Mastering the game of Go with neural networks and tree search

Review by Safak Ozkan
shafaksan@gmail.com

1. INTRODUCTION

The current paper presents a short analysis of the AI program AlphaGo that plays a game of Go on a standard board of size 19x19. The program combines deep neural networks and tree search to successfully beat the European Go champion Fan Hui in 2015 winning all of the 5 formal games.

The challenges of designing an AI agent that plays Go resides in the high branching factor (b=250), and depth values (d=150) of the game which in combination poses an intractable search space, and also in constructing an effective heuristic value function at nodes where the search tree is truncated. AlphaGo mitigates these challenges by using:

- **Convolutional Neural Networks** to represent the game states and evaluating the positions by taking an image of 19 by 19 pixels as input, and
- **Monte Carlo Tree Search** (MCTS) to estimate the value at each state, where a Monte Carlo Simulation plays out to the end of the search tree without branching, but by sampling sequences of actions for each player following a certain policy.

2. METHODOLOGY

Several different policy networks are used in this work:

1. Two separate Supervised Learning policy networks were trained. Both networks sampled directly from expert human moves provided by the KGS Go server. They used 13 convolutional layers to represent the board states. The slow rollout policy p_{\sigma} used a softmax classifier to predict a probability distribution over possible actions. A second, and 3 orders of magnitude faster rollout policy p_{\pi} was also constructed with a linear softmax classifier.
2. A Reinforcement Learning policy network p_{\rho} is used to improve on p_{\sigma} by self-play. A reward function r(s) is used which assigns a reward value z_t = r(s_L) at the end of the game.
3. Finally, a value network v_{\theta}(s) is trained to predict the winner in self-play games of RL policy network.

These networks are eventually combined by the MCTS. The weights are learned by SGD updates in the direction that maximizes the likelihood.

3. RESULTS AND EVALUATION

The competition tournaments between single machine AlphaGo and several other contemporary AI programs resulted in a win by AlphaGo in 99.8% of all games showing that it is many ranks stronger than its counterparts. In the absence of policy networks and using value networks alone was enough for AlphaGo to outperform the other AI opponents. This showed that the value networks provide a viable alternative to Monte Carlo evaluation.

During a match with Fan Hui, AlphaGo evaluated thousands of times fewer positions than Deep Blue did in its chess match against Kasparov in 1997. AlphaGo selected only certain actions more intelligently, using the policy network, and evaluated them more precisely using the value network.

Finally, this work inspires hope to leverage AlphaGo's novel techniques to help achieve human-level performances in other AI domains that was previously seen as unconquerable.