Yiren (Aaron) Zhao

yaz21@cam.ac.uk

UK Contact: (0044) 07547842218

Experience

Assistant Professor, Imperial College London *2022-Now*
Lead a lab with 20+ people, secured more than 5M funding from both industry and government in two years
Research Fellow, University of Cambridge *2021-2022*
Research Scientist (Part-time), Apple AI and ML Team (Cupertino) *Dec 2021-June 2022*
Research Scientist (Part-time), Microsoft Research New England (Boston) *June 2019-Feb 2020*
Microsoft Research New England, Research Intern *June 2019-Oct 2019*
Microsoft Research Redmond, Research Intern *June 2018-Oct 2018*
Microsoft Research Cambridge, Research Intern *June 2017-Oct 2017*

Education

PhD in Computer Science, University of Cambridge *Grad. 2021*
Mphil in Advanced Computer Science, University of Cambridge *Grad. 2017*
BEng in Electrical & Electronic Engineering, Imperial College London *Grad. 2016*

Selected Awards and Honors

Accelerating Foundation Models Research Award, Microsoft *2023*
Best Paper Award, The 2023 ICML Workshop on Computational Biology, ICML-WCB'23 *2023*
Toby Jackman Prize, St Edmund's College, University of Cambridge *2022*
Junior Research Fellowship at St John's College, University of Cambridge *2021*
Best Paper Award, DLG-AAAI'21 *2021*
Apple Scholar in AI and ML, 12 in total annually worldwide *2020*
EPSRC International Doctoral Studentship joint Qualcomm Premium Scholarship *2017*
Willis Jackson Medal and Prize *2016*

Selected Papers

Summary: I have published over 70 peer-reviewed papers (the list only contains papers in the past 3 years, full list available on Google Scholar) and my group maintains a consistent appearance on top ML conferences (eg. ICML, NeurIPS, CVPR, EMNLP, ACL etc.).

Refining Salience-Aware Sparse Fine-Tuning Strategies for Language Models
X Liu, A Thomas, C Zhang, J Cheng, Y Zhao, X Gao;
The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)

Hardware and Software Platform Inference
C Zhang, H Foerster, R Mullins, Y Zhao, I Shumailov;
2025 International Conference on Machine Learning (ICML 2025)

Cached Multi-Lora Composition for Multi-Concept Image Generation
X Zou, M Shen, C Bouganis, Y Zhao;
2025 International Conference on Learning Representations (ICLR 2025)

QERA: an Analytical Framework for Quantization Error Reconstruction
C Zhang, J TH Wong, C Xiao, G Constantinides, Y Zhao;
2025 International Conference on Learning Representations (ICLR 2025)

Architectural Neural Backdoors from First Principles
H Langford, I Shumailov, Y Zhao, R Mullins, N Papernot
IEEE Symposium on Security and Privacy 2025 (S&P 2025)

GV-Rep: A Large-Scale Dataset for Genetic Variant Representation Learning
Z Li, V Subasi, G Stan, Y Zhao, B Wang
The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2024, Datasets and Benchmarks Track)

Absorb & Escape: Overcoming Single Model Limitations in Generating Heterogeneous Genomic Sequences
Z Li, Y Ni, G Xia, W Beardall, A Das, G Stan, Y Zhao

The Thirty-eight Conference on Neural Information Processing Systems (NeurIPS 2024)

AI Models Collapse when Trained on Recursively Generated Data
I Shumailov, Z Shumaylov, Y Zhao, Y Gal, N Papernot, R Anderson
Nature 2024 (Front Cover)

Enhancing Node Representations for Real-World Complex Networks with Topological Augmentation
X Zhao, Z Li, M Shen, G Stan, P Lio, Y Zhao
European Conference on Artificial Intelligence (ECAI 2024)

HASS: Hardware-Aware Sparsity Search for Dataflow DNN Accelerators
Z Yu, S Sreeram, K Agrawal, J Wu, A Montgomerie-Corcoran, C Zhang, J Cheng, C Bouganis, Y Zhao
The International Conference on Field-Programmable Logic and Applications (FPL 2024)

