Yihuai Hong

Email: yihuaihong@gmail.com Phone: +86 15815289871 Homepage: https://yihuaihong.github.io

Education

Bachelor of Engineering, Computer Science Sept 2020 – Present
South China University of Technology(**SCUT**), Guangdong, China Expected Graduation: June 2024
GPA: 3.59/4.00 (Top **10**)
Relevant Courses: Algorithm design and analysis(94), Probability & Mathematical Statistics(94), Discrete Mathematics(95), Data Structure(92), Database System(90), Java Programming(93), Advanced Language Program Design(97), Advanced Topics of Information Technology(98)

Research Experience

Research Intern, Supervisor: Dr. *Mor Geva Pipek from Google Deepmind* Feb 2024 - Present
Tel Aviv University
- Conduct the research on the area of Machine Unlearning in Large Language Model
- Propose a new Benchmark for evaluating LLM's Unlearning problems
- Perform the experiments and analysis toward top-tier NLP conferences

Research Intern, Supervisor: Dr. *Haiqin Yang* Dec 2023 - Present
DataStory AI lab, International Digital Economy Academy (IDEA)
- Conduct the research on the area of Hallucination Detection in Large Language Model
- Perform the experiments and analysis toward top-tier NLP conferences

Research Intern, Supervisor: Dr. *Aldo Lipani* June 2023 - Dec 2023
Web Intelligence Group & Space Time Lab, UCL
- Conduct the NLP research on the Knowledge Editing of Large Language Model
- Perform the experiments and analysis toward top-tier NLP conferences
- Submit a paper to ACL Rolling Review on December 2023 as first author

Research Intern, Supervisor: Dr. *Ziqian Zeng* June 2022 - Aug 2023
School of Computer Science and Engineering, SCUT
- Conduct the NLP research on Pretrained Language Model and Information Extraction
- Perform the experiments and analysis toward top-tier NLP conferences
- Publish a paper that has been accepted by AAAI2024 as co-first author

Publications

Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces June 2024
Yihuai Hong, Yu Lei, Shauli Ravfogel, Haiqin Yang, Mor Geva
Under review for NeurIPS, 2024

Dissecting Fine-Tuning Unlearning in Large Language Models June 2024
Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, Hainqin Yang
Under review for EMNLP, 2024 (During Internship in IDEA)

Interpretability-based Tailored Knowledge Editing in Transformers Dec 2023
Yihuai Hong, Aldo Lipani
Under review for EMNLP, 2024 (During Internship in UCL)

ConsistentEE: A Consistent and Hardness-Guided Early Exiting Method for Accelerating Language Models Inference Aug 2023

Ziqian Zeng*, **Yihuai Hong***, Huiping Zhuang, Cen Chen, HongLiang Dai

The 38th Annual AAAI Conference on Artificial Intelligence, 2023

Honors and Awards

AAAI 2024 Student Scholarship Dec 2023

Top Ten Excellent Students Nomination Award of South China University of Technology Nov 2023

China National Scholarship Sept 2023

Meritorious Winner of The Mathematical Contest in Modeling (MCM) May 2023

Kaggle **Silver** medal (Top 5%) July 2021

- CommonLit Readability Prize： Rate the complexity of literary passages for grades 3-12 classroom use Kaggle

Kaggle Bronze medal (Top 6%) Mar 2022

- Evaluating Student Writing: Analyze argumentative writing elements from students grades 6-12

SCUT First Prize Scholarship Oct 2022

Projects

Explanation-Guided Large Language Model Unlearning. June 2023 – Present
(Aiming for NeurIPS 2024)

- During the training process of large language models, insufficient data cleansing often results in the model encountering and inadvertently memorizing private or harmful data. However, the substantial costs associated with retraining the model necessitate consideration of lightweight methods for purging this knowledge from the model.
- From the perspective of interpretability within language models, we analyze the patterns of the corresponding components' functions, attempting to accurately erase the knowledge contained within these components. This approach aims to achieve a more efficient effect of knowledge unlearning.

Knowledge Editing of Large Language Model via Data Augmentation. June 2023 – Dec 2023
(Aiming for EMNLP 2024)

- Large language models can store and predict factual statements about the world. However, as time goes by there will be some outdated knowledge or errors, and retraining the language models to correct these errors demands a great time and resources. So the challenge lies in how to correct and edit the models without too much extra training.
- We gained inspiration from the work ROME and MEMIT, and designed a more effective method to edit a certain knowledge in the language models' parameters from several perspectives using data augmentation. We also perform a gradient-based analysis to find out the layers responsible for different knowledge to execute a smarter edition.

Dynamic Early-exit based on Consistent and Hardness-Guided Training strategy. Dec 2022 – Aug 2023
(Aiming for AAAI 2024, **Accepted**)

- Dynamic Early-exit, allowing samples to exit earlier without passing through the entire model, is an effective method to speed up inference of PLM. The challenges lie in enhancing performance and effectiveness simultaneously.
- We propose an early exiting method that can achieve consistency during training and inference by formulating the early exiting problem as a reinforcement learning problem. The experimental results show that our method can out-perform other baselines on natural language understanding and generation tasks.

Patent

Self-supervised pre-training method, system and medium for Chinese Pinyin spelling correction. Sept 2022

IP No: 202211156374.3

SKILLS

Programming Languages: Python, C++, C, Matlab, Java, Latex
Framework & Tools: PyTorch, TensorFlow, Django
Languages: English, Chinese (Native)