Mastering the Game of Go without Human Knowledge

The goal of the paper is to master the game of Go without human knowledge, that is, to demonstrate a pure reinforcement learning approach is fully feasible: even in the most challenging areas, without human guidance (supervised learning), it is possible to train to superhuman level given no knowledge beyond basic rules. Using approaches in the paper, AlphaGo zero defeated all other versions of AlphaGo, which were trained from human data using handcraft features, by a large margin. The followings are some technique AlphaGo zero used.

Search algorithm

1. AlphaGo zero uses simpler Monte-Carlo search tree (MCTS) as its search algorithm to search the state space, it is a simpler MCTS without rollout. Instead, it relies on the latest neural network f_{θ_i} to guide its simulation. When selecting each move, AlphaGo zero uses MCTS with 1600 simulations.

Heuristic function to evaluate the game

1. Unlike the previous version of AlphaGo, AlphaGo zero only uses one neural network, which is the residual convolutional neural network. It is input features is 19*19*17 image stack comprising 17 binary feature planes. 16 feature planes consist of a binary value indicating the presence of players' moves (each color has 8 feature planes), the last one is a constant value indicating the color of the current player. One thing to notice is that, unlike the previous version of AlphaGo, AlphaGo zero didn't use any handcraft features. The features space only consists board states, that is, without any human knowledge.
2. AlphaGo zero starts with the untrained neural network f_{θ_0} with random parameters θ_0, the parameters θ of the neural network were continuously updated after certain amount of games to better guide the MCTS.
3. The neural network is optimized by stochastic gradient descent with momentum and learning rate annealing; the data is sampled uniformly at random from most recent 500000 games. The optimization process produces a new checkpoint every 1,000 training steps. This checkpoint is evaluated by the evaluator, and it may be used for generating the next batch of self-play games.

Self-play

AlphaGo Zero's self-play training pipeline consists of three main components, all executed asynchronously in parallel. Neural network parameters θ_i are continually optimized from recent self-play data; AlphaGo Zero players $\alpha\theta_i$ are continually evaluated; and the best performing player so far, $\alpha\theta_*$, is used to generate new self-play data.

Result

Using the above approaches, AlphaGo Zero defