Review of AlphaGo Paper

Goals:

Due to the enormous size of the search space of the game of Go, the DeepMind Team set out to limit the search space using a combination of Deep-Learning Networks and Monte Carlo Search Trees.

Network Design:

DeepMind made four neural nets to decide which moves to play. These were referred to as the Supervised Learning (SL) policy network, the Reinforcement Learning (RL) policy network, the Fast Rollout (FR) policy network, and the Value network.

The SL network predicts the probability of each legal next move. It was structured as a 13-layer deep Convolutional Neural net consisting of alternating layers of convolution and rectifier non-linarites for the first 12 layers followed by a final softmax layer. This network was trained on 30-million game positions from the KGS servers.

The RL network predicts the best move using the same structure as the SL network. Instead of being trained using existing data, this network played against random previous iterations of itself.

The FR network combined the SL network with the Monte Carlo Tree Search as described below. The SL network was chosen because it performed better than the RL network.

The Value network estimates the probability of winning from the current state. Using the KGS dataset this network overfitted, learning outcomes instead of moves. Instead, DeepMind generated a new data set consisting of 30 million distinct positions. Each game was then played between the RL policy network and itself until the game terminated. This significantly reduced the overfitting.

Tree Search:

AlphaGo combines SL, RL, and FR networks with the Value network in a Monte Carlo search algorithm. This algorithm consists of four phases: the Selection Phase, the Expansion Phase, the Evaluation phase, and the Backup Phase.

The Selection Phase chooses the branch from the current state with the highest *Q*, where *Q* is the “score” of that branch.

The Expansion Phase runs the SL network to determine a strong move from the legal moves available.

The Evaluation Phase first runs the Value network to evaluate the position of this branch for the probability of winning. It then uses the FR network to playout the rest of the game from this branch as many times as possible before it hits a time limit.

The Backup Phase propagates the results of the Evaluation Phase adjusting the *Q* value of the branch up or down depending on whether the branch did well or not.

Results:

The results presented indicate that:

- Both versions of AlphaGo presented (non-distributed and distributed) outperform previous AIs and the best (human) European player.
- Just using the Value network AlphaGo manages to stay on par with other AIs.
- The performance of AlphaGo scales as hardware performance is increased.

Source:

https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf