Search Review for:
Mastering the game of Go with deep neural networks and tree search

Goals
The goal of the paper is to introduce a new approach to compute Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves, the approach is effective enough to defeat a human professional in the full-size game of Go.

Techniques
Go has been viewed as the most challenging of classic games for artificial intelligence due to search space and difficulty of evaluating board positions and moves. Compare to chess, which has about a breadth of 35 and depth of 80, Go has 250 and 150, respectively. Exhaustive search is infeasible for large games like chess and Go. In order to reduce the search space, AlphaGo combines Monte Carlo tree search (MCTS) with value and policy networks. The team uses a training pipeline as follows:

1. Supervised learning of policy networks
For the first stage of the training pipeline, the team built the supervised learning (SL) policy network is a 13-layer policy network using 30 million positions from the prior work on predicting expert moves using supervised learning. Policy network takes a representation of the board position *s* as its input, passes it through many convolutional layers with parameters, and outputs a probability distribution over legal moves *a*. The SL policy network alternates between convolutional layers and rectifier nonlinearities. A final softmax layer outputs a probability distribution over all legal moves *a*.

2. **Reinforcement learning of policy networks**
The second stage of the pipeline aims at improving the policy network by policy gradient reinforcement learning (RL). The RL policy network is identical in structure to SL policy network, and its weights ϱ are initialized to the same values. The team plays games between the current policy network and a randomly selected previous iteration of the policy network, with a reward function. Weights are updated at each time step by stochastic gradient ascent in the direction that maximizes expected outcome. This resulted an RL policy network that won more than 80% of games against the SL network.

3. **Reinforcement learning of value networks**
The final stage of the training pipeline focuses on position evaluation, estimating a value function that predicts the outcome from position of games played by using policy *p* for both players. The value network has a similar architecture to the policy network, but outputs a single prediction instead of a probability distribution.

With the training pipeline, the team combines policy and value network with MCTS algorithm that selects actions by lookahead search. The leaf note is evaluated by both value network

and by outcome of a random rollout played out until terminal step using the fast rollout policy, combined using a mixing parameter.

Results

With the novel combination of supervised and reinforcement learning, AlphaGo achieved a 99.8% win rate against other Go programs. It also defeated Fan Hui, a 2 dan professional, this is the first time that a computer Go program has defeated a human professional player, without handicap, in the full game of Go, a feat that was previously believed to be at least a decade away.