LQER: Low-Rank Quantization Error Reconstruction for LLMs
C Zhang, J Cheng, G Constantinides, Y Zhao
International Conference on Machine Learning (ICML 2024)

ImpNet: Imperceptible and Blackbox-undetectable Backdoors in Compiled Neural Networks
E Clifford, I Shumailov, Y Zhao, R Anderson, R Mullins
2nd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2024)

MiliPoint: A Point Cloud Dataset for mmWave Radar
H Cui, S Zhong, J Wu, Z Shen, N Dahnoun, Y Zhao
Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)

Revisiting Block-based Quantisation: What is Important for Sub-8-bit LLM Inference?
C Zhang, J Cheng, I Shumailov, GA Constantinides, Y Zhao
Empirical Methods in Natural Language Processing (EMNLP 2023)

Dynamic Stashing Quantization for Efficient Transformer Training
G Yang, D Lo, R Mullins, Y Zhao
Empirical Methods in Natural Language Processing (EMNLP 2023)

Revisiting Automated Prompting: Are We Actually Doing Better?
Y Zhou, Y Zhao, I Shumailov, R Mullins, Y Gal
61st Association for Computational Linguistics (ACL 2023)

Architectural backdoors in neural networks
M Bober-Irizar, I Shumailov, Y Zhao, R Mullins, N Papernot
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR 2023)

Adaptive Channel Sparsity for Federated Learning Under System Heterogeneity
D Liao, X Gao, Y Zhao, CZ Xu
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR 2023)

DAdaQuant: Doubly-adaptive quantization for communication-efficient Federated Learning
R Honig, Y Zhao and R Mullins
International Conference on Machine Learning 2022 (ICML 2022)

Rapid Model Architecture Adaption for Meta-Learning
Y Zhao, X Gao, I Shumailov, N Fusi and R Mullins
Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)

Funding

Tracking Evovling AI Compute: Uniting Models, Algorithms, and System Innovators with a Top-Down Evolutionary Bencharmak
E Ponti, N Brown, A Jackson, B Grot, W Li, L Mai, Y Zhao
ARIA (Scaling Compute), 2M GBP, 2025-2027

AIxSIM: A Scalable and Modular Simulation Framework for AI Accelerator Systems
Y Zhao (PI), M Luo, R Mullins, T Jones, G Constantinides, W Luk, M O'Boyle
ARIA (Scaling Compute), 4.9M GBP, 2024-2027

AMD Research Funding Donation for Developing AMD AI Engine Compilation
Y Zhao, G Constantinides
AMD Donation, 80K GPB cash donation + 30K equipment, 2023

Graph AI for Science: Helping LLMs to Understand Complex Data in Genomics
Y Zhao

Accelerating Foundation Models Research Award, Microsoft, 100K GBP Compute Credit, 2023

UKRI Access to HPC

Y Zhao

Total 100K GPU hour on EPSRC tier 2 resources, 2022, 2023 and 2024

Patents

Neural Network Activation Compression with Narrow Block Floating-point

D Lo, A Phanishayee, E S Chung, Y Zhao and R Zhao; US patent, US20200210838, 2020

Neural Network Activation Compression with Non-uniform Mantissas

D Lo, A Phanishayee, E S Chung, Y Zhao US patent, US20200242474, 2020

Neural network Activation Compression with Outlier Block Floating-point

D Lo, A Phanishayee, E S Chung, Y Zhao and R Zhao US patent, US20200210839, 2020

Adjusting Acitvation Compression for Neural Network Training

D Lo, B D Rouhani, E S Chung, Y Zhao, A Phanishayee and R Zhao US patent, US20200264876, 2020

Affliations

Imperial Artificial Intelligence Network *2024-Now*

UKRI Artificial Intelligence for Engineering Biology Consortium *2022-Now*

CaRAML - Cambridge Research and Applications in Machine Learning *2021-Now*

Academic Services

Technical Program Committee Chair for EuroMLSys (2024)

Area Chair for NeurIPS and ICML (2023 - Now)

Program Committee for AAAI (2024), and EuroMLSys(2022 - Now)

Reviewer for ICLR, ACL, EMNLP and NAACL (Since 2018)

PhD Supervision

Summary: Currently supervising 14 PhD students (both primary and secondary) in the areas of Efficient AI, AI Safety, and AI Hardware Acceleration. I have also demonstrated EDI commitment by supervising students from under-represented groups (eg. LGBTQ).

