Format

Exploring AI Safety via Deep Reinforcement Learning Applications on Games and Scientific Applications

Supervisor
Wei, Ting Han, Ph.D
Professor, Informatics

1. Objective

This project is aimed at:

My primary research experience is in the field of game AI, with a specific focus on the integration of deep reinforcement learning (RL) with heuristic search. These methods have found particular success in games because games are well-defined environments that are much simpler than real world problems, while retaining interesting properties such as complex situations, hidden information, and stochasticity. RL is also closely related to the topic of AI safety. For example, reinforcement learning with human feedback (RLHF) is commonly used to fine-tune large language models so that its behaviour is aligned with human interests. This research project aims to reproduce and investigate successful RL applications in games and scientific problems while proposing methods to avoid or eliminate unintended agent behaviour.

2. Research Theme Outline

To that end, the research will consist of the following phases:

(a) Investigation of various games and benchmarks that may pose interesting research questions related to unintended agent behaviour and alignment, e.g. reward hacking, explainability, etc.
(b) Reproducing notable published results from other research teams to observe examples of these unintended behaviour.
(c) Propose methods to align agent behaviour with desirable actions.

3. Expected Performance

The successful candidate would be expected to:

(a) Survey and present up-to-date research papers in the field of reinforcement learning, deep reinforcement learning, game AI research, and AI safety/alignment.
(b) Learn and understand common reinforcement learning algorithms, with a focus on model-based methods such as AlphaZero and MuZero, and optionally, model-free methods such as DQN, A3C, PPO, etc.
(c) Apply the above algorithms to new environments (including but not limited to games)
(d) Interpret the application's impact on AI safety and alignment.
(e) Propose and test methods that eliminate undesirable agent behaviour.
(f) Publish results and present them at international top conferences such as NeurIPS, ICML, ICLR, AAAI, or IJCAI.

4. Required Skills and Knowledge

The successful candidate will have the following knowledge and skills:

(a) Background in Machine Learning: We use deep reinforcement learning in many of the projects in our lab. Experience in fundamental machine learning concepts and algorithms are therefore necessary. Experience in deep learning will be helpful.
(b) Reinforcement Learning: Familiarity in topics including the following is preferred, but not necessary. Monte

Carlo Tree Search, the AlphaZero and MuZero algorithms, Q-learning, and reward shaping. Experience in deep reinforcement learning is highly desirable (e.g. DQN, A3C, PPO, etc.)
(c) Programming Skills: Proficiency in Python (especially PyTorch) is preferred. Experience in other languages, especially C++, is a strong plus.
(d) Mathematical Proficiency: A strong understanding of linear algebra, probability, and statistics is highly desirable.
(e) Research and Language Skills: Proficiency in English is required. You must be able to pose research questions, design and conduct experiments, and interpret and analyze experiment results autonomously. Past experience in academic writing is a plus, but not necessary.
(f) Teamwork and Collaboration: We may collaborate with research teams in Taiwan and Canada, so you must be able to communicate clearly and work as part of a team.

Contact

E-mail: tinghan.wei@kochi-tech.ac.jp