Research review - AlphaGo

JuHyung Son

Go is one of the most complicating game in the world. It has 19*19=384 boxes on the board. Because of the high complexity, the previous game AI, that is used in chess, othello, checker can not be used. But now, Alpha Go is the top Go player in the world (recently Alpha Go zero came out). In this paper, they presented an approach that uses ‘value networks’ to evaluate game board position and ‘policy networks’ to select moves. Also, there is a search algorithm that combines Monte Carlo simulation with value and policy networks.
Monte Carlo tree search uses Monte Carlo rollouts to estimate the value of each state in a search tree. With large number of simulations, the policy used to select actions during search is improved over time. Asymptotically, this policy converges to optima play.
The Alpha Go team employ CNN for the Go for reducing the effective depth and breadth of the search tree. They train the neural networks using a pipeline consisting of several stages of machine learning. Beginning by train a supervised learning policy network from expert human. Here the SL policy network alternates between convolution layers with weights, rectifier nonlinearities, and then softmax layer outputs a probability dist. over all legal moves. Next, they train a reinforcement learning policy network that improves the SL policy network by optimizing the final outcome of games of self play. The RL policy network is identical in structure to the SL policy network. They play games between the current policy network and a randomly selected previous iteration of the policy network. Also, there is a reward function r(s) that is zero for all non-terminal time steps. The outcome is the terminal reward at the end of the game from the perspective of the current player at time step t: +1 for winning and -1 for losing. Weights are updated at each time step by stochastic gradient ascent. Finally, they train a value network that predicts the winner games played by RL policy network against itself. Actually they would like to know the optimal value function under perfect play, but it is impossible, so instead, they estimate the value function for strongest policy, using the RL policy network. Alphago team uses the value function using a value network with weights theta. This neural network has a similar architecture to the policy network, but outputs a single prediction. AlphaGo team combines the policy and value network in an MCTS algorithm that selects actions by lookahead search. After training, alphago plays GO to evaluate playing strength with alphago and several other go programs, all of them are based on high-performance MCTS algorithms. The results of the tournament suggest that single-machine AlphaGo is many dan ranks stronger than any previous Go program winning 99.8%. The distributed version of AlphaGo was significantly stronger, winning 77% of games against single-machine AlphaGo and 100% of its games against other programs. Finally, alphago team evaluated the distributed version of AlphaGo against Fan Hui, a professional 2 dan, and the winner of the 2013~2015 European Go championships. Alphago and Fan Hui competed in a formal five-game match. AlphaGo won the match 5 games to 0. This is the first time that a computer Go program has defeated a human professional player, without handicap. Finally, alphago is based on a combination of deep neural networks and tree search that makes it plays at the level of the strongest human players.