Article: Mastering the game of GO with deep neural networks and tree search

Summary: The paper proposed an innovative method to build an intelligent Go agent that would achieve superhuman performance,defeating the best human player. The issue of enormous search space and difficulty in evaluating board positions and moves were efficiently solved with deep neural networks being trained by a combination of supervised learning from human expert games and reinforcement learning from games of self-play.

Frameworks: A novel approach for reducing the effective depth and breath of a search tree based on applying neural networks in a pipeline manner consisting of several stages of machine learning is presented and it consists of following stages:

1. Supervised learning for policy networks: policy network was used to provide a probability distribution over all legal moves and thus, to predict expert human moves. A 13-layer policy network was trained from a dataset of 30 million positions and the trained network could predict the expert move at an accuracy of 57% if all input features are used and 55.7% if only raw board position and move history are used as inputs. The result outperformed the current state-of-the-art accuracy of 44%.

2. Improvement of policy network by policy gradient reinforcement learning: games were performed between the current policy network and a randomly selected previous iteration of the policy network and the weights within the network were updated based on whether the current policy wins or losses. Then, a tournament between the current policy (improved, RL policy) and those from before this reinforcement learning (SL policy) was run in order to evaluate the improvement. The result is that the RL policy network won more than 80% of the time. Furthermore, when played with the strongest open-source Go program (Pachi), the RL policy network won 85% of games.

3. reinforcement learning for value networks: the final stage of the training pipeline focused on position evaluation with value networks. The value function wsa approximated using value network which is based on RL policy network and indeed, has similar architecture to the RL policy network. The only difference was that the output of value network is a single prediction instead of a probability distribution as in policy network. To mitigate the issue of overfitting caused by strongly correlated, new self-play dataset consisting of 30 million distinct positions sampled from separate games were used. The result showed that the position evaluation accuracy of the value network is consistently more accurate than Monte Carlo rollout using the fast network and especially, the computation time is 15000 times lower.

Results: To quantitatively evaluate AlphaGo, the tournament against the state-of-the-art commercial and open-source programs was performed. The timeout per move was 5s. The results were ground-breaking as AlphaGo achieved 99.8% winning rate (494 winning out of 495 match) against other Go program and in the case of more challenging games with four handicap stones, AlphaGo won 77% up to 99%. Finally, AlphaGo was compared to Fan Hui player, a championship human player and AlphaGo set the world record with the result of 5 games to 0.