Research Review: A simple summary of the AlphaGo Nature paper

Summary

The game of Go is viewed as one of the more challenging games for AI, due to the large number of possible moves/strategies. The papers authors have used deep neural nets to develop "value networks" that provide an evaluation of the boards state relative to the player and "policy networks" that will provide an evaluation of the next move. The nets are trained using a combination of supervised learning based on historical games, and reinforcment learning from games played against itself. The resulting networks are then combined using Monte Carlo Tree Searcg (MCTS) to acheive a new state-of-the-art.

Techniques

The AlphaGo team has developed a training pipeline with several stages:

1. Supervised learning of policy networks - In this stage, historical games are used to predict moves(also known as 'actions). Stochastic gradient ascent is used on randomly sampled state/action pairs from the historical games to select the most likely action given a state.
2. Reinforcement learning of policy networks- In order to refine the predictions from the prior stage, the prior policy network is used to play batches of games, with the opponents being the most current and a randomly selected prior version. The result of the games are used to update the weights of the network, and new batches played, as long as training progresses. T
3. Reinforcement learning of value networks - In this stage of the training pipeline, A deep neural netof similar architecture to the the prviouse is used to predict the game outcome based on current board position, with the primary difference being the inputs(state/action, vs state/outcome) and the output. In this network they output consits of a single prediction instead of a probability distribution. To minimize overfitting, the researchers generated a new self-play data set from the games previously played, with each distinct state coming from a different game. They then used this dataset to train the value network

   After training, AlphaGo uses a customized MCTS algorithm that uses combines the policy and value networks as an evaluation function . Evaluating this algorithm requires significantly more computation than other methods, and so AlphaGo developed methods that may be run as multithreaded or even distributed to multiple machines

Results

The researchers ran an internal tournament of various Go playing programs, and evaluated the Elo rating of each program. The Elo system is a method to calculate the relative skill of competitors. The base score for the calulation used profesional Go player Fan Hui's rating of 2908 to calculate it's core score. The tournament results suggest that the AlphaGo is a much stronger player than the other current programs, with the distributed version winning 100% of games against other programs. They also evaluated variants that used just the value network, etc, and found that they also performed highlt against opponents, winning as much as 95% of games. Finally, the distributed version of AlphaGo played a 5 game match against Fan Hui and won all 5 games, the first time a computer player has beaten a profesional player, and believed previously to be impossible.

After publication of the paper, AlphaGo went on to win 4-1 against Lee Sedol, who is the 2nd highest rated Go player world wide