# AlphaGo Research Review

Mu Chen

Paper Title:
Mastering the game of Go with deep neural networks and tree search
by David Silver et al.

Go game has been long believed to be very challenging for computers to play due to its huge search space. The paper introduced a new approach to computer Go that uses value networks to evaluate board positions and policy networks to select moves. The new computer Go passes in the board position as a 19 x 19 image and uses convolutional layers to construct a representation of the position. Then the policy network and value network are trained through several stages of machine learning.

First, they train a 13-layer policy network directly by supervised machine learning from expert human moves of 30 million positions. The network predict expert moves with an accuracy of more than 55%, which has exceeded the best result from previous research groups. They also trained a faster but less accurate (24%) policy that can rapidly sample actions during rollouts. To improving the policy network, reinforcement learning is applied on policy gradient that optimize the final outcome of games of self-play. In this way, the policy network is aimed at achieving the correct goal of winning the game, instead of maximizing predictive accuracy. At this stage, the SL and RL combined network has outperformed any other SL only program. Using no look-ahead search, the RL policy network won 80% of games against the strongest open-source Monte-Carlo search program. The last stage of training is to find a value function that predicts the outcome of the game from current game state. To avoid overfitting, instead of feeding in successive positions from one game, a new self-play data set consisting of 30 million distinct positions, each sampled from a separate game, is used for training. The value network is consistently more accurate than Monte Carlo rollouts using the fast rollout policy. AlphaGo selects actions by look-ahead search using Monte Carlo tree search algorithm. It combines policy and value networks in the search. Since evaluating policy and value networks requires several orders of magnitude more computation than traditional search heuristics, AlphaGo uses a multi-threaded search that executes simulations on CPUs and computes the policy and value networks in parallel on GPUs.

The result of tournament against other programs shows that single-machine AlphaGo is much stronger than any previous Go program. The multi-machine AlphaGo is even stronger, winning 77% of games against single-machine AlphaGo. By assessing variants of AlphaGo that evaluated positions using just value network or just rollouts or a mixture of both, they found that the two position evaluation mechanisms are complementary. The value network is strong but slow while the Monte Carlo rollout policy is fast but weak. A mixture of both gives the best performance.

Through a combination of supervised learning and reinforcement learning, they have developed the effective move selection and position evaluation function for Go. The new search algorithm that combines neural network with Monte Carlo rollouts makes AlphaGo the best Go program so far that can compete against the strongest human players.