Ziyang Yu

ericyu8857@gmail.com +1 548-577-0264 github.com/Ziyang-Yu

Education Background

University of Waterloo **Ontario, Canada**
M.Eng, Electrical and Computer Engineering *Sep. 2023-Present*
GPA: 3.96/4.00 | Coursework: Deep Learning, Computer Network

Southern University of Science and Technology (SUSTech) **Guangdong, China**
B.Sc. (Hons), Mathematics and Applied Mathematics *Aug. 2019-Jun. 2023*
GPA: 3.75/4.00 | Coursework: Convex Optimization, Linear Algebra

Research Experience

AI Generative Content for Health Care Data **Yale University**
Research Assistant | Advisor: Rex Ying *Apr. 2024- Aug. 2024*

- **Pipeline Design**: Fine-tuned GPT-2 using Hugging Face, designed and implemented a text-based drug prediction task, significantly improving model accuracy.
- **Data Processing**: Preprocessed raw MIMIC III data, unifying database formats with text data, and created efficient prompts.
- **Data Augmentation**: Employed Retrieval-Augmented Generation (RAG) to retrieve similar patient diagnostic information, utilizing in-context learning to substantially enhance the model's overall performance.

Distributed Co-training of LLM and GNN **Emory University**
Research Assistant | Advisor: Liang Zhao *Apr. 2024- Aug. 2024*

- **Algorithm Implementation**: Used Huggingface and Pytorch to implement synchronous training of BERT/LLaMA-7b and GraphSAGE, significantly improving the accuracy of paper classification tasks.
- **Model Optimization**: Applied DeepSpeed technology for pipeline parallelism, accelerated model training and inference, and used LoRA and Offloading techniques to reduce training parameters and increase batch size.
- **Experimental Design**: Conducted experiments using NVIDIA A100 80GB, achieving 74.06% accuracy on the OGB-Arxiv dataset and 78.19% on the Cora dataset.

Systematic Survey of Resource-Efficient Large Language Models **Emory University**
Research Assistant | Advisor: Liang Zhao *Sept. 2023-Oct. 2023*

- **Paper Research**: Investigated low-rank decomposition techniques in large language models, conducted mathematical theoretical analysis, and wrote research papers.

Staleness-Alleviated Distributed Graph Neural Network Training **Emory University**
Research Assistant | Advisor: Liang Zhao *Jul. 2022-Jan. 2023*

- **Algorithm Implementation**: Used PyTorch to implement distributed training for Graph Neural Networks, significantly accelerating training on large-scale graph data, and utilized Plasma technology to speed up reading and updating.
- **Algorithm Optimization**: Proposed using an LSTM-GCN model to capture the temporal evolution of node embeddings, reducing data staleness issues, and employing pre-training techniques to improve model transferability across different datasets, significantly reducing training time.
- **Experiment Verification**: The proposed framework achieved an F1 score of 97.02% on the Reddit dataset and 80.21% on OGB-Products, significantly enhancing the performance and convergence speed of distributed GNN training.

Publications

- **Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models**
  Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, **Ziyang Yu**, Mengdan Zhu, Yifei Zhang, Carl Yang, Yue Cheng, Liang Zhao.
- **Staleness-Alleviated Distributed Graph Neural Network Training via Online Dynamic-Embedding Prediction**
  Guangji Bai*, **Ziyang Yu***, Zheng Chai, Yue Cheng, Liang Zhao.

Skills

- **Computer Languages**: C/C++, Python, Assembly Language, Java, JavaScript, Bash, Matlab.
- **Machine Learning Libraries**: NumPy, Sci-kit Learn, HanLP, Pandas, Matplotlib, Pytorch, Pytorch Geometric.
- **Mathematics**: Algebra, Analysis, Geometry, Topology.