HAIKANG DENG

+1 919-260-9698
frankdenghaikang@gmail.com

EDUCATION

University of North Carolina, Chapel Hill Aug 2019 - May 2023
B.S. in Computer Science & Statistics. Overall GPA: 3.96/4.0

RESEARCH & PROFESSIONAL EXPERIENCE

University of North Carolina, Chapel Hill Aug 2022 - Present
Research Assistant Advised by Prof. Colin Raffel

- Benchmarked various Learning from Human Feedback methods and studied their overoptimization problem
- Introduced an efficient weighted decoding method that aligns text to a given attribute with uni-directional reward model
- Explored language models' knowledge-learning process and their QA performance relative to their pre-training data
- Analyzed language model hallucination and tracked wrong answers in training corpus

Amazon May 2022 - Aug 2022
Software Engineer Intern Bellevue, WA

- Built a Horizonte Service for Local Landing Page which displays local products available for pick up
- Deployed the service to production and verified its reliability with production data
- Onboarded downstream dependencies to fetch data and extended JSP to render user interface
- Configured shopping portal page type and added routing rules from amazon.com

Lenovo May 2021 - Aug 2021
Software Engineer Intern Beijing, China

- Trained Encoder-Decoder LSTM for anomaly detection on time series data
- Participated in the design of Control Chart and Anomaly Detection Module
- Performed model tuning and data grouping which improved f1 score from 0.41 to 0.48

Zhongchao Credit Card Industry Development Co., Ltd Jun 2020 - Aug 2020
Software Engineer Intern Hangzhou, Zhejiang, China

- Built an Ethereum smart contract for medical data management with user interface
- Deployed the smart contract to private chain network and explored various data structures

PUBLICATIONS & PROJECTS

Benchmarking Learning from Human Feedback Methods July 2023 - Present

- Incorporated LLaMA2 reward model into OpenAssistant's code base and trained a GoldRM on RLHF datasets
- Used the GoldRM to create synthetic data for later experiments on LHF overoptimization

Controllable Text Generation with Uni-directional Reward Model Jan 2023 - July 2023

- Introduced Reward-Augmented Decoding (RAD), an efficient, performant, and generalizable weighted decoding method that steers text generation toward a desired attribute using a unidirectional reward model trained on task-specific data
- RAD outperforms other weighted decoding methods on detoxification and sentiment-controlled generation
- Applied RAD on various model families and sizes (e.g. GPT2 and LLaMA) and demonstrated its ability to generalize
- Publication: Reward-Augmented Decoding: Efficient Controlled Text Generation With a Unidirectional Reward Model
  Haikang Deng and Colin Raffel
  The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Knowledge Memorization of Large Language Models Aug 2022 - Jan 2023

- Demonstrated correlational and causal relationships between the number of relevant documents during pre-training and a model's question-answering accuracy
- Created a parallelized pipeline for entity linking and relevant document counting
- Showed that model scaling is inefficient in improving QA performance and explored retrieval augmentation as an alternative
- Publication: Large Language Models Struggle to Learn Long-Tail Knowledge
  Nikhil Kandpal, **Haikang Deng**, Adam Roberts, Eric Wallace, and Colin Raffel
  40th International Conference on Machine Learning (ICML)

Neural Methods of Image Captioning Jun 2021 - Dec 2021

- Compared Vanilla-LSTM, LSTM with attention, and Transformer on Image Captioning
- Achieved BLEU-1, BLEU-2 score of 67.1, 44.3 with LSTM with attention on MS COCO

TECH SKILLS

Programming Skills: Python, Java, MySQL, R, Matlab, HTML; Pytorch, Tensorflow
Models and Algorithms: Transformers, RNN/LSTM, CNN, MLP, SVM
Topics and Concepts: Regression, Time Series, Multimodality, Prompt Engineering, Controlled Text Generation