Summary of AlphaGo Paper: "Mastering the game of Go with deep neural networks and tree search"

Introduction: The challenging game Go has been known for its very large search space, approximately 250^{150} sequences of moves, and the difficulty of evaluating board positions and moves. The AlphaGo team has achieved tremendous success against other Go programs, and defeated the human European Go champion thanks to some novel techniques. They trained value networks and policy networks using a combination of supervised learning (SL) and reinforcement learning (RL). They also used combined Monte Carlo simulation with value and policy networks in the search algorithm.

Below are the major stages:

- First, the team trained a policy network on 30 million positions by expert human players using SL. The input *s* was a simple representation of the board state and output was a probability distribution over all legal moves *a*. The policy network was trained on randomly sampled state-action pairs (*s, a*), using stochastic gradient ascent to maximize the likelihood of the human move *a* selected in state *s*. The accuracy was 57.0% using all input features, which was much higher than the highest rate of 44.4% at the time by other research groups. They also trained another policy which was less accurate but able to select an action very fast during rollouts.

- Second, since the goal was to win games, the team improved the SL policy network using policy gradient RL. The RL policy network's architecture was the same as the SL policy network. The team let the current RL policy network play against a randomly selected previous iteration of the policy network. The RL policy network won more than 80% of games against the SL policy network, and 85% of games against Pachi, the strongest open-source Go program

- Finally, they trained a value network to evaluate board positions. Its structure was the same as the policy network, but its output was a single prediction instead of probability distribution. The training data set contained 30 million positions sampled from games played between the RL policy network. AlphaGo combined the policy and value networks with Monte Carlo Tree Search. It traversed the tree by simulation and at each step selected an action that maximized action value plus a bonus, which encouraged exploration. AlphaGo evaluated the leaf nodes using both the output of the value network and the outcome of a random self-play game of the fast rollout policy. At the end of the search, AlphaGo chose the most visited move from the root position.

Results and conclusion: AlphaGo won 99.8% of games against other Go programs, and won the match against human European Go champion 5 games to 0, making Go the first program that defeated a human expert player. This was thanks to the novel combination of supervised and reinforcement learning, an a new search algorithm than combines Monte Carlo rollouts and neural networks. The approach was closer to how humans play that that by Deep Blue.