Haocheng Xi

Berkeley AI Research, University of California, Berkeley | xihc@berkeley.edu
xijiu9.github.io

EDUCATION

University of California, Berkeley Berkeley, CA
Ph.D. in Computer Science, Berkeley AI Research (BAIR) 09/2024 – Present
Advisor: Prof. Kurt Keutzer

Tsinghua University Beijing, China
B.Eng. in Computer Science & Technology 09/2020 – 06/2024
Institute for Interdisciplinary Information Sciences (IIIS)
Yao Class, led by Prof. Andrew C.C. Yao

University of Washington Seattle, WA
Visiting Student, Paul G. Allen School of Computer Science & Engineering 02/2023 – 08/2023
Advisor: Prof. Sheng Wang

Beijing No.8 High School Beijing, China
Experimental class for gifted and talented young, Excellent Graduate 09/2015 – 07/2020

RESEARCH INTERESTS

My research interests lie in efficient machine learning and model quantization. I aim to push the boundaries of how we can effectively compress and accelerate deep learning models while maintaining their accuracy and robustness.

SELECTED PUBLICATIONS

Sparse VideoGen: Accelerating Video Diffusion Transformers with Spatial-Temporal Sparsity
Haocheng Xi, Shuo Yang*, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, Song Han*
International Conference on Machine Learning (ICML), 2025. [arxiv] [code] [website]

QuantSpec: Self-Speculative Decoding with Hierarchical Quantized KV Cache
Rishabh Tiwari, Haocheng Xi, Aditya Tomar, Coleman Hooper, Sehoon Kim, Maxwell Horton, Mahyar Najibi, Michael W. Mahoney, Kurt Keutzer, Amir Gholami
International Conference on Machine Learning (ICML), 2025. [arxiv]

COAT: Compressing Optimizer states and Activation for Memory-Efficient FP8 Training
Haocheng Xi, Han Cai, Ligeng Zhu, Yao Lu, Kurt Keutzer, Jianfei Chen, Song Han
Under reviewed, 2024. [arxiv] [code] [website]

Jetfire: Efficient and Accurate Transformer Pretraining with INT8 Data Flow and Per-Block Quantization
Haocheng Xi, Yuxiang Chen, Kang Zhao, Kai Jun Teh, Jianfei Chen, Jun Zhu
Internation Conference on Machine Learning (ICML), 2024. [arxiv] [code]
Selected as **Spotlight Paper** in ICML 2024. [poster]

Training Transformers with 4-bit Integers
Haocheng Xi, Changhao Li, Jianfei Chen, Jun Zhu
Conference on Neural Information Processing Systems (NeurIPS), 2023. [arxiv] [code]

INTERNSHIP EXPERIENCE

Nvidia Research, Research Intern 03/2024 – 08/2024
Advisor: Prof. Song Han

COAT: Compressing Optimizer states and Activation for Memory-Efficient FP8 Training

- Introduced COAT, a framework that quantizes optimizer states and activations to FP8 precision, significantly reducing memory usage during large-scale model training.
- Proposed Dynamic Range Expansion for Optimizer states and Mixed-Granularity Activation Quantization, achieving outstanding accuracy and efficiency.
- Achieved a 1.54× reduction in training memory footprint and a 1.43× speedup compared to BF16 training, also doubled the training batch size to utilize GPU better.
- Training loss curve and downstream task performance were consistent with BF16 training, across language models and vision language models.

RESEARCH EXPERIENCE

Tsinghua Univeristy, Tsinghua Statistical AI & Learning Group (TSAIL) Beijing, China
Advisor: Prof. Jianfei Chen, Prof. Jun Zhu 06/2021 – 06/2024

Jetfire: Efficient and Accurate Transformer INT8 Pretraining

- Proposed a new framework for pretraining Transformer models using INT8 data flow, enabling quantization of activations, weights, and gradients within Transformer layers into 8-bit integers.
- Introduced a block-wise quantization strategy to accommodate low-precision training, maintaining accuracy comparable to FP16 baselines while reducing memory usage.
- Achieved significant gains on LLM, including 1.42× training speed-up and 1.49× lower memory usage.

Training Transformers with 4-bit Integers

- Presented the first framework for training transformer-based neural networks using 4-bit integers that is able to quantize all of the activations, weights, and gradients appearing in linear layers into INT4
- Identified the challenge of outliers in activations for ultra-low bit quantization, and proposed a Hadamard quantizer that greatly improves the training accuracy on NLP and CV transformer models
- Leveraged sparsity in gradients, and designed a sampling algorithm to de-bias the quantization and reduce the multiply-accumulate (MAC) computation to achieve speed up
- Implemented a prototypical implementation of our algorithm, achieving up to 2.2× speed up for the linear layer, up to 6.48× speed up for inference, and up to 1.35× for end-to-end training

University of Washington, Paul G. Allen School of Computer Science & Engineering Seattle, WA
Advisor: Prof. Sheng Wang

Corpus Deletion for Pre-Trained Language Models 02/2023 – 09/2023

- Aimed at removing the information in a subset of the training data from the large language models, motivated by privacy concerns and eliminating erroneous information in the data

HONORS

Fellowship of Tsinghua Xuetang Talents Program Among top 300 / 3000 Tsinghua students each year
Athletic Excellence Scholarship In 2022
First Prize of National Senior High School Mathematics Competition In 2019

SKILLS

Language: TOFEL: Total 110 (Reading 29, Listening 29, Speaking 24, Writing 28)
GRE: Quantitive 170, Verbal 158, Writing 4.0
Programming and Software: Python, CUDA, C++, Bash, Git, LaTeX
Deep Learning Package: PyTorch, Transformers, Triton, PEFT, TransformerEngine