HETONG WANG

+44 7715097314 ⋄ hetong.wang809@gmail.com ⋄ https://erikaawang.github.io/

EDUCATION

University of Edinburgh, UK *Sep 2022 – Nov 2023*
MSc in Artificial Intelligence GPA: 3.9/4.0

- *Relavant courses:* Machine Learning and Pattern Recognition (Theory&Practical), Natural Language Processing (Foundation&Advanced), Probabilistic Modelling and Reasoning

University of Liverpool, UK *Sep 2018 – Jul 2022*
BSc (Hons) in Computer Science GPA: 4.0/4.0 (**top 1%**)

- *Relavant courses:* Artificial Intelligence(Foundation&Advanced), Complexity of Algorithms, Calculus, Linear Algebra
- Exchanged at Xi'an Jiaotong-Liverpool University, China

PUBLICATIONS

[1] Probing the Emergence of Cross-lingual Alignment during LLM Training, [Code]
Hetong Wang, Pasquale Minervini, Edoardo Ponti, Under Review at **ACL 2024**.
[2] Accelerating Laboratory Automation Through Robot Skill Learning For Sample Scraping, arXiv: 2209.14875 Under Review at **CASE 2024**.
Gabriella Pizzuto, **Hetong Wang**, Hatem Fakhruldeen, Bei Peng, Kevin S. Luck, Andrew I. Cooper

ACADEMIC EXPERIENCE

THUNLP, Tsinghua University
Research Intern in Open Lab for Big Model Base(OpenBMB) *Nov 2023 - Present*
work on utilizing LLM-based multi-agent systems in AI alignment

Leverhulme Research Centre, University of Liverpool
Research Intern in Robotics and Chemistry Automation Group *Jun 2022 - Sep 2022*
work on utilizing deep reinforcement learning algorithms in lab automation

RESEARCH PROJECTS

Alignment with Multi-agent Systems towards Evolving Social Norms (Ongoing)

- This project aims at developing a data-efficient alignment method that could utilize feedback from multi-model reward proxy and supplementary data from tools using
- Experimented both intra-critic and inter-critic rewarding schemes: explored the feedback from self-rewarding and utilizing reward from models with different scales, and the response sampled with different temperatures

Understanding the Cross-lingual Alignment in Large Language Models(LLMs) - [1]

- This project aimed at understanding the emergence of cross-lingual alignment throughout LLM pretraining, and how its zero-shot ability correlates accordingly
- Hypothesized that the zero-shot transfer capabilities are predicated on the ability of language models to implicitly align different domains even without parallel data, e.g. different languages will activate the same sub-networks of a multilingual LLM during inference

- Applied Intrinsic Probing technique to pinpoint the set of neurons that is most informative to a specific morphosyntactic feature. Inspected the trend of alignment by comparing the overlap of neuron sets of different languages in BLOOM throughout pretraining
- Measured the zero-shot cross-lingual transfer capability of BLOOM concurrently by modifying the XTREME benchmark. Observed a great correlation between the neuron overlap rate and the zero-shot transfer performance throughout training, which confirms our hypothesis

Sparse Fine-tuning Towards Multi-tasked Instruction Following - Report

- This project aimed at developing an instruct-tuning method to improve task-specific capabilities of general-purpose LLMs while avoiding competitive or degrading performance on other tasks
- Based on the modularity of Deep Neural Network(DNN): the architecture of DNN could be disentangled into identifiable modules that correspond to specific task abilities, e.g. datasets of the same task from different sources will modify the same subset of parameters while fine-tuning, which we assumed to be the task-sensitive sub-network
- Proposed a parameter-efficient Sparse FineTuning(SFT) method that involves two stages: *Parameter Selection Stage* – selecting the intersection of parameter sets that met the largest modification while fine-tuning on a specific task; *SFT Stage* – masked fine-tuning on the selected parameters with the corresponding task data
- Evaluated our method by comparing with full FT and random-masked FT, which showed that our method could avoid the catastrophic forgetting problem in full FT, and improve zero-shot performance compared with random-masked FT(an improvement of 7.4 in ROUGE score on unseen data).

Reinforcement Learning for Robot Laboratory Skills - [2]

- This project aimed at researching and developing a robot learning framework to accelerate lab automation using deep reinforcement learning for the acquisition of new contact-rich skills
- Modified panda-gym, a simulation environment for robotic learning based on the Physical engine Pybullet and OpenAI gym to fit our laboratory environment.
- Experimented with different deep reinforcement learning algorithms(DDPG, SAC, TQC+HER) and visualised their performances to improve our reward function.
- Applied the learning framework of the simulation to the real robot task. Designed a whole experiment and evaluated the trained model by its generalization performance.

HONORS & AWARDS

British Computer Society(BCS) Prize *2022, University of Liverpool*
Department of Computer Science, top 1 student with excellent academic performance

University Academic Excellence Award *2020, University of Liverpool*
£5,400 each academic year, top 1% students across the University departments

SKILLS

Research Skills

- Deep Learning Toolkits: Huggingface, PyTorch, Numpy, Keras, Scikit-learn, OpenAI gym
- HPC Cluster: Slurm, SGE(Sun Grid Engine), Shell, Linux
- Tools: LaTeX, HTML

Programming Languages Python, Java, R, C/C++, SQL