Research Review on Alpha Go

Written by Yuen Pok Henry. Ho, AI Nano Degree Term 1.

Goals and Techniques introduced

The game of Go is one of the most challenging of classic games in AI with its enormous search space and difficulty of evaluating board positions and moves. AlphaGo tackle this game with a different approach, involved the use of value networks to evaluate board positions and policy networks to select moves. The neural networks are trained through supervised learning (SL) from human experts and Reinforcement Learning from self-play games. Using techniques such as Monte Carlo tree search, the value of each state in the search tree can be estimated, and the policy used to select actions is improved over time with the evaluations converge to the optimal value function. Deep convolutional neural network used in visual domains is also employed in the form of board positions as a 19 x 19 image, arranged as overlapping tiles in layers to represent positions. In turn positions are evaluated using value network then sampling actions using a policy network.

Supervised learning (SL) of policy networks outputs a probability distribution over all legal moves a, where the input s is a simple representation of the board state. The policy network is trained on random sampled state-action pairs using stochastic gradient ascent to maximise the likelihood of the human move a selected in state s.

Reinforcement learning (RL) of policy network is used as the second stage of the training pipeline to improve the policy network, where the structure is identical to the SL policy network. Game is played between the current policy network against randomised previous iteration of policy networks, and a reward function with weights updated at each time step. The performance is then evaluated in game play.

The third and final stage is the Reinforcement learning (RL) of value networks, estimating a value function that predicts the outcome from a position s. We approximate the value function using value network, which again has a similar architecture as the policy network, outputting a single prediction instead. To eliminate overfitting, each gain was played between the RL policy network and itself until the game terminated, with indication of minimum overfitting.

Results

The Supervised learning of policy network was able to predict expert moves on a held out test set with an accuracy of 57% and 55.7% using only raw board position and move history, small improvements in such accuracy led to large improvement in playing strength.

The RL of policy networks was able to achieve more than 80% of winning games against the SL policy network, also winning against the strongest open-source Go program Pachi with a winning rate of 85%.

The RL of value network was able to achieve the accuracy of Monte Carlo rollouts using RL policy network, but using 15,000 times less computation.

The combination of deep neural networks and tree search enabled AlphaGo to play against some of the strongest human player. The well publicised game against Fan Hui, a professional 2 dan human player – a winner of the 2013 – 2015 European Go championships was a success with a 5 – 0 victory. The first time a computer Go program is able to defeat a human professional player without handicap in a full game of Go.