Xiaohan Xu, 许肖汉

(+86) 17753073352 · Homepage · shawnxxh@gmail.com · Github

Bio

I am a second-year graduate student at Information Institute of Engineering (IIE), UCAS, mentored by Prof. Hongbo Xu. I am self-motivated and actively dedicated to my studies. Furthermore, I possess good planning skills and have the ability to set clear goals. My research interests focus on language models and knowledge graphs.

Research Interests

Natural Language Processing, Large Language Model, Knowledge Graph Representation Learning.

Education

University of Chinese Academy of Sciences *M.S.* China, 2021.9 - 2024.6 (expected)

- Information Institute of Engineering, IIE (GPA: 3.84/4)

Beijing Institute of Technology *B.S.* China, 2017.9 - 2021.6

- College of Automation (GPA: 89.1/100, Ranking: 5/32)
- ***National Scholarship (<2%), National Encouragement Scholarship (<5%)***

Illinois Institute of Technology, IIT, *Summer Research* U.S., 2019.7 - 2019.9

- Statistics Learning and Big Data.

University of Alberta, *School-sponsored Exchange Program* Cananda, 2018.7 - 2018.8

Publication

- **Xiaohan Xu**, Peng Zhang, Yongquan He, Chengpeng Chao, Chaoyang Yan. *Subgraph Neighboring Relations Infomax for Inductive Link Prediction on Knowledge Graphs*. **Long presentation (25%) at IJCAI 2022**
- **Xiaohan Xu**, Xuying Meng, Yequan Wang. *PoKE: Prior Knowledge Enhanced Emotional Support Conversation with Latent Variable*. **Preprint 2023**

Awards & Honors

- **National Scholarship (< 2%)** 2018
- **National Encouragement Scholarship (< 5%)** 2019
- BIT Excellent Student Model (< 2%) 2018

Research Experience

Baidu Inc. | XiaoDu Cloud Research Intern 2022.11-now

- **LLM**: Build Large Language Model (LLM) aligned with human. Work on supervised fine-tuning (SFT) stage.
  - *SFT Data*: Participate in building 1M+ Chinese SFT data by self-instruct and filter out low-quality data.
  - *SFT Model*: 1) Fine-tune LLM on SFT data, 2) evaluate the SFT model from multiple views, 3) explore the influence of data quality and size.
  - The trained SFT model can follow the instructions and generate well-organized text.
- **Intelligent Exam Solving Model**: Unify and solve K-9 multi-subject & multi-type exams by instruct-tuning and knowledge injection.
  - *Prompt Build*: 1) Build diverse prompts for exam data; 2) Fine-tune SFT model on exam data.
  - *Knowledge Injection*: 1) Probe LM scarce knowledge; 2) Inject corresponding knowledge by post-training. 3) Train model in distributed manner by deepspeed and mpi.
  - Achieve accuracy of 82% on online English questions and 61% on Chinese questions.

BAAI | Recognition & Data Group Research Intern 2022.4-2022.10

- Supervised by Dr. Yequan Wang
- **Emotional Support Conversation (ESC)**: Design a dialogue system for emotional support by mining prior knowledge for quality and latent variable for diversity.
  - *Prior Knowledge Enhance*: Mine rich prior knowledge of response and strategy transition from relevant conversations.

  - *Diversity & Denoising*: Introduce latent variables to model one-to-many relationship of strategy, and thus improve diversity and denoise exemplars.
  - Get SOTA result on both automatic and human evaluation.

- **Pretrained LM & KG**: Explore to pre-train language model integrated with knowledge graphs by contrastive learning and multi-task learning.
  - *Contrastive Learning*: Do in-batch contrastive learning of representations between text and graph.
  - *Multi-task Learning*: Conduct MLM task for text and link prediction task for KG.

IIE | Personal Research **IJCAI 2022** Long Presentation GitHub 2021.7-2022.1

- **Inductive KG Link Prediction**: Work on predicting missing links between unseen nodes by their subgraph, and model neighboring relations and MI maximization.
  - *Neighboring Relations*: Exploit neighboring relations to characterize node feature and model the consistency of target relational path.
  - *MI Maximization*: Maximize mutual information between local and global views for modeling relations from a global perspective.
  - Achieve SOTA performance and improve by an average of 4.5 in Hits@10, especially on sparse graphs.

PROJECTS

IIT | Summer Research Project Leader GitHub 2021.7-2022.1

- **Win Rate and Duration of LOL Prediction**: Predict the win rate and time duration of LOL and do optimization of model.
  - *Model Design*: 1) Reduce redundancy of feature by view correlation matrix. 2) Explore and comopare different algorithms for win rate prediction, including LR, LDA and KNN.
  - *Optimization*: Use forward selection and the lasso to optimize model, reducing the number of predictors.
  - Get test error of 4.12% and 4 significant predictors of model eventually.

SKILLS

- **Programming Languages**: Python, Shell, C++, HTML, JavaScript.
- **NLP&Graph**: Pytorch, Transformers, Deepspeed, DGL.
- **Tools**: LaTeX, Git, Linux, GDB.
- **English**: CET6 (554), GRE (324)
- **Highlighted Courses**: Mathematical Analysis For Engineering (96/100), Linear Algebra (95/100), C++ Programming (100/100), Theory & Practice of Machine Learning (90/100), Deep learning for natural language processing (91/100)

ACTIVITIES

- **Volunteer Teaching**: I volunteered at a local school for children of migrant workers once a week during my 1st and 2nd years. In my 2nd year, I also went to teach in a rural area in Bozhou, Anhui.