Research Review for Build a Game-Playing Agent project (AIND)

Neil Mistry

May 16, 2018

For the research review portion of this project, I have reviewed the paper, “Mastering the game of Go with deep neural networks and tree search”, published Jan 27, 2016 on Nature, International Journal of Science.

This paper explores using ‘value networks’ to evaluate board positions and a ‘policy network’ to explore move selection in playing the game ‘Go’. Both value networks and policy networks are deep neural networks which are trained by supervised learning (from data taken from games played by expert ‘Go’ players). The algorithm is further refined by using reinforcement learning gained from each game played by the algorithm.

The developers also explored using Monte Carlo simulation in combination with deep neural networks and managed to achieve a win rate of 99.8% against other ‘Go’ algorithms. Additionally, this algorithm defeated the European ‘Go’ champion player, 5 games to null.

Much like the Isolation game, ‘Go’ has a large number of moves possible for every node in the search tree. Completing a search of the entire tree can be exhaustive. The developers used two general principals to reduce the search size.

First, the search tree is truncated (depth), by using an approximation for the value of each state to that predicts the outcome of the node without going further down the tree. Second, the breadth of the tree is reduced by sampling actions over a policy that is probabilistically distributed through a set of actions in each state using Monte Carlo simulation.

Supervised learning is used to create a neural network that has a final layer which outputs a probabilistically distributed over all legal moves in a state. The network is trained on randomly selected action and state pairs using stochastic gradient descent.

Reinforcement learning is used in a second stage to improve the policies using an identical structure to the supervised learning neural network. Similarly in value networks, the reinforcement learning models are deployed to evaluate the outcome from each position given the policy of both players.

The algorithm then combines the value network and policy networks using Monte Carlo Simulation. The Monte Carlo Simulation looks ahead in the search tree. An action is selected by maximizing the action value and a bonus term that is proportional to the prior probability with some decay (reinforcement learning).