Research review of the AlphaGo paper by DeepMind

Go has been one of the most challenging games for AI due to his huge sarch space (around to `250 ^ 150`). The guys from deep mind introduced a new approach that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. They also introduced a new search algorithm tha combines Monte Carlo simulation with the value and policy networks. They achieved a 99.8% winning rate against other Go programs and also defeated the human European Go champion by 5-0. Experts were sure that this achievement was a at least a decade away.

So, like I said, the biggest issue with go is it's huge search space. The DeepMind guys focused on reducing the effective search space by general principles:

1. Reducing the depth of the search by evaluating the position by truncating the search tree at state `s` and replace it by an approximate value function `v(s)`.
2. Reduce the breadth of the search by sampling actions from a policy `p(a|s)` that is a probability distribution over the possible moves `a` in position `s`.

And they achieved this reduction of the search space using deep neuronal networks for both policy network and value fn.

Training

Supervised learning of policy networks

They trained a 13 layer policy network (SL policy network) with an input `s` being a simple representation of the board state. It was trained with 30 million positions from the KGS Go Server.

Reinforcement learning of policy netwroks

In this stage the goal was to improve the policy network of the previous stage by gradient reinforcement learning (RL policy network). It's structure is identical to the SL policy network and it's weights are initialized to the same values. Then they play games between the current policy network and randomly selected previous iterations of itself.

Reinforcement learning of value networks

This is the final stage of the training pipeline. It focuses on estimating a value function `v(s)` tha predicts the outcome from position `s` of games played by using policy `p` for both players. This neural network has a similar structure to the policy network but outputs a single prediction. Predicting game outcomes from data consisting of complete games leads to overfitting. To mitigate this, they generated a new self-play data set of games played between the RL policy and itself.

Searching with policy and value networks

AlphaGo combines the policy and value networks in an Monte Carlo tree search (MCTS) algorithm.

Discussion

During the match against Fan Hui, AlphaGo evaluated thousands of times fewer positions than Deep Blue did in its chess match against Kasparov, compensating by selecting positions more intelligently using the policy networks and the value network. Is an approach that looks more like humans think. Since we cannot make as many calculations per second as a computer, we compensate this by learning and experience. This has been a huge achievement for AI and might lead us a bit further an AI that can make decisions more like humans does.