Udacity Artificial Intelligence Nanodegree
AlphaGo Paper Summary - Research Review
Submission By: Lim Si Jie

Goals

The game of Go has been viewed by experts as the most challenging of classic games for artificial intelligence. This is largely due to its enormous search space and the difficulty of evaluating board positions and moves. The consensus is that it will take another decade to reach the level of human intelligence to win the game. The goal of AlphaGo was to design and engineer a system, a new set of algorithms and techniques that can challenge (and hopefully beat) the professional human Go player, including the world's top Go player champion.

Techniques

AlphaGo thinks fast and slow using a combination of Deep Neural Network and new search algorithm that uses Monte Carlo rollouts. It "intuitively" evaluates board positions and potential moves through deep neural networks. During live play, it refines its intuition by simulating likely progression of moves. There are three main components in the design of AlphaGo: Policy Network, Value Network and Rollout Policy.

1. Policy Network

Policy Network is used to evaluate the best next move based on the current state of the board. The Policy Network was initially trained using tons of human players data. The Supervised Learning policy network is a 13-layer convolutional neural network trained on 30 million moves of human experts.

After which, it was trained with Reinforcement Learning by playing with itself. The aim of the Reinforcement Learning is to prevent overfitting of weights in neural network and also adjusts the policy towards the correct goal of winning games rather than maximizing predictive accuracy.

2. Value Network

Based on the moves suggested from the Policy Network, the Value Network will determine the probability of winning the game. Similar to the Policy Network, the Value Network is first trained using human players data using Supervised Learning and then fortified using Reinforcement Learning to prevent overfitting.

3. Rollout Policy

Compared to prior work of Go computer players that focus on sophisticated Rollout, AlphaGo uses simple Rollout that focuses on improving quality of the Value Network and Policy Network.

To ensure a high-performance Rollout, AlphaGo uses a fast and simple Rollout technique with a new search algorithm that involves the combination of Monte Carlo simulation with its Value Network and Policy Network. This is known as "asynchronous policy and value MCTS" (APV-MCTS).

APV-MCTS evaluates leaf node s_L in two different ways. The first is by the Value Network v_θ (s_L) and the second by the outcome of simulations z_L. These evaluations are then combined using a mixing parameter λ tuned to 0.5, into a leaf evaluation V (s_L).

According to the AlphaGo paper, the Supervised Learning Policy Network performed better in APV-MCTS than the stronger Reinforcement Learning Policy Network. The hypothesized reason for this is that

Supervised Learning Policy Network better reflect the diverse beam of promising moves humans select, whereas Reinforcement Learning Policy Network optimizes for a single best move. However, the situation is reversed for the value function. The value function derived from the stronger Reinforcement Learning Policy Network perform better than the value function derived from Supervised Learning Policy Network.

Results

To evaluate the strength of AlphaGo, an internal tournament was conducted among variants of AlphaGo and several other Go programs like Crazy Stone and Zen, Pachi and Fuego. The tournament showed that AlphaGo is many dan ranks above any of the previous Go program with a win rate of 99.8% (out of 495 games) against other Go programs.

Interestingly, AlphaGo exceeded the performance of all other Go programs even without rollouts. This demonstrates that Value Networks provide a viable alternative to Monte Carlo evaluation in Go.

Another important finding from the AlphaGo paper is that two position-evaluation mechanisms are complementary. The Value Network approximates the outcome of games played by the strong but impractically slow Policy Network, while the Rollouts can precisely score and evaluate the outcome of games played by the weaker but faster Rollout policy.

In late 2015, AlphaGo and Fan Hui competed in a formal five-game match which AlphaGo won 5 games to 0. This is the first time that a computer Go program has defeated a human professional player without any handicap.