Fuzhao XUE

+65 83434586 | xuefuzhao@outlook.com | https://xuefuzhao.github.io | Singapore

EDUCATION

Doctor of Philosophy (Ph.D.) Dec. 2024 (expected)
National University of Singapore Singapore

- Majored in Computer Science
- Supervisor: Yang You
- Google Ph.D. Fellowship
- President's Graduate Fellowship
- AAAI 2023 Distinguished Paper Award (Co-author)
- Published papers at ICML, NeurIPS, ACL, AAAI, CVPR
- Internship at Google (Brain/DeepMind Team) and worked with Yi Tay and Mostafa Dehghani
- Internship at NVIDIA (GEAR Team) and worked with Linxi (Jim) Fan and Yuke Zhu
- Thesis: Towards Efficient Transformer Scaling

Master of Engineering (M.Eng.) Jul. 2021
Nanyang Technological University Singapore

- Majored in Computer Science and Engineering
- Supervisor: Eng-Siong Chng and Aixin Sun
- Published paper at AAAI, ICML and ICASSP.
- Thesis: Refining latent multi-view graph for relation extraction
- GPA: 5.0/5.0

RESEARCH INTEREST

- **Machine Learning:** Foundation Model Scaling, Conditional Computation, Adaptive Computation
- **Natural Language Processing:** Large Language Model Pre-training and Instruction Tuning
- **High Performance Computing:** Model Parallelism and Large-scale Deep Learning System for Long Sequence.

HIGHLIGHT RESEARCH

- **OpenMoE:** The FIRST FULLY OPEN MoE-based Decoder-only LLM Trained over Chinchilla Scaling Law.
  - ➢ OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models
    - ✧ **Fuzhao Xue**, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou and Yang You
- **Token Crisis:** The FIRST Project Studying Training LLM with Limited and Repeated Data
  - ➢ To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis. NeurIPS 2023
    - ✧ **Fuzhao Xue**, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, Yang You
- **AdaTape:** Adaptive Computation Foundation Model Supporting Elastic Input Sequence and Dynamic Read&Write,
  - ➢ Adaptive Computation with Elastic Input Sequence. ICML 2023
    - ✧ **Fuzhao Xue,** Valerii Likhosherstov, Anurag Arnab, Neil Houlsby, Mostafa Dehghani, Yang You
- **Sequence Parallelism and Ring Self-Attention:** The FIRST Work Trying to Solve Long Sequence Training with Distributed System Instead of Improving Attention Layer Efficiency.
  - ➢ Sequence Parallelism: Long Sequence Training from System Perspective, ACL 2023
    - ✧ Shenggui Li*, **Fuzhao Xue***, Yongbin Li, Yang You
- **WideNet:** The SECOND Most Cited MoE-based Vision Transformer based on Google Scholar until Mar 2024
  - ➢ Go Wider Instead of Deeper, AAAI 2022
    - ✧ **Fuzhao Xue**, Ziji Shi, Futao Wei, Yong Liu, Yang You
- Full Publication List: https://xuefuzhao.github.io/publications/

INDUSTRY RESEARCH EXPERIENCE

NVIDIA | Work Remotely from Singapore ***Research Intern*** Jan. 2023-Present

In charge of multi-model foundation model scaling and robotics foundation model. Working with Linxi(Jim) Fan and Yuke Zhu.

- Training video foundation model for general video captioning with more high-quality data and compute.
- Designing robotics foundation model agent based on the general video foundation model.

Google | Work Remotely from Singapore ***Student Researcher*** Jul. 2022-Nov 2022

In charge of adaptive computation with adaptive model depth and elastic input sequence. Worked with Yi Tay and Mostafa Dehghani. Part of the work is accepted at ICML 2023.

- Implemented universal transformer and PonderNet on vision transformer based on Scenic and made vision universal transformer open source.
- Proposed AdaTape to achieve adaptive computation with elastic input sequence, which is a novel strategy that enables dynamic computation in neural networks via adaptive tape token. (ICML 2023)

ACADEMIC RESEARCH EXPERIENCE

National University of Singapore | Singapore ***Ph.D. Candidate*** Jul. 2021-Present

In charge of efficient and effective general training framework for both computer vision and natural language processing to scale transformer. Published at NeurIPS 2023, ICML 2023, ACL 2023, AAAI 2022.

- Studied the relationship between transformer configuration and training objectives. (ICML 2023)
- Proposed sequence parallelism to train transformer with longer sequence from system perspective. (ACL 2023)
- Proposed to go wider instead of deeper, compressing along depth by parameter sharing and scaling along width by mixture-of-experts to construct a parameter-efficient framework. (AAAI 2022)

Nanyang Technological University | Singapore ***M.Eng. Student*** Jul. 2020-Jun. 2021

In charge of design of improving dialogue-level relation extraction by adaptive graph pooling proposed. Published paper at AAAI 2021, ICASSP 2022.

- Identified indictive words using Dynamic Time Wrapping Pooling in an unsupervised manner with high accuracy.
- Improved the state-of-the-art of dialogue-level relation extraction by 6% on DialogRE.

National University of Singapore | Singapore ***Research Intern*** Oct. 2019-Apr. 2020

In charge of designing hybrid speech recognition framework with Pytorch and Kaldi. Published paper at ICML 2020.

- Used Pytorch underlying operations to implement acoustic models such as DNN, LSTM, SRU, RRN and RTN.
- Modeled relational thinking via Deep Graph Process, reducing WER by 10% relatively on Chime2&5 and SWBD.

SKILLS

- **Programming:** Python, C/C++, C#, Matlab
- **Frameworks:** PyTorch, JAX, TensorFlow, Keras, Theano, Scikit-learn, pandas
- **Advanced:** T5x, Flaxfomer, Scenic, SeqIO, Deepspeed, Megatron, Huggingface

SERVICES

- **Conference 2024:** ICLR Reviewer, ICML Reviewer, COLM Reviewer
- **Conference 2023:** EMNLP Reviewer, NeurIPs Reviewer, ACL PC Member, CVPR Reviewer
- **Conference 2022:** EMNLP Reviewer, ICML Reviewer, SIGIR PC Member,
- **Journal:** TKDE Reviewer