Research Review of AlphaGo: Mastering the game of Go with deep neural networks and tree search

Url to paper.

The “Google DeepMind Challenge Match” was a Go match between `AlphaGo`, a computer program capable of playing Go and 18-time Go world champion, Lee Sedol. It was a five game series, in which AlphaGo won all but the fourth game.

This paper was published by the DeepMind team after AlphaGo became the first AI Computer Go program to beat a human professional Go player without handicaps on a full-sized 19×19 board. This feat was previously thought to be at least a decade away.

Issues faced

Go has approximately 250 legal moves per position and a average game length of 80 moves. It has been estimated that over 10^761 games of Go are possible, compared to approximately 10^120 games for chess. Because of its large branching factor traditional AI methods such as alpha–beta pruning, tree traversal and heuristic search become unfeasible.

AlphaGo Approach

1. A supervised learning approach is taken to train a 13-layer convolutional neural network, to predict expert moves using 30 million positions from the KGS Go Server. The network outputs the probability of each legal move given an input of a 19x19 board state. This provides immediate feedback, high-quality gradients, and is used for improving board evaluation.

2. Next, a Reinforcement Learning policy network is trained to optimize the final outcomes of the SL stage via “policy gradient reinforcement learning”. This new policy has the correct goal of winning the games, rather than maximizing the predictive accuracy of each move. RL policy network won more than 85% of games against Pachi - the strongest open-source Go program with a sophisticated Monte Carlo search algorithm.

3. The final stage focuses on estimating a value function, which can be used to predict the outcome from a game state using a policy for both players. This function is similar to the heuristic functions, but are not 100% accurate and are estimated via reinforcement learning. This function was trained via regression, and when trained on a dataset generated by playing against itself, the network acheived accuracies of 0.226 and 0.234 on the training and test set respectively, indicating minimal overfitting.

4. Monte Carlo Tree Search is used to run many game simulations guided by the policy and value networks.

Results

In the end the Supervised Learning policy network performed better than the stronger Reinforcement Learning policy network. However the value function derived from the RL policy evaluated the game state better. The most commonly used system for comparing the strength of players is the Elo rating system. AlphaGo’s Elo rating is estimated at 2890. Furthermore the distributed version of AlphaGo, which runs on 1202 CPUs and 176 GPUs is estimated to be 3586. The only human with a higher Rlo rating is Ke Jie, at 3621.