Research Review

Frederick Chyan

July 23, 2017

Selected Game Paper: AlphaGo by the DeepMind Team.

Introduction

Just like the game of Isolation, Go is also finite, deterministic, and perfect information. However, unlike Isolation, Go has much larger search space which makes designing an optimal value function by hand very difficult. The AlphaGo paper is selected to understand how state of the art AI tackles games that's previous thought impossible for a machine to beat expert human player. AlphaGo uses novel approaches to combine neural networks and Monte Carlo tree search to achieve this.

Techniques

Policy function p(a|s) estimates the utility of selecting action a in game state s. Value function v(s) estimates the outcome of the game under perfect play by both players. The greater the value the better. AlphaGo first uses deep convolutional neural networks to estimate the policy and value network. Deep convolutional neural networks were used to in visual domains such as image classification and face recognitions, here it passes the board positions as 19 x 19 image, and use convolutional layers to construct a representation of the position. Next, it uses policy and value networks to carry out Monte Carlo tree search by evaluating the position using value network and sampling actions using policy network.

Policy Network and Value Network

AlphaGo's training pipeline works as follows. First, a 13 layer network is trained from 30 million positions from the KGS Go Server, this is the Supervised Learning policy network p_sigma, it has 55.7% accuracy and 3ms to select an action. Also trained is a fast rollout policy network which has lower accuracy of 24.2% but only 2us to select an action. Next a Reinforcement Learning policy network, p_rho is trained by making it play against a randomly select previous iteration of the policy network. Finally, the value network v_theta is trained from the self-play data from RL policy network regressing on (state, outcome) pair. The reason why the value network is trained this way rather than using available data of complete games is because of overfitting due to successive positions are strongly correlated.

Combining Neural Networks in Monte Carlo Tree Search (MCTS)

Perhaps the most interesting and powerful technique in AlphaGo is using Monte Carlo tree search to improve the prediction. It simulates and sample matches by rolling out searches in state spaces. It is highly parallelizable, and can be stopped at any time to achieve a balance of speed and accuracy. It introduces a random element in selecting next action, and constantly updates the parameters in estimation function, effectively making the AI smarter continuously and asynchronously. At every edge of the tree, there is an action value Q(s, a), visit count N(s,a), and prior probability P(s,a). The prior probability is based on SL policy network p_sigma. The MCTS is carried out in four steps. It first selects a node to expand, based on the bonus function u(s,a) which is proportional to the prior probability over the visit count. This ensures other actions will also be explored if this node gets expanded too much. When it reaches a leaf node at step L, it then expand the node and process it using SL policy network, now the output probability will be stored as the prior. Next, the leaf node will be evaluated using two different ways. 1. the value network v_theta and 2. playing the game until termination using fast roll out policy network p_pi. The resulting score are then weighted and summed together. Finally, the result will back propagate to the root node updating the visit count N, and action value Q. The AI then chooses the most visited node as the action to take.

Key Result

Utilizing deep learning, Monte Carlo tree search, and reinforcement learning, AlphaGo studied and learned from professional players, then it evolves by playing and simulating countless number of games continuously, updating itself while the game in progress. Using just the value network, Alpha Go achieves a rating that places it in the Amateur dan. Also, it's also able to beat the strongest AI at the time by allowing 4 free moves. With Rollouts, Value Network and Policy network, AlphaGo can win >95% against other players. Meaning the search mechanism successfully combines the strong but impractical reinforcement learning policy network and weaker but faster rollout policy network. Winning several world champions AlphaGo achieved what was thought to be impossible previously.