AlphaGo

AlphaGo created a novel neural network based game playing architecture for perfect information games. It combines a CNN based deep neural network pipeline with the Monte Carlo Tree Search (MCTS) algorithm to create a new game playing architecture which is both, more powerful and more computationally efficient, than the traditional tree search based algorithms.

AlphaGo Design

AlphaGo uses four CNN based neural networks in its training pipeline.

1. **Supervised learning (SL) Policy Network** is a 13 layer CNN which uses 30 million positions from the Go expert moves dataset to train a model to predict human expert moves by using a softmax classifier that assigns probabilities to possible Go moves.
2. **Rollout Network** is trained along with the SL policy network (in #1. above) which faster than SL but less accurate. It is used for action selection during game tree search.
3. **Reinforcement Learning (RL) Network** is initialized with the SL Policy Network (trained in #1 above) and it improves the policy gradient learning. Training data for this network is generated by playing RL in self-play mode with the older versions of the SL Policy Network. The goal of the RL network is to maximize the outcome of “winning more games” on the leaf nodes of the game tree.
4. **The Value Network** is trained by regression to predict a value to the expected outcome, which is whether the current player wins, in the positions from the self-play data (# 3. above).

The above models are utilized in the MCTS in the following manner

1. **Selection**: The tree is traversed by selecting the edge at a given node with the maximum action value (Q+u(P)), u(P) is the bonus that depends on the prior probability stored for the edge.
2. **Expansion**: A leaf node is expanded based on the prediction from the SL Policy Network, the prediction is stored as prior for each action.
3. **Evaluation**: The leaf node is then evaluated using the Value Network and the Rollout Network to compute the winner
4. **Backup**: Action values Q are updated to track the mean value of all evaluations in the subtree below that action.

AlphaGo Results

The single machine version of AlphaGo was sufficient to beat state of the art MCTS based GnuGo, winning 99.8% of all games. Against the highest ranked human professional player Fan Hui the distributed (more powerful) AlphaGo won a 5 game series 5-0. This was the first time a Go program beat a human professional player with no handicap, a feat previously believed to be

at least a decade away. In addition, AlphaGo evaluated thousands of times fewer positions compared to DeepBlue, the other perfect information game(chess) playing program that beat the highest rated human professional player.