HETONG WANG

hetong.wang809@gmail.com ◇ erikaawang.github.io

EDUCATION

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

Supervisor: Edoardo Ponti and Pasquale Minervini
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%**)

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

PUBLICATIONS

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

ACADEMIC EXPERIENCE

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

Leverhulme Research Centre, University of Liverpool
Research Intern in Robotics and Automation Group *Jun 2022 - Sep 2022*
Supervisor: Gabriella Pizzuto and Andy Cooper
work on utilizing deep reinforcement learning algorithms in lab automation

RESEARCH PROJECTS

Utilizing Multi Base Model Information in AI Alignment (Ongoing)

- This project aims at developing an alignment approach that could utilize multi-model settings in RLHF reward modelling.
- Experimented both intra-critic and inter-critic rewarding schemes: explored the feedback from self-rewarding and utilizing reward from different base models.

Exploring the Interpretability of Cross-lingual Generalisation in LLMs - [1]

- This project aimed at understanding the pre-training dynamics of cross-lingual alignment on shared neurons, and how this trend differs over model scales.
- Claimed that the zero-shot transfer capabilities are predicated on the ability of LMs to implicitly align different domain without parallel data, e.g. different languages will activate the same sub-networks of a multilingual LLM during inference.

- Applied Intrinsic Probing technique to identify the set of language-specific neurons that is most informative to some linguistic features. Inspected the trend of alignment by comparing the overlap rate of neuron sets through languages in a sequence of BLOOM checkpoint models.
- Measured the single-source transfer ability of checkpoint models concurrently. Observed a great correlation between the structural overlap and the generalisation performance throughout pre-training to support our claim.

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

- This project aimed at developing an instruct-tuning approach to improve task-specific capabilities of general-purpose LLMs while avoiding competitive or degrading performance on other tasks.
- Based on the assumption of deep network modularity: Neuron models could be disentangled into modules that correspond to different task abilities, e.g. datasets of the same task but different sources will modify the same subset of parameters while fine-tuning, which we assumed to be the task-specific 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