Research Review

Dexiang Xu

Abstract:

This paper introduces a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement earning from games of self-play. Also, this paper introduces a new search algorithm that combines Monte Carlo simulation with value and policy networks. Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs, and defeated the human European Go champion by 5: 0

Deep convolutional neural networks:

The team uses a deep convolutional neural networks architecture for the game of Go. They pass in the board position as a 19×19 image and use convolutional layers to construct a representation of the position. In this way, it could reduce the effective depth and breadth of the search tree: evaluating positions using a value network, and sampling actions using a policy network.

And They train the neural networks using a pipeline consisting of three stages of machine learning. And program combines the policy and value networks with MCTS.

Stage1. Supervised learning of policy networks:

This is a prior work on predicting expert moves in Go game using supervised learning. They trained a 13-layer policy network, which is the SL policy network, from 30 million positions from the KGS Go Server. They also trained a faster but less accurate rollout policy, using a linear softmax of small pattern features.

Stage2. Reinforcement learning of policy networks:

The second stage of the training pipeline aims at improving the policy network by policy gradient reinforcement learning. Games were played between the current policy network and a randomly selected previous iteration of the policy network. Randomizing from a pool of opponents in this way stabilizes training by preventing overfitting to the current policy.

A reward function is set for terminal time steps, from the perspective of the current player. Weights are then updated at each time step t by stochastic gradient ascent in the direction that maximizes expected outcome.

Stage3. Reinforcement learning of value networks:

This stage focuses on position evaluation, estimating a value function that predicts the outcome from position s of games played by using policy p for both players. This neural network has a similar architecture to the policy network, but outputs a single prediction instead of a probability distribution. And to mitigate the overfitting problem, a new self-play data set consisting of 30 million distinct positions is generated, and each of them are sampled from a separate game. Each game was played between the RL policy network and itself until the game terminated.

Searching with policy and value networks:

AlphaGo combines the policy and value networks in an MCTS algorithm that selects actions by lookahead search. And to efficiently complete this step, they use an asynchronous multi-threaded search that executes simulations on CPUs, and computes policy and value networks in parallel on GPUs.

Evaluating the playing strength of AlphaGo(result):

AlphaGo has much stronger performance against previous Go program based on high-performance MCTS algorithms. AlphaGo also won 5-0 against a human professional player.