Wei Shen

Age: 24

Apply for a position: PhD application

China, Shanghai, Yangpu District, 200441

+86-18607404976(WeChat)

wshen21@m.fudan.edu.cn

weyshioncn@gmail.com

GitHub WebPage

Homepage

Education

Fudan University *2021-2024*

Computer Science, Master's degree, Fudan NLP Lab, Advisor: Xuanjing Huang CGPA/Percentage: 3.51/4.0

Huazhong University of Science and Technology *2016-2020*

Computer Science and Technology, Bachelor's degree CGPA/Percentage: 3.27/4.0

Publication

Mitigating Length Bias in Reinforcement Learning from Human Feedback (EMNLP 2023 Findings)

Wei Shen* *et al.*

In this paper, we propose an innovative solution, applying the Product-of-Experts framework to separate reward modeling from the negative influence of sequence length.

Improving Generalization of Alignment with Human Preferences through Group Invariant Learning (arxiv)

Rui Zheng,* ***Wei Shen**** *et al.*

The paper introduces a unified framework that ensures distributionally robust alignment by dynamically adapting to various data groups, enhancing the model's ability to handle different data distribution. (This paper is submitted to ICLR 2024)

Secrets of RLHF in Large Language Models Part I: PPO (Instruction Workshop @ NeurIPS 2023)

Rui Zheng, Shihan Dou*, Songyang Gao*,* ***Wei Shen*** *et al.*

RLHF training has challenges and limitations, including PPO This paper studies how different factors in PPO affect policy training. In particular, it identifies policy constraints as one of the major components that affect the performance of PPO. With this observation, we proposed PPO-max, which stabilized the PPO training.

Experience

ByteDance AI Lab - Responsible AI team - Research Internship *2023.8 - present*

- **Supervisors:** Yang Liu and Xiaoying Zhang
- **RLHF with imperfect reward model**: RLHF with noise reward (a paper is ongoing)

Fudan CISL Lab - Research Intership *2021.9 - 2022.3*

- **Supervisor:** Li Shang
- **ML in agile EDA:** chip placement and routing with graph neural network

Erricsson - Software Engineer Internship *2021.6 - 2021.8*

UCLA - ML course in Python, Exchange Student (A) *2018.9 - 2018.10*

Research Interest

LLM Alignment

My current research interest primarily revolves around LLM Alignment, with a specific focus on Reinforcement Learning from Human Feedback (RLHF). RLHF, despite being time-consuming and costly, possesses immense potential for AI alignment. Furthermore, the RLHF pipeline can be scalable and facilitate continual improvement with the assistance of human oversight. RLHF has promising applications across various domains, such as tool learning, task planning, and even debates.

Reward Modeling

Reward model is a crucial module for alignment in both SFT and RLHF, showcasing its potential for the traditional learning paradigm by providing robust and informative preferences to select optimal trajectories or sentences from multiple models' or human responses. However, the generalization of reward models heavily relies on curated pairwise training data and the presence of vague scalar values can lead to issues such as reward hacking. Therefore, I am committed to exploring a more robust and scalable reward model and reward function design to address these challenges.

Project

MOSS-RLHF *2023.5 - 2023.7*

- **Project Overview:** An open-source RLHF project is to facilitate seamless alignment for LLMs, enabling practitioners to achieve alignment more effortlessly.
- **Project Link:** https://github.com/OpenLMLab/MOSS-RLHF (800+ stars)
- **Content:** Contribute to the complete implementation of RLHF, encompassing reward model training and subsequent reinforcement learning processes. Primarily responsible for implementing the English branch and exploring the effects of various techniques for PPO.

TextFlint *2021.10 - 2022.1*

- **Project Overview:** A comprehensive multilingual robustness evaluation platform for natural language processing that offers diverse functionalities such as text transformation, filtering, adversarial attack, and more, enabling a thorough analysis of model robustness.
- **Project Link:** https://www.textflint.io/textflint (600+ stars)
- **Content:** Conducted performance evaluations on multiple open-source Chinese pre-training models and NLU models, yielding comprehensive results.

Skills

Programming Framework: Pytorch **Programming Languages**: Python, C++, LaTeX

Natural Language: Mandarin(native) and English

Interests

Sports: ping pong, badminton and basketball. **Hobbies**: photography, eSports, ukulele

Achievement

Fudan University Computer Science and Technology Academic Scholarship *2022, 2023*

Huazhong University of Science and Technology Self-improvement Scholarship *2019*