Primary Supervisor

Cheng Zhang (2022-Now, post-ESA, co-supervised with Prof. George Constantinides)
Victor Zhao (2022-Now, post-ESA, co-supervised with Prof. Pietro Lio)
Mingzhu Shen (2023-Now, post-ESA, co-supervised with Prof. Christos Bouganis)
Pedro Gimense (2023-Now, with Prof. Robert Mullins)
Eleanor Clifford (2023-Now, with Prof. Robert Mullins)
Can Xiao (2023-Now, with Dr. Jianyi Cheng)
Jeffrey Tsz Hang Wong (2024-Now, with Prof. Wayne Luk)
Przemyslaw Forys(2025-Now, with Prof. Wayne Luk)
Tony Liu (2025-Now)

Secondary Supervisor

Zehui Li (2022-Now, post-ESA, co-supervised with Prof. Guy-bart Stan)
Timon Schneider (2023-Now, co-supervised with Prof. Guy-bart Stan and Prof. Tom Ellis)
Hanna Foerster (2024-Now, co-supervised with Prof. Robert Mullins)
Ying Yu (2024-Now, co-supervised with Dr. Fei Teng)
Keran Zheng (2024-Now, co-supervised with Prof. Christos Bouganis)

Teaching and Administration

MSc in Advanced and Digital IC Design Course Director (Shadowing) *2024-Now*

Organize the MSc program, including course design, teaching structuring and admissions.

Information Processing (ELEC70009) *2023-Now*

Deliver both lectures and labs for the module.

Advanced Deep Learning Systems (ELEC70109/EE9-AML3-10/EE9-AO25) *2023-Now*

Deliver both lectures and labs for this new module. Designed the entire course from scratch and now delivering it to MEng, MSc in AML and MSc in ADIC

ADIC Design Labs (ELEC70093) *2024-Now*

Help design and organize the labs.

Projects

Summary: I love and have written a lot of code. This is a small list of projects that I am proud of. Some of them are not research related, but I believe they are fun to mention.

Project MASE

MASE is a tool-chain that wraps given Pytorch Models and provide a series of modular torch FX Graph based manipulations to lower these models into either a hardware description language or a more efficient model through software optimizations.

Making Second-hand GPUs Shine

The project has not been open-sourced yet, but we have successfully booted LLaMA 400B models on our custom data center with second-hand V100s with our custom inference engine. The inference engine is coded in Rust from ground up and is outperforming SOTA inference engines such as vLLM.

Project Mayo

We built a wrapper for the entire TensorFlow (think about it as Keras), we make training, evaluation and hyper-parameter tuning all controllable through YAML files. We later archived this project when Tensorflow 2.0 was released, because all interfaces are changed.

Media Coverage and external recognition

Model Collapse

Our work on AI models collapse when trained on recursively generated data was covered by a great number of media outlets, including MIT Technology Review, The Atlantic, IEEE Spectrum, BBC Science, WSJ New York Times, Financial Times, and so on. The paper is featured on Nature's front cover. This article is in the 99th percentile (ranked 22nd) of the 302,761 tracked articles of a similar age in all journals and the 99th percentile (ranked 3rd) of the 1,044 tracked articles of a similar age in Nature.

Sponge Example as DDoS Attack on Neural Networks

We have identified the existence of sponge examples and validated this on popular machine learning (ML) services like Microsoft Azure Translation. Upon informing Microsoft about this vulnerability, the company promptly fixed it. This attack method was included in the renowned MITRE Attack Framework and garnered coverage from The Register and recognition in the Cybersecurity of AI and Standardisation report by ENISA (The European Union Agency for Cybersecurity)