ZHIHANG YUAN

E-mail：hahnyuan@gmail.com Birthday: 15/8/1995
Website：zhihang.cc github.com/hahnyuan Google Scholar

Education

2017 - 2022: Ph.D., School of Computer Science, Peking University.(Advisor: Guangyu Sun)
2013 - 2017: Bachelor's degree, School of Pharmaceutical Sciences, Peking University.

Research Experiments

He specializes in Efficient AI. He has published articles in conferences and journals such as CVPR, NeurIPS, ICLR, ECCV, ISSCC, DATE, HPCA, and TCAD, with his articles being cited over 480 times according to Google Scholar.

Presently, he leads a research team dedicated to developing model compression algorithms and an engineering team responsible for the AI chip toolchain at a startup called Houmo AI.

His research primarily revolves around two core areas:

- **Compression of Neural Networks**
  - Quantization, low-rank decomposition, and software-hardware co-optimization
- **Simpler and Smarter Models**
  - Dynamic networks, network architecture search, and efficient Transformer design

Publications

* indicates equal contribution, ✉ indicates communication author

- Shang Y*, **Yuan Z***, et al. PB-LLM: Partially Binarized Large Language Models. ICLR 2024.
- Zhang C, **Yuan Z**, et al. Algorithm-hardware co-design for Energy-Efficient A/D conversion in ReRAM-based accelerators. DATE 2024.
- **Yuan Z***, Shang Y*, et al. ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models. arXiv 2023.
- Shang Y*, **Yuan Z***, et al. MIM4DD: Mutual Information Maximization for Dataset Distillation, NeurIPS, 2023.
- **Yuan Z***, Lin N*, Liu J, et al. RPTQ: Reorder-based Post-training Quantization for Large Language Models. arXiv preprint arXiv:2304.01089, 2023.
- Niu L, Liu J, **Yuan Z**✉, et al. Improving Post-Training Quantization on Object Detection with Task Loss-Guided Lp Metric. arXiv preprint arXiv:2304.09785, 2023.
- **Yuan Z***, Liu J*, Wu J, et al. Benchmarking the Reliability of Post-training Quantization: a Particular Focus on Worst-case Performance. AdvML-Frontiers 2023.
- Shang Y*, **Yuan Z***, Xie B, et al. Post-training Quantization on Diffusion Models. CVPR 2023.
- Liu J, Niu L, **Yuan Z**✉, et al. PD-Quant: Post-Training Quantization based on Prediction Difference Metric. CVPR 2023.
- Han Y*, **Yuan Z***, Pu Y, et al. Latency-aware Spatial-wise Dynamic Networks, NeurIPS 2022.

- Li X*, **Yuan Z***, Guan Y, et al. Flatfish: a Reinforcement Learning Approach for Application-Aware Address Mapping. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2022.
- Li X, Bing Z, Guang Y, et al. Enabling High-Quality Uncertainty Quantification in a PIM Designed for Bayesian Neural Network. HPCA, 2022.
- **Yuan Z***, Xue C*, Chen Y, et al. PTQ4ViT: Post-Training Quantization Framework for Vision Transformers. European Conference on Computer Vision (ECCV), 2022.
- **Yuan Z**, Chen Y, Xue C, et al. PTQ-SL: Exploring the Sub-layerwise Post-training Quantization. arXiv preprint arXiv:2110.07809, 2021.
- **Yuan Z**, Jingze L, Xingchen L, et al. NAS4RRAM: Neural Network Architecture Search for Inference on RRAM-based Accelerators. SCIENCE CHINA Information Sciences, 2021.
- **Yuan Z***, Wu B*, Sun G, et al. S2DNAS: Transforming Static CNN Model for Dynamic Inference via Neural Architecture Search. European Conference on Computer Vision (ECCV oral), 2020.
- **Yuan Z**, Liu X, Wu B, et al. ENAS4D: Efficient Multi-stage CNN Architecture Search for Dynamic Inference. arXiv preprint, 2020.
- Guan Y, Sun G, **Yuan Z**, et al. Crane: Mitigating Accelerator Under-utilization Caused by Sparsity Irregularities in CNNs. IEEE Transactions on Computers (TC), 2020.
- Guan Y, **Yuan Z**, Sun G, et al. FPGA-based accelerator for long short-term memory recurrent neural networks. Asia and South Pacific Design Automation Conference (ASP-DAC), 2017.
- Wu B, Liu Z, **Yuan Z**, et al. Reducing overfitting in deep convolutional neural networks using redundancy regularizer. International Conference on Artificial Neural Networks (ICANN), 2017.

Work Experiments

2022 - Present | Senior Algorithm Engineer, Houmo AI Technology Co., Ltd., Beijing

- LLMs deployment: He developed specialized software for large language models and successfully deployed them on in-house AI chip, Houmo H30. This involved performing comprehensive quantization of the LLMs and ensuring seamless integration with the chip.
- Development of AI chip toolchain: The toolchain can deploy AI applications on the Compute-In-Memory AI accelerator developed by Houmo AI. It supports various models including vision-based models and LLMs. He is responsible for model parsing, model quantization, frontend graph optimization, and frontend IR design.
- Compute-In-Memory AI accelerator design: a) Design the quantization and dataflow targeting characteristics of Compute-In-Memory hardware. b) Design the vector module and special function modules and for efficiently supporting various complex operators. c) Develop simulators to evaluate the performance of executing neural networks.
- Lead the research team: Mentor 4 full-time interns and 7 joint training interns in total.

2016 - 2022 | Undergraduate/Ph.D. at Peking University

- Ph.D. Dissertation: Research on Neural Network Compression Algorithms for Practical Deployment.
- Design of SNN Accelerator: Developed the ANN2SNN tool for hardware design, which can convert ANN into SNN networks. Optimized hardware design to address issues encountered during deployment. Participated in the design and chip fabrication of two chips.
- Design of ultra-low-power asynchronous convolutional neural network (CNN) chips and contributed to the production of two chip tapeouts.

- Design high-energy-efficient face detection/recognition network architectures and vehicle recognition network architectures for application in security and surveillance products.

Honors

Peking University Outstanding Research Award

Peking University Public Welfare Practice Award

Peking University Innovation Practice Project Award