Yizhe Xiong (熊翊哲)

xiongyizhe2001@163.com · Google Scholar · linkedin.com/in/bostoncake · xiongyizhe.xyz

Education

Tsinghua University, Beijing, China 2022 – Present (expected June 2027)

Ph.D. in Software Engineering, GPA 3.9/4.0
Research focus: Transfer Learning, Pretraining and Fine-Tuning Foundation Models

Tsinghua University, Beijing, China 2018 – 2022

B.S. in Computer Science and Technology, GPA 3.7/4.0

Skills

- Programming Language: C, C++, Python
- Deep learning development: PyTorch, torchvision, Huggingface transformers, timm, MMCV, etc.

Experience

Kuaishou Technology October 2023 – Present

Machine Learning Intern @ Department of Foundation Models and Multimedia

Researching in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), focusing on pretraining, fine-tuning, and enhancing model capabilities.

- Submitted two first-author articles: Temporal Scaling Law for Large Language Models and UniAttn: Reducing Inference Costs via Softmax Unification for Post-Training LLMs, both are currently under review.
- Contributed to the development of the tokenizer and Mixture-of-Experts (MoE) in the KuaiYii model.
- Developed the hyperparameter selection method for the KwaiYii model using Temporal Scaling Law for Large Language Models.
- Developed the fine-tuning process to enable long-context support for the KwaiYii model.

Microsoft (China) Co. Ltd May 2021 – August 2021

Software Engineering Intern @ Cloud and AI (C+AI), Microsoft Azure

Enhanced Synapse and developer tooling in Azure Powershell.

- Developed and maintained key features for managing Synapse workspaces.
- Streamlined the PowerShell development process by optimizing the integration pipeline for the AutoRest tool.

Publications

PYRA: Parallel Yielding Re-Activation for Training-Inference Efficient Task Adaptation

European Conference on Computer Vision (ECCV) 2024

Authors: **Yizhe Xiong**, Hui Chen, Tianxiang Hao, Zijia Lin, Jungong Han, Yuesong Zhang, Guoxin Wang, Yongjun Bao, Guiguang Ding

Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation

International Conference on Computer Vision (ICCV) 2023

Authors: **Yizhe Xiong**, Hui Chen, Zijia Lin, Sicheng Zhao, Guiguang Ding

Breaking the Stage Barrier: A Novel Single-Stage Approach to Long Context Extension for Large Language Models

International Conference on Computational Linguistics (COLING) 2025

Authors: Haoran Lian*, Junmin Chen*, Wei Huang*, **Yizhe Xiong***, Wenping Hu*, Guiguang Ding, Hui Chen, Jianwei Niu, Zijia Lin, Fuzheng Zhang, Di Zhang (* denotes equal contribution)

Scaffold-BPE: Enhancing Byte Pair Encoding for Large Language Models with Simple and Effective Scaffold Token Removal

AAAI Conference on Artificial Intelligence (AAAI) 2025

Authors: Haoran Lian, **Yizhe Xiong**, Jianwei Niu, Shasha Mo, Zhenpeng Su, Zijia Lin, Peng Liu, Hui Chen, Guiguang Ding

LBPE: Long-token-first Tokenization to Improve Large Language Models

International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2025

Authors: Haoran Lian, **Yizhe Xiong**, Zijia Lin, Jianwei Niu, Shasha Mo, Hui Chen, Peng Liu, Guiguang Ding

CartesianMoE: Boosting Knowledge Sharing among Experts via Cartesian Product Routing in Mixture-of-Experts

2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025) Main conference

Authors: Zhenpeng Su, Xing Wu, Zijia Lin, **Yizhe Xiong**, Minxuan Lv, Guangyuan Ma, Hui Chen, Songlin Hu, Guiguang Ding

Mitigating Hallucinations in Multi-modal Large Language Models via Image Token Attention-Guided Decoding

2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025) Main conference

Authors: Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, **Yizhe Xiong**, Zijia Lin, Jungong Han, Guiguang Ding

Other Selected Papers

UniAttn: Reducing Inference Costs via Softmax Unification for Post-Training LLMs

arXiv:2502.00439 (Under Review)

Authors: **Yizhe Xiong**, Wei Huang, Xin Ye, Hui Chen, Zijia Lin, Haoran Lian, Zhenpeng Su, Jungong Han, Guiguang Ding Keywords: Large Language Model (LLM), Language Modeling, Model Compression, Post-Training, Supervised Fine-Tuning (SFT)

Temporal Scaling Law for Large Language Models

arXiv:2404.17785 (Under Review)

Authors: **Yizhe Xiong**, Xiansheng Chen, Xin Ye, Hui Chen, Zijia Lin, Haoran Lian, Zhenpeng Su, Wei Huang, Jianwei Niu, Jungong Han, Guiguang Ding Keywords: Large Language Model (LLM), Language Modeling, Scaling Law

Awards

Academic Scholarship, School of Software, Tsinghua University — December 2024
First Place and Gold Prize, VISION'24 Data Challange, ECCV 2024 — September 2024
Academic Scholarship, School of Software, Tsinghua University — October 2023
Outstanding Graduate, Department of Computer Science and Technology, Tsinghua University — June 2022

Additional Information

- GitHub: https://github.com/bostoncake
- Fluent in English (TOEFL 113)