research_review

Table of Contents

1 Review of "Mastering the game of Go with deep neural networks and tree search"

1.1 Introduction

This paper dewscribed the algorithms and strategies used to build AlphaGo compouter program which can beat humans and other AI programs. Exhaustive search for optimal move is infeasible in Go. (150 to power 250).

1.2 Techniques used

AlphaGo uses Monte Carlo tree search (MCTS) which needs a policy to give values to different moves. Instead of using the policies which predicts the human experts move as in current systems, AlphaGo utilizes convolution neural network. These neural networks were on on a pipeline of ML stages. First a supervised learning (SL) policy was trained on human moves. Then a reinforcement learning policy network improves the previous stage from self play. SL policy network was 13 layer trained on 29.4 million positions from 160,000. This gave the accuracy of 57% as against the previous best of 44.4%. Larger networks gave better performance but at the cost of higher time in evaluation. A smalller network was also trained which achieved 24.2% in 2us against 3ms. Symmetries of Go wasn't exploited as it hurt the performance of large neural networks. Instead, it is used at run time. RL policy is similar in structure to SL policy and was initialised with same weight. Games were played between this network and a randomly selected previous iteration of policy. Then weight were updated in the direction using stochastic gradients which maximises theexpected outcomes. This network won 80% of the games head to head against SL policy network. Final stage in pipeline is reinforcement learning of value network. In this, prediction of outcome is given using a policy for a given state of boards. This network is trained on state-outcome pairs using stochastic gradient descent to minimise the mean square error. To mitigate the overfitting, an additional 30 million data of self play was used. This approached the accuracy of RL policy network but using 1500 times less computation. Next policy and value networks are combined with MCTS.In AlphaGo, SL policy network outperformed RL policy network but value network trained from RL policy was more accurate. Evaluating policy and value networks require lots of computations than traditional search heuristics. Final version of AlphaGo uses 40 search threads, 48 CPUs, and 8 Gpus in asynchronous multi-threaded which executes simulation in CPUS and networks in GPUs, The distributed AlphaGo uses 40 search threads, 1,202 CPUs and 176 Gpus.

1.3 Conclusion

- AlphaGo 494 out of 495 games (99.8%) against other Go AI programs.
- Asynchronous single machine AlphaGo won 77%, 86%, and 99% of handicap games against Crazy Stone, Zen and Pachi with 4 free moves.
- Distributed AlphaGo won 100% of it's games and 77% against the single machine AlphaGo.
- Performance of AlphaGo with only value networks and only rollouts was also tested.
- Only value networks won against all other Go programs.
- Mixed evaluation of value networks and rollouts performed the best.
- AlphaGo beat European Champion Fan Hui 5-0.
- During the match against Fan Hui, AlphaGO searched thousands of times fewer positions than the Deep Blue against Kasparov.
- Also evaluation functions in AlphaGo aren't hardcrafted like in DeepBlue but trained through supervised learning and reinforcement learning.

Author: Khurram Beigh
Created: 2017-02-20 Mon 02:35
Emacs 25.1.1 (Org mode 8.2.10)
Validate