Zangwei Zheng

✉ zhengzangw@gmail.com · 🏠 zhengzangw.github.io

Education

National University of Singapore Aug. 2021 – Jun. 2025 (expected)
Ph.D. in Computer Science, supervised by Prof. Yang You *Singapore*

- Research Achievement Award of NUS

Nanjing University Sep. 2017 – Jun. 2021
B.S. in Computer Science and Technology, National Elite Program in Computer Science *Jiangsu, China*

- **GPA:** 4.61/5.00 (92.2/100, top 2%)

Research Interest

Video Generation: large-scale video generation model, video control and alignment.
Efficient Machine Learning: computation-efficient training (accelerated optimizer, large batch training, incremental training), memory-efficient training, efficient inference.
Large-scale Deep Learning Optimization: optimizer design (faster, robust, memory-efficient, etc.), optimizer explanation, data-model-algorithm connections.

Academic Research Experience

National University of Singapore (HPC-AI Lab) May 2019 – Present
Ph.D. student, supervised by Prof. Yang You *Singapore*

- Large language model inference acceleration by predicting response length and sequence scheduling.
- Continual learning of vision-language model to prevent zero-shot performance degradation.
- Acceleration of recommendation system training by large batch training.
- Introduce prompt learning for domain generalization with vision transformer.

University of California, Berkeley (iCyPhy, DOP Center) Apr. 2020 – May 2021
Research intern, supervised by Prof. Alberto Sangiovanni-Vincentelli & Dr. Xiangyu Yue *(remote) CA, US*

- Few-shot Domain Adaptation via Self-supervised Learning with Clustering
- Proposed scene-aware learning with better backbones and data augmentations for radar object detection.

Industry Research Experience

HPC-AI Tech Mar. 2024 – Present
Research intern, team lead & first author of the video generation model **Open-Sora 17k stars** *Singapore*

- Design, develop and train Transformer-based video generation model from scratch.
- Design and develop the data processing pipeline. Incorporate features including rectified flow, temporal VAE, image-conditioned generation, dynamic resolution support, etc.

ByteDance Jun. 2021 – Jun. 2022
Research intern, in charge of large batch training for click-through rate prediction model *Singapore*

- Transformed the asynchronous CTR training model into the large-scale synchronous training framework.
- Deployed CowClip algorithm with batch size 512k and improved the AUC of CTR prediction (**AAAI 2023**).

Publications

1. **Helen: Optimizing CTR Prediction Models with Frequency-wise Hessian Eigenvalue Regularization**
   Zirui Zhu, Yong Liu, Zangwei Zheng, Huifeng Guo, Yang You **WWW 2024**
2. **Openmoe: An early effort on open mixture-of-experts language models** Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, Yang You **ICML 2024**
3. **InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning** Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Daquan Zhou, Yang You **ICLR 2024**
4. **Response Length Perception and Sequence Scheduling: An LLM-Empowered LLM Inference Pipeline**
   Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, Yang Luo, Xin Jiang, Yang You **Neurips 2023**

5. **To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis**
   Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, Yang You **Neurips 2023**
6. **Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models**
   Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, Yang You **ICCV 2023**
7. **A Study on Transformer Configuration and Training Objective**
   Fuzhao Xue, Jianghai Chen, Aixin Sun, Xiaozhe Ren, Zangwei Zheng, Xiaoxin He, Yongming Chen, Xin Jiang, Yang You **ICML 2023**
8. **CAME: Confidence-guided Adaptive Memory Efficient Optimization** Yang Luo, Xiaozhe Ren, Zangwei Zheng, Xin Jiang, Zhuo Jiang, Yang You **Distinguished Paper Award (0.8%), ACL 2023**
9. **CowClip: Reducing CTR Prediction Model Training Time from 12 hours to 10 minutes on 1 GPU** Zangwei Zheng, Pengtai Xu, Xuan Zou, Da Tang, Zhen Li, Chenguang Xi, Peng Wu, Leqi Zou, Yijie Zhu, Ming Chen, Xiangzhuo Ding, Fuzhao Xue, Ziheng Qing, Youlong Cheng, Yang You **Distinguished Paper Award (0.1%), AAAI 2023**
10. **Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation**
    Xiangyu Yue*, Zangwei Zheng*, Shanghang Zhang, Yang Gao, Trevor Darrell,
    Kurt Keutzer, Alberto Sangiovanni-Vincentelli **CVPR 2021**
11. **Scene-aware Learning Network for Radar Object Detection**
    Zangwei Zheng, Xiangyu Yue, Kurt Keutzer, Alberto Sangiovanni Vincentelli **ICMR-W 2021**

SKILLS

Languages	Python, C, C++, LaTeX
Frameworks	PyTorch, TensorFlow, Huggingface, OpenCV, Scikit-learn, NumPy

Last update: 28 May 2024