Research Review

Adam Liu

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

In the long history of human evolution, Go has been a game that has the most challenging reputation, and it has been challenging to computers that used to solve it due to the enormous possibilities evaluating board positions and moves (the board is a 19 x 19 grid board). In this paper the authors introduced a new approach that uses ‘value networks’ to evaluate board positions and ‘policy networks’ to select moves, which are trained in combination of supervised learning from human expert games, and reinforcement learning from games of self-play.

1. Supervised learning of policy networks

    In this step, the authors used supervised learning to train the neural network based on 160,000 games played by KGS 6 to 9 dan human experts including 35.4% handicap games in the total dataset. Some of the benefits of this approach is that it provides fast and efficient learning progress. In the whole training process of the AlphaGo program, this network is improved by the RL policy network.

2. Reinforcement learning of policy networks

    The most important benefits of policy network is it helps to understand which move gives the best overall game state.

    With the reinforcement learning network, the self-play outcomes of the current state games are efficiently evaluated. The efficiency of this network is outstanding comparing to the SL policy network. When used to play against the SL policy network, this network won more than 80% of the games.

3. Reinforcement learning of value networks

    The reinforcement learning value networks are the evaluation method that evaluates the overall game state after playing stone in coordinate X, predict the winner of the games played. In this approach, the RL policy network was playing each game with itself until the game terminates, which leads to MSEs of 0.226 and 0.234, indicating minimal overfitting. The policy networks & value networks are combined with with MCTS (MCTS is the Monte Carlo Tree Search which uses Monte Carlo rollouts to estimate the value of each game tree state, a very important element of AlphaGo). With the implementation of MCTS, the estimated winning probabilities would become more accurate along with the growth of the tree size.

4. Searching with policy and value networks

In AlphaGo, the policy networks & value networks are combined by the MCTS algorithm, which looks ahead and selects actions.

5. Evaluating the playing strength of AlphaGo

AlphaGo resulted as many dan ranks stronger than any opponent Go program, winning 494 out of 495 games (99.8%) against other Go programs in fair games. In handicap stones that allow opponents to take free moves, AlphaGo won 77% against Crazy Stone, 86% against Zen, and 99% against Pachi.

Conclusion

By utilising both policy networks and value networks, AlphaGo obtains the best move according to the current game state, by achieving Elo score of 2890 and 3140 (distributed version). AlphaGo demonstrated the possibilities deep learning brings, both in efficient evaluation and optimised decision making.

As a quote famously spreaded: Artificial Intelligence is the search of the most optimised approaches, the potential of deep learning is definitely going to be discovered more through the time.