HETONG WANG

+44 7715097314 ⋄ s2308470@ed.ac.uk ⋄ 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 (top 1%)

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

RESEARCH PROJECTS

Understanding the dynamics of implicit alignment in Large Language Models(LLMs) [1]
Master Thesis in the University of Edinburgh, Supervisor: Edoardo Ponti. *May 2023 - Present*

- 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, a variational Bayesian model to pinpoint the set of neurons that is most informative to a specific morphosyntactic feature. Inspected the trend of alignment emergence 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

Task-oriented Sparse Finetuning for Instruction-tuned Language Model - Report
Research Course Project in the University of Edinburgh. *Jan 2023 - May 2023*

- 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 on instruction-tuned language model, which 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 normal fine-tuning and random-masked SFT, result showed that our method could avoid the catastrophic forgetting problem in traditional fine-tuning, and improve zero-shot performance compared with naive SFT(an improvement of 7.4 in ROUGE score on unseen data).

Reinforcement Learning for Robot Laboratory Skills [2] *Jun 2022 – Sep 2022*
Research Intern in Leverhulme Research Centre, Supervisor: Gabriella Pizzuto.

- 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.

Reinforcement Learning: AI helper in Yahtzee Game *Sep 2021 – May 2022*
Bachelor Thesis in the University of Liverpool, Supervisor: Michele Zito.

- This project aimed at developing a software embedded with an AI helper, which can analyse the current game state and provide action suggestions to users.
- Gained a global understanding of reinforcement learning principles, such as the Markov decision process, Monte Carlo Method. Familiar with commonly used algorithms and their principles, such as Q-learning, Sarsa, Deep Q Network.
- Developed a Yahtzee reinforcement learning environment following OpenAI gym interface, built a Deep Q Network model with Python library TensorFlow. Handled data transferring and processing, also designed a simplified state representing scheme to reduce computational cost.

PUBLICATION

[1] Probing the Dynamics of Cross-lingual Alignment throughout Training in Multilingual Language Models
Hetong Wang, Edoardo Ponti, To Be Submitted to ACL 2024.
[2] Accelerating Laboratory Automation Through Robot Skill Learning For Sample Scraping, arXiv: 2209.14875
Gabriella Pizzuto, **Hetong Wang**, Hatem Fakhruldeen, Bei Peng, Kevin S. Luck, Andrew I. Cooper

RESEARCH 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 algorithm in lab automation

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
- Paper Writing: LaTex, Academic English Writing and Presenting
- HPC Cluster: Slurm, SGE(Sun Grid Engine), Shell, Linux

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