# Haocheng Xi

Yao Class, IIIS, Tsinghua University | xihc20@mails.tsinghua.edu.cn

## EDUCATION

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

**Relevant Coursework**

Probability and Statistics (4.0) | Advanced Topics in Linear Algebra (4.0) | Discrete Mathematics (1) (2) (4.0) | Mathematics for Artificial Intelligence (4.0)

Machine Learning (4.0) | Deep Learning (4.0) | Computer Vision (4.0) | Natural Language Processing (4.0) | Quantum Computer Science | AI+X Computing Acceleration (4.0) | General Physics (2) (4.0)

**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

## PUBLICATIONS

**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]
GitHub repo. [code]
Selected as Spotlight Paper. [poster]

**Training Transformers with 4-bit Integers**
*Haocheng Xi, Changhao Li, Jianfei Chen, Jun Zhu*
Conference on Neural Information Processing Systems (NeurIPS), 2023. [arxiv]
GitHub repo received 100+ stars. [code]
Selected as huggingface daily paper. [link]

## RESEARCH EXPERIENCE

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

**Pretraining with 8-bit Integers**

- Reduce the communication latency and training time of neural networks by quantizing the data flow to INT8
- Expecting to implement INT8 training recipe using block-wise quantization

**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
- Completed a first-author paper accepted by NeurIPS 2023

**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
- Constructed an unlearning algorithm that alternates ascent and descent steps on forget dataset and retain dataset respectively
- Identified the gradient conflicting problem between ascent and descent steps, and proposed to preserve the perpendicular component of the ascent steps
- Designed an ascent-aware descent step to improve the forgetting ability of the unlearning algorithm
- Effectively removed thousands of data instances while maintaining model stability and performance, in contrast of hundreds of data instances in previous works

## PROJECT EXPERIENCE

**Multi-Core DNN Accelerator based on Network-on-Chip (NoC) on FPGAs** 06/2022 – 09/2022
Course Project of AI+X Computing Acceleration
Supervisor: Prof. Kaisheng Ma

- Implemented a convolution module and an interconnection network for concurrent communication
- Designed a NoC-based Multi-Core DNN accelerator, achieving 4x speed up when using multi-core

**Reinforcement Learning on Sichuan Mahjong** 02/2022 – 07/2022
Course Project of Deep Learning
Supervisor: Prof. Yi Wu

- Applied Reinforcement learning to Mahjong, an imperfect-information extensive-form game
- Trained a DQN network, resulting in a close to human-level skill in Mahjong

**Artistic Image Synthesis using SinGAN** 02/2022 – 07/2022
Course Project of Computer Vision
Supervisor: Prof. Yang Gao

- Generated a synthesized image with unique artistic style with only an origin image and a style for input
- Combined AdaIN and SinGAN to train a GAN with one image

## 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

## 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

## EXTRACURRICULAR ACTIVITIES

**Sports:** Members of the school's ultimate frisbee team and the department's soccer team
**Volunteer work:** Participating in the Program Buddy project, providing coding assistance to beginners