Zehuan Zhang

Research Group: cc.doc.ic.ac.uk
Website: zehuanzhang.github.io
Email: zehuan.zhang22@imperial.ac.uk

Department of Computing
Imperial College London
South Kensington Campus

EDUCATION BACKGROUND

Imperial College London, London, UK 04/2023 –on
Ph.D., Computing Research | Supervisor : Prof. Wayne Luk
- 4-year PhD Scholarship

Tianjin University, Tianjin, China 08/2019 – 03/2022
M.S. of Engineering in Microelectronics and Solid Electronics | Supervisor : Prof. Qiang Liu
- Thesis: Design and Hardware Implementation of Lightweight Zero-Shot Learning Based on Attribute
- GPA: 90.38/100 (Ranking: Top 1)
- 2-year Graduate Research Assistantship

Tianjin University, Tianjin, China 09/2015 – 07/2019
B.E. in Integrated Circuit Design and Integrated System
- GPA: 3.85/4.0 (Ranking: Top 1)

RESEARCH INTERESTS

- Hardware System for ML:
  - Accelerator for Trustworthy AI
  - Reconfigurable Hardware for Complex-Valued Neural Network
  - Heterogeneous Acceleration for LLM
- ML for Hardware System
  - Co-Design for Reconfigurable Accelerator
  - LLM-based Code Generation and Optimization

RESEARCH EXPERIENCES

Research on Heterogeneous Architecture for Reward-Guided Speculative Reasoning 01/2025 – now
- (System) Propose a novel heterogeneous architecture featuring FPGA-based drafting and GPU-based verification for reward-guided speculative reasoning, leveraging the reconfigurability and low-power characteristics of FPGAs alongside the high computational capacity of GPUs.
- (Algorithmic) Integrate a backtracking mechanism into reward-guided speculative reasoning, enabling the revisitation of earlier predictions and the refinement of step-wise results to enhance accuracy.
- (Hardware) Map a small-scale draft model onto the AMD V80 FPGA via high-level synthesis (HLS) design. The developed accelerator accommodates both accepted and rejected draft candidates; for rejected cases, the hardware design employs a progressive processing scheme, thereby mitigating extensive redundant computations.

Research on Algorithm and Hardware Design for Bayesian Complex-Valued Neural Network 09/2024 – 11/2024
- Propose a novel Dropout-based BayesCVNN that enables reliable uncertainty prediction for complex-valued applications with theory guarantees. This complex-valued dropout module can be used as a plug-and-play method to enhance uncertainty estimation in any CVNNs.
- Design an automated search approach to effectively find the dropout configurations for real and imaginary parts, considering both algorithmic targets and hardware constraints. The approach is capable of identifying the optimal configuration within an exponentially larger design space than conventional BayesNNs.
- Develop a framework to generate customized accelerators for BayesCVNNs via a series of building blocks, achieving high hardware performance.
- The generated accelerator achieves 23.5x higher energy efficiency and 3x faster speed than GPU implementations.

Research on Acceleration of Pruned Complex-Valued Neural Networks 12/2023 – 08/2024
- Demonstrated both the real and imaginary parts can be pruned in complex-valued neural networks, which may lead to performance variability with different pruning rates for each part.
- Proposed a novel progress heterogeneous pruning technique for real and imaginary parts of weights in CVNNs, significantly improving the overall sparsity ratio while maintaining the algorithmic performance.
- Experiments to show that the heterogenous pruning achieves a 51.23% reduction in FLOPs with only a 0.63% accuracy loss. The adaptable accelerator achieves 6.7x and 12.2x speedup over GPU and CPU implementations, demonstrating significant improvements in processing speed and energy efficiency.

