Google AlphaGo Research Review

Go is a perfect information board game that has long been considered infeasible to solve due to its large search space (Go has approx. breadth = 250 legal moves position and a depth = game length of approx. 150, resulting in approximately 250^150 possible search positions). The AlphaGo team implemented a new novel approach using neural networks to reduce the effective breadth and depth of the search tree – a policy network to select moves and a value network to evaluate board position, achieving a 99.8% winning rate against other Go programs and defeating a human European Go champion, Fan Hui, 5-0.

The AlphaGo team implemented a three-stage training pipeline for the value network and the policy network mentioned above.

The first stage of the pipeline is a supervised learning (SL) policy network. This is a 13-layer network trained using 30 million expert human moves/position from the KGS Go Server. The input to this layer is a simple representation of the board state (consisting of the features and number of planes for the features) and outputs a probability distribution over all legal moves. The network predicted expert moves on a test set with an accuracy of 57%

The second stage of the pipeline is reinforcement learning (RL) policy network. This network is identical in structure to the SL, and its key purpose is to improve the predictions of the previous stage SL policy network. For the purpose of training, games are played between current policy network and randomly selected previous iteration of the policy network. A reward function awards 0 for all non-terminal time-steps, +1 for winning and -1 for losing and the weights are updated in the direction that maximize winning. The RL policy network won more that 80% of its games again the SL policy network, and 85% of games against Pachi (a state of art open-source Go program).

The final stage of the pipeline is a value network that focuses on position evaluation. The purpose of this network is to predict an outcome for the strongest policy from the RL policy network. This network has a similar architecture to the policy networks and is trained using regression on the game state and game outcome tuple from the RL network.

In addition to the 3-stage network AlphaGo also uses a Monte Carlo Tree Search (MCTS) to select actions by lookahead search which choose the moves based on evaluation of each leaf node using the value network and using policy network.

AlphaGo was evaluated against the strongest commercial programs (Crazy Stone and Zen) as well as open-source programs (Pachi and Feugo) which are based on the MTCS algorithm. The results shows AlphaGo winning 99.8% (494 to 495) games against other Go programs.

Finally, AlphaGo was evaluated against Fan Hui a 3-time European Go champion in formal 5 match game resulting in AlphaGo winning 5-0 against its human opponent. During the match against Fan Hui, AlphaGo evaluated 1000 times fewer position compared to IBM Deep Blue, primarily because of selecting more intelligent position using the policy network and evaluating them more precisely using the value network.