Research Review:

AlphaGo using deep neural networks and tree search by DeepMind

The game of "Go" is considered to be a difficult game in the domain of AI due to challenging decision-making task, an intractable search space : b^d ($b \approx 250$, $d \approx 150$) where b is the game's breadth and d is its depth (game length), an optimal solution quite complex which seems infeasible to directly approximate using a policy or value function.

In order to tackle the challenges of enormous search space and the difficulty of evaluating board positions and moves.DeepMind introduces a new approach that uses "Value networks" to evaluate board positions and "Policy networks" to select moves.

AlphaGo is based on a combination of deep neural networks and tree search, that plays Go beyond the strongest human player level.

These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play.

DeepMind team solves these problems using two trained Neural Network

1. **Policy Network**: Policy network provides the probability distribution of moves. It learns to predict - for any particular position what's the most likely moves were to be played. So instead of looking at one position of all possible legal moves, it looks for top 3 or top 5 most likely moves were taken into consideration.This reduces down the breadth of the search space. Policy network is first trained by Supervised learning using 30 million positions data from the KGS Go Server. (13 convolutional layers policy network). The second stage of policy network is trained using policy gradient reinforcement learning. This improved the policy network further.

2. **Value Network** : Value network evaluate a particular position and determines who is winning (0 - white or 1 - black). The value network is trained using reinforcement training. Initially using KGS data, the value network memorised

the game outcomes rather than generalising to new positions. Achieving a minimum MSE(mean squared error) of 0.37 on the test set, compared to 0.19 on the training set. This is mainly because of the complexity of Go game - as successive positions are strongly correlated, differing by just one stone. To mitigate this problem, alpha-go is made to play against each other and generated new self-play data set consisting of 30 million distinct positions, each sampled from a separate game. Later, the best version of Alpha is trained by playing against the previous best version of AlphGo Eventually value network got better.

AlphaGo combines the policy and value networks in an Monte Carlo tree search algorithm that selects actions by lookahead search. At each state, the selected move will maximize the action value and prior probability, minimize visit count. It is worth noting that the SL policy network performed better in AlphaGo than the stronger RL policy network, presumably because humans select a diverse beam of promising moves, whereas RL optimizes for the single best move.

To evaluate AlphaGo, DeepMind ran an internal tournament among variants of AlphaGo and several other Go programs, including the strongest commercial programs Crazy Stone and Zen, and the strongest open source programs Pachi and Fuego. The results of the tournament suggest that single machine AlphaGo is many dan ranks stronger than any previous Go program, winning 494 out of 495 games (99.8%) against other Go programs. The result also suggests that the two position-evaluation mechanisms are complementary: the value network approximates the outcome of games played by the strong but impractically slow, while the rollouts can precisely score and evaluate the outcome of games played by the weaker but faster rollout policy.

Conclusion

Notable matches of AlphaGo

1. Mar 2016 - AlphaGo vs Fan Hui (AlphaGo won by 5-0)

2. Oct 2016 - AlphaGo vs Lee Sedol (AlphaGo won by 4-1)

AlphaGo was able to beat the strongest human player, thereby achieving one of artificial intelligence’s grand challenges.

AlphaGo evaluated thousands of times fewer positions than Deep Blue did in its chess match against Kasparov compensating by selecting those positions more

intelligently, using the policy network, and evaluating them more precisely, using the value network—an approach that is perhaps closer to how humans play.

Unlike Deep Blue the neural networks of AlphaGo are trained directly from gameplay purely through general-purpose supervised and reinforcement learning methods.