Research on Automatic Search for Dropout-Based Bayesian Neural Network 04/2023 – 11/2023

- Proposed a novel neural dropout search framework with one-shot supernet training and an evolutionary algorithm to automatically optimize both dropout-based Bayesian neural networks and the associated FPGA-based accelerators given the target applications and constraints.
- Designed FPGA-based implementations of four types of dropout methods, enabling the acceleration of dropout-based Bayesian neural networks with different dropout combinations.
- Produced FPGA designs achieving up to 65× and 33× higher energy efficiency over CPU and GPU implementations, despite the more advanced technology (DAC 2024).

Research on MRI Analysis with Reliable Uncertainty Estimation 09/2022 – 03/2023

- Proposed an algorithm–hardware co-optimization flow that converts a deep neural network to a hardware-efficient mask-based Bayesian neural network.
- Applied the design flow to the medical model (IVIM-NET) to produce uIVIM-NET providing uncertainty information for MRI analysis.
- Developed a novel customized FPGA-based accelerator for the uIVIM-NET with mask-zero skipping strategy and batch-level scheme optimizations to enhance performance and reduce power consumption.
- Demonstrated the co-design approach can satisfy the uncertainty requirements of MRI analysis, while achieving 7.5× and 32.5× speedup on an Xilinx VU13P FPGA compared to GPU and CPU implementations with 34.4× and 82.8× reduced power consumption (ASAP 2024).

SELECTED PUBLICATIONS

First-author

- Hardware-Aware Neural Dropout Search for Reliable Uncertainty Prediction on FPGA
  Zehuan Zhang, Hongxiang Fan, Hao Mark Chen, Lukasz Dudziak, Wayne Luk.. ACM/IEEE Design Automation Conference (**DAC**).
- Accelerating MRI Uncertainty Estimation with Mask-based Bayesian Neural Network
  Zehuan Zhang, Matej Genci, Hongxiang Fan, Andreas Wetscherek, Wayne Luk.. IEEE International Conference on Application-specific Systems, Architectures and Processors (**ASAP**).
- Harnessing Heterogeneous Sparsity and Adaptable Acceleration for Complex-Valued Neural Networks
  Zehuan Zhang, Zhengyan Liu, Qiang Liu, Wayne Luk.. (Under Review)
- Extending Dropout-based Bayesian Approximation to Complex Domains: Mapping Bayesian Complex-Valued Neural Networks on FPGA
  Zehuan Zhang, Hao Mark Chen, He Li, Wayne Luk.. (Under Review)

Co-authored

- Advancing AI-assisted Hardware Design with Hierarchical Decentralized Training and Personalized Inference-Time Optimization
  Hao Mark Chen, **Zehuan Zhang**, Wanru Zhao, Nicholas Lane, Hongxiang Fan.. (Under Review)
- CODESCA: Co-Design for Spectral Clustering Acceleration
  Zhengyan Liu, Ce Guo, **Zehuan Zhang**, Wayne Luk.. (Under Review)

AWARDS AND HONORS

Young Fellow @ DAC Conference, 2024, USA 2024
4-year Ph.D Scholarship, China Scholarship Council (CSC) , China 2023 – 2027
Outstanding Graduates of Tianjin University, Tianjin University, China 2018 – 2019
18th "Student Science and Technology Talent" of Tianjin University, Tianjin University, China 2018 – 2019
"People's Government" Scholarship, Tianjin Education Commission, China 2016 – 2017
Advanced Individuals in Scientific and Technological Innovation, Tianjin University, China 2016 – 2017
Merit Student, Tianjin University, China 2015 – 2016, 2016 – 2017
Advanced Individuals with Excellence (5%), Tianjin University, China 2015 – 2016

SKILLS

Programming Languages: Python, C++, MATLAB.
Framework: PyTorch, TensorFlow, Keras
Remote and Collaboration Tools: Skilled in SSH, TeamViewer, Remote Desktop, Microsoft Teams, Zoom, and Slack for efficient remote collaboration and system management.
Productivity and Documentation: Experienced with Google Docs, Microsoft /Excel/PowerPoint/Word, Overleaf.