Orlando Trejo November 26, 2017

AIND-Isolation: Research Review

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

Summary / Techniques:

In a game-changing paper, the DeepMind team report how neural networks and tree searches can be leveraged to create a computer program, AlphaGo, capable of defeating a professional human player in a full-sized game of Go. Their approach can be decomposed into two steps. The first step consists of training deep neural networks using supervised and reinforcement learning strategies. This machine learning step creates policy and value networks that can be used to select moves. The second step consists of complementing the policy and value networks with a Monte Carlo tree search (MCTS) that determines actions through a lookahead search.

The machine learning step can be subdivided into three sub-steps: 1) supervised learning (SL) of policy networks; 2) reinforcement learning (RL) of policy network; and 3) RL of value networks. A policy network is a method to help an agent decide on an action given the state it is in, while a value network helps evaluate the position of a given player for a given outcome. The SL of policy networks consist of a 13-layer network trained on 30 million positions of expert human moves. This network can achieve up to a 57% prediction accuracy of expert moves. The next step is RL of policy networks that is trained in a network similar to the SL network. However, in this case, games are played between networks and the policy network is trained towards maximizing the expected outcome (i.e. winning). The RL policy network can beat the SL network 80% of the time when played against each other. The final step of the training procedure consists of RL to evaluate positions based on their predicted outcomes. The RL value network has similar architecture to the RL policy network; however, it differs in that it predicts a single outcome instead of a probability distribution. Using a RL value network can outperform Monte Carlo rollouts in terms of accuracy and computation time. Integrating this machine learning pipeline with a MCTS provides the best game-playing performance.

The main second step of AlphaGo combines the policy and value networks with an MCTS algorithm that helps select moves by simulating the results of possible moves. By traversing the search tree through simulations, it is possible to determine an action by estimate the value of a state. The value of state is a linear combination between the respective values from the value network and outcome from a rollout. At the end of the simulation, the selected move is the one that was visited the most times during the search. In the final version AlphaGo, 40 search threads were implemented.

Results:

To gauge its performance, AlphaGo was played against various Go programs, which are based on state-of-the-art MCTS or search algorithms. With 5 seconds of computation time allowed for all contestants, AlphaGo outperformed all other programs 99.8% percent of the time. The last and most-telling test was done by playing again Fan Hui, who was a recent multi-year winner of the European Go championships. From October 5 to October 9 of 2015, AlphaGo beat Fan Hui in all planned 5 games, marking a critical point in history.