Derek Wang
AI Nanodegree
Jan 2, 2018
Research Review of AlphaGo Paper

Summary

This paper examines the new techniques implemented in Google DeepMind's AlphaGo to create a Go playing agent that was able to compete against the top human Go players as well as 99.8% of all other Go playing programs. Using supervised learning (SL) techniques utilizing games containing human expert moves as well as reinforcement learning (RL) techniques involving self-play, Google was able to train AlphaGo's deep neural network to use "value networks" and "policy networks".

Initially the Google Team created a fast policy and a supervised learning policy network that were then trained to predict the probability of human expert moves. The fast policy was considerably faster at predicting a move at the cost of accuracy while the SL policy network was able to predict expert moves with a 55-57% accuracy, significantly beating other research groups who generally achieved around 44% accuracy. From here, the a RL policy network is initialized in the same structure as the SL policy network, but it develops through self play where the RL policy network is set to play against a randomly selected previous iteration of the policy network. This randomization allows the policy network to combat over-fitting to any particular policy. After a large number of sample games conducted, the RL policy network was able to achieve over 80% win rates against the SL policy network as well as an 85% win rate over a competing program Pachi. Pachi relied on sophisticated Monte Carlo Search and was considered the strongest open source Go Playing agent. In comparison, other Go agents that only relied on supervised learning techniques could only achieve an 11% win rate against Pachi, showing the superiority of the RL policy network.

Following the development of the RL policy network, a value network was developed to predict the outcome of a given position. This value function was defined to use the same policy for move selection of both players and outputs a single prediction of a move. Using a custom training set of distinct positions to encourage generalization of new positions and discourage over-fitting, the value function was able to provide consistently more accurate evaluations of positions at 15,000 times less computational power. This value network and the previously defined policy network allowed for the development of a search algorithm that integrates the both networks and could outperform any Go playing program ever before seen. AlphaGo was able to defeat every other program consistently, and could even win while handicapped. Furthermore, AlphaGo was able to defeat a European Go Champion in a formal game without a handicap.

Key Results

The key result was the fact that a "grand challenge" of AI was completed decades ahead of prior predictions. The use of machine learning techniques to train deep neural networks shows that human-level performance is quickly becoming obtainable in artificial intelligence programs. This gives the community confidence that such success can soon be replicated in other applications. Additionally, the success of AlphaGo demonstrates the breakthrough capabilities of deep neural networks as well as the potential that machine learning techniques have when applied to hard problems. Being able to approach a solution using these techniques when tackling a problem that is too complex to evaluate traditionally will have a huge impact on the way problems are solved in the future.