Ziheng Qin

zihengq@comp.nus.edu.sg · (+65) 86948476 · ziheng-qin-635551184

Education

National University of Singapore (NUS), Singapore 2021 – Present

PhD student in Computer Science, expected 2025

University of Southern California (USC), CA, USA 2020 – 2021

Master in Computer Science

University of Michigan, Ann Arbor (UM), MI, USA 2017 – 2019

B.S.E in Computer Science

Shanghai Jiao Tong University (SJTU), Shanghai, China 2015 – 2019

B.S.E in Electronic and Computer Engineering (ECE)

Industry Experience

TikTok Pte. Ltd. Singapore 2024 Aug - Present

Research Engineer Intern Manager: Kanchan Sarkar

Efficient Model-Data Co-evolution. Solution for continually improving deep learning model and maintaining a high-quality growing dataset with efficient and affordable automatic cleaning and active annotation. Granted 200k SGD research funding from TikTok.

- Propose and lead the research with collaboration between HPC-AI Lab and TikTok.
- Designed to be able to efficiently 1. boost the quantity and quality of annotations 2. manage growing data 3. improve sample efficiency 4. estimate annotation source quality 5. maintain an improving model

Ali Tongyi Lab, Research Collaboration. Singapore 2023 Mar - 2024 July

Student Researcher Instructor: Yang You Manager: Baigui Sun

Data curation for Data-centric AI. Propose and implement research ideas, experiment and write research papers. Publishing at ICLR 2024 (Oral) and ECCV 2024 as first author.

- (ICLR 2024 Oral) InfoBatch. Improved the sample efficiency with negligible overhead and lossless performance during training. Saved cost on various tasks by 20% $\sim$ 40%, used for business model training.
- (ECCV 2024) Proposed and implemented an efficient online algorithm for data cleaning and selection to deal with growing data.

Google LLC. California, USA 2021 May-July

Software Engineer Intern Manager: Yixin Shi

DLVM Cluster for VM-Based Distributed Training: A VM-based distributed training solution.

- Designed and implemented a configurable system based on Terraform for automatically setting up distributed training environments on the cloud. The system could create and set up a cluster of cloud VMs, scale the training cluster, execute commands in batch, and manage files with a simple YAML file and command line. It supports different Machine learning frameworks and distributed learning frameworks.
- Provided a choice with higher freedom and availability than the existing solution (TPU, GKE clusters etc).

Research Experience

ECCV 2024 2023-2024

First Author Supervisor: Yang You Collaboration: Ali Group

Dataset Growth

- Proposed an efficient online algorithm for data cleaning and selection, resulting in a growing dataset that keeps up to date with an awareness of cleanliness and diversity, and information for efficient data sampling.
- Could improve data quality/efficiency on single-modal and multi-modal tasks, which could benefit Vision Language Model training (using 25% data to achieve better results than original data).
- The method was designed to be scalable and efficient.

ICLR 2024 Oral (1.2%) 2022-2023

First Author Supervisor: Yang You Collaboration: Ali Group

InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning

- Proposed a novel framework to achieve lossless training acceleration by unbiased dynamic data pruning.
- Consistently obtains lossless training results on classification, semantic segmentation, object detection, vision pertaining (MAE and Diffusion), and instruction tuning (LLaMA) with 20% $\sim$ 40% saving.
- Negligible overhead, accelerate multiple hours with cost in seconds. Practical for real usage.
- Compatible with various optimizers, coreset selection methods, and LoRA.

ICCV 2023 2022-2023

Co-author Supervisor: Yang You

Preventing zero-shot transfer degradation in continual learning of vision-language models

- Zero-shot ability of Vision Language Model like CLIP would significantly degrade during continual learning due to catastrophic forgetting. We proposed a novel method ZSCL to prevent zero-shot transfer degradation in the continual learning of vision-language models in both feature and parameter space.
- We proposed a more challenging Multi-domain Task Incremental Learning (MTIL) benchmark to evaluate different methods. ZSCL outperforms other methods in the traditional class-incremental learning setting and the MTIL by 9.7% average score.

AAAI 2023 Outstanding Paper 2021-2022

Co-author Supervisor: Yang You

CowClip: reducing CTR prediction model training time from 12 hours to 10 minutes on 1 GPU

- This work identified the bottleneck of large batch training in the recommendation system and solved it with the proposed CowClip algorithm, which greatly (72x) reduced the training time and cost.
- My contribution: help to analyze the large batch training bottleneck in the recommendation system, namely how the long-tailed feature frequency in a batch affects weight norm and stability during batch size scaling, and participated in the algorithm design.

Shanghai Jiao Tong Univesity & Intel Shanghai, China 2019 Summer

Project Member Instructor: Yong Long(SJTU), Forrest Zhao(Intel), Ruoyu Ying(Intel)

Cloud Gaming Architecture based on StarlingX and Akraino: An edge cloud gaming architecture.

- Used StarlingX and Kubernetes to build an edge server that did remote graphical computing in containers for clients during cloud gaming with low latency, high compatibility, and fast deployment.
- Attended Open Infrastructure Summit (Shanghai 2019) and did a presentation for this project as a speaker.
- Won Golden prize of SJTU-UMJI 2019 Summer Design Expo.

Skills

- Programming Languages: Python, C/C++, Java
- Platform: Linux/CentOS
- Development: Machine Learning/Reinforcement Learning/Deep Learning/Data Mining/Distributed Training, MySQL, Spark, XGBoost, CSS, JavaScript, Html, Hadoop, MATLAB, Mathematica, PPT, LaTeX

Miscellaneous

- GitHub: https://github.com/henryqin1997
- Google Scholar: https://scholar.google.com/citations?user=I04VhPMAAAAJ
- Languages: English, Chinese(Mandarin)