

Jiayi Tian

+1 (805) 245 0298 | jiayi_tian@ucsb.edu | github.com/ttttttris | linkedin.com/in/jiayi-tian-32b9652a5/

Focus on efficient training and inference of LLMs (Tensor decomposition, Pruning, Quantization, Knowledge Distillation)

EDUCATION

University of California, Santa Barbara, Ph.D. in Computer Engineering | CA, USA **3.9/4.0**

Fall 2023 - ongoing

Nanjing University, B.Eng. in VLSI Design & System Integration | China **4.5/5.0**

Fall 2019 - Jun 2023

INDUSTRIAL EXPERIENCE

Intel Corporation, Research Intern | Portland, OR

June. 2024 - Sep. 2024

- Proposed a tensor-compressed LLM training accelerator using FPGA with optimized compute ordering, dataflow, and memory allocation.
- Achieved up to $48\times$ memory efficiency and $3.6\times$ energy efficiency compared to Nvidia RTX 3090.
- Resulting paper under review at IEEE TCAD.

AMD-Xilinx Technology, Co-Op/Intern | Beijing, China

June 2023 - Sep 2023

- Developed a C++/HLS Transformer training framework with custom tensorized linear layers and nonlinear operations for LLM acceleration.
- Achieved $30\times \sim 52\times$ saving in model size for end-to-end Transformer training.

SKILLS & RESEARCH INTERESTS

Languages & Tools Python, PyTorch, TensorFlow, Huggingface, C/C++, High-level Synthesis (HLS), Vivado/Vitis/XRT

ML & NLP

Large Language Models (LLMs), Efficient Training/Inference Speedup (Model Compression, Pruning, SVD/Tensor-decomposition, Distillation, Quantization)

PUBLICATIONS & PREPRINTS

BEBERT: Efficient and robust binary ensemble BERT

Tian, Jiayi, Chao Fang, Haonan Wang, and Zhongfeng Wang, ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

Ultra Memory-Efficient On-FPGA Training of Transformers via Tensor-Compressed Optimization

Jiayi Tian, Jinming Lu, Hai Li, Xiangwei Wang, Cong (Callie) Hao, Ian Young, Zheng Zhang, under review at IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. arXiv preprint arXiv:2501.06663.

FETTA: Flexible and Efficient Hardware Accelerator for Tensorized Neural Network Training

Jinming Lu, Jiayi Tian, Hai Li, Ian Young, Zheng Zhang, under review at IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

RESEARCH PROJECTS

Structural Pruning for Efficient LLM Inference leveraging Tensor Decomposition

Aug. 2024 - Current

- Explored structural pruning of LLM that leveraging the low-rankness of data to prune model weights.
- Proposed head-wise SVD, joint PCA, and Nyström methods across decoder modules.
- Demonstrated superior results on LLaMA and other large-scale LLMs over existing pruning baselines.

Training Accelerator Design for Tensor-Compressed Transformer Models

Sep. 2023 - Dec. 2024

- Designed a tensor-decomposition-based training scheme that reduces parameter count by $30\times \sim 52\times$.
- Introduced bidirectional tensor contraction to enhance memory and compute efficiency, especially in long-sequence training and inference.
- Built an HLS-based Transformer training engine achieving up to $48\times$ memory efficiency and $3.6\times$ energy efficiency compared with Nvidia RTX 3090.

Binary-Quantized Ensemble LLM for Fast and Robust Language Model Inference

Apr. 2021 - June. 2023

- Developed BEBERT, a novel quantization-ensemble strategy enabling efficient and accurate 1-bit BERT inference.
- Leveraged efficient knowledge distillation strategy for high training efficiency.
- Achieved $13\times$ model size reduction and $15\times$ compute savings over standard BERT with minimal accuracy loss.
- Proposed early-exit inference variant, further cutting compute by $20\% \sim 40\%$ on GLUE benchmark.