Expert Iteration in OXO using UCT as an Expert Agent

Robert Marc James-Stroud
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
Email: rj18801@essex.ac.uk

Abstract

I. Introduction

General Game Playing (GGP) is a field of research in AI involving programs that play games such as chess and Go [1]. Historically GGP focuses on two-player zero sum games with finite states. Programs developed for playing games used to be specialised (Deep Blue), only able to play that one game often with handcrafted rules.

Programs such as Deep Blue have little value in AI research, they are able to evaluate a state for one game only before making a decision about the moves [1]. However more recently researchers have begun developing AI agents that are able to take a ruleset and state at runtime and play games to successful completion. Genesereth describes general game players as:

> '*Systems able to accept descriptions of arbitrary games at runtime and able to use such descriptions to play those games effectively without human intervention*' [1]

As general game playing agents are unable to predetermine any policies about the game it is going to be playing, it is essential that a classifier is determined quickly, an algorithm that has seen great success in this area is Monte Carlo Tree Search (MCTS). Neural networks on the other hand are not trained quickly, they require lots of training data until they can output a result with high success.

Humans approaching a game for the first time use two kinds of thinking - dual process theory. The first aspect of this system is a fast thinking, intuition. The second aspect is slow thinking, reasoning [2]. Exploiting dual processes theory for should result in strong agents independently trained from human influence removing any human play bias from the agents learning.

Reinforcement learning (RL) agents take actions without looking ahead and tree searches, to determine the best action must evaluate branches, with even simple games like OXO having a high branching factor. Expert Iteration uses MCTS to train a neural network, which in turn guides the tree search [2].

II. Background

Supervised learning has been common place in AI, using datasets provided by human experts which agents will then try to replicate. Having humans involved in deployment of agents, according to Hui is troublesome [3]. Therefore it is advisable to remove the human bias from training in all aspects. The benefit of this is two-fold, not only is the human removed from the agent, the agent will also be able to discover non-human approaches which might be of more strategic value than imitating and trying to beat a human based agent. Another benefit to using non-human expert datasets is that expert human datasets are 'often expensive, unreliable or simply unavailable' [4].

If the human datasets are available. and considered reliable; using them provides a ceiling on the performance of an agent. RL removes this by being trained on their own experience, allowing them to surpass human expertise [4].

Expert Iteration mimics the human learning process dual-process theory. Dual-process theory has two processes an automatic implicit happening at the subconscious level, and an explicit conscious logical and reasoning process. Humans when first encountering a new game exploit both processes [5].

Both [2] and [4] use a very similar RL process to training their agents.

III. Methodology

IV. Experiments

V. Discussion

VI. Conclusion

References

[1] M. Genesereth, "Overview of general game playing," 2005.
[2] T. Anthony, Z. Tian, and D. Barber, "Thinking fast and slow with deep learning and tree search," *CoRR*, vol. abs/1705.08439, 2017. [Online]. Available: http://arxiv.org/abs/1705.08439
[3] J. Hui.
[4] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.
[5] J. S. B. Evans, "Heuristic and analytic processes in reasoning," *British Journal of Psychology*, vol. 75, no. 4, pp. 451–468, 1984.