AlphaGo Summary

Danny Luo

Summary

Silver, Huang et al. 2016. Mastering the game of Go with deep neural networks and tree search.

AlphaGo begins with a supervised learning policy network to predict human expert moves and a reinforcement learning policy network to improve the SL policy network by optimizing the weights to what it thinks will win the match. The policy network takes in a representation of the board as the input and outputs a probability distribution $p(a|s)$ over legals moves a given position s.

The SL policy network consists of constitutional layers with 128, 192 or 256 filters on raw board positions, explicit symmetry ensembles and specific features such as liberties and capture size.

The RL policy network is identical in structure to the SL policy network, with the final weights of the SL policy network as the weight initialization for RL policy network. A game is played between the current policy network and a random previous iteration of the policy network. The policy network predicts the outcome of the game from the current time step t: $z_t = \pm 1$ for win and loss respectively. The weights ρ are then updated:

$$\delta\rho \propto \frac{\partial \log p_\rho(a_t|s_t)}{\partial\rho} z_t$$

The final stage is the position evaluation, which estimates value functions $v^p(s)$ that predicts the outcome from position s using policy p for both players. The weights of the value network are trained by regression on (s, z) pairs.

AlphaGo searches for the right move by combining policy and value network in a Monte Carlo Tree Search (MCTS) algorithm. The selection criteria for move a at time t is $a = \arg\max(Q(s_t, a) + u(s_t, a))$ where Q is the action value and u is a bonus:

$$u(s, a) \propto \frac{P(s, a)}{1 + N(s, a)}$$

where P is the prior probability and N is the visit count. A fast rollout policy is used to simulate the game.

Questions

1. What are the "explicit symmetry ensembles" used in the SL policy network?
2. In the RL policy network, is z_t forecasted? Is a game played and then the weights are then adjusted.
3. What is this action value $Q(s, a)$?