Ting Cheng
IEEE Transactions on Circuits and Systems (TCAS) 2022

Transferable Interactiveness Knowledge for Human-Object Interaction Detection
Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, **Xijie Huang**, Liang Xu, Cewu Lu
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2021

Latent Fingerprint Image Enhancement based on progressive generative adversarial network
Xijie Huang, Peng Qian, Manhua Liu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 Biometric Workshop

PaStaNet: Toward Human Activity Knowledge Engine
Yong-Lu Li, Liang Xu, Xinpeng Liu, **Xijie Huang**, Shiyi Wang, Hao-Shu Fang, Cewu Lu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 (Acceptance Rate: 22.09%)

Transferable Interactiveness Knowledge for Human-Object Interaction Detection
Yong-Lu Li, Siyuan Zhou, **Xijie Huang**, Liang Xu, Ze Ma, Hao-shu Fang, Yanfeng Wang, Cewu Lu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019 (Acceptance Rate: 25.15%)

NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations
Xijie Huang, Moustafa Alzantot, Mani.B.Srivastava
Preprint 2020

HAKE: Human Activity Knowledge Engine
Yong-Lu Li, Liang Xu, Xinpeng Liu, **Xijie Huang**, Ze Ma, Hao-Shu Fang, Cewu Lu
Preprint 2020

SELECTED ACADEMIC ACHIEVEMENTS

National Scholarship (Top 2% students in Shanghai Jiao Tong University)	2017
A Class Scholarship (Top 2% students in Shanghai Jiao Tong University)	2017
Second Prize in China Undergraduate Mathematical Contest in Modeling, Shanghai Division.	2017
Endress+Hauser Scholarship, Endress+Hauser Inc.	2018
Meritorious Winner in MCM & ICM, Comap.	2018
CSST Scholarship (USD $5,343) University of California, Los Angeles	2019
Best Presentation Award (Among 90 research interns at UCLA)	2019
RongChang Academic Scholarship (The highest honor in SJTU, Top 20 of 16000 students)	2019
A Class Oversea Research Fellowship	2019
8th place in ICCV 2019 Person In Context Human-Object Interaction Challenge	2019
RedBird Scholarship (HKD $40000) Hong Kong University of Science and Technology	2020-2023
Postgraduate Studentship, Hong Kong University of Science and Technology	2020-2023
AAAI-22 Student Scholarship	2022

SERVICES AND EXPERIENCES

Research Internship

- Ph.D. Research Intern in Microsoft Research Asia (MSRA) May 2023 - Feb 2024

Open Source Project

- Awesome LLM Compression (383 stars, 12/11/2023)

Reviewer

- Conference: ICLR 2024, EMNLP 2023, NeurIPS 2023, ICCV 2023, CVPR 2023, AAAI 2022-2023, WACV 2022-2023, ICML 2022 (Top 10% Reviewer), ECCV 2022

Teaching Assistant

- COMP 2211 (Exploring Artificial Intelligence), Fall 2022, Lecture: Professor Desmond Tsoi
- COMP 5421 (Computer Vision), Spring 2021, Lecture: Professor Dan Xu
- COMP 1021 (Introduction to Computer Science), Fall 2021, Lecturer: Professor David Rossitor

Volunteer

- Shanghai International Marathon (2016)
- China-Korea Symposium on Artificial Intelligence and Brain Science (2019)

COMPUTER AND LANGUAGE SKILLS

Natural Languages	Chinese (native), English (fluent)
Programming Languages	Python, MATLAB, C, C++
Deep Learning Framework	PyTorch, TensorFlow, Keras
Miscellaneous Skills	LaTeX, Altium Designer, Proteus, LabVIEW

STANDARD TEST

TOEFL	105 (Reading:28 Listening:30 Speaking:24 Writing:23)
GRE	322 (Q170+V152) + 3.5(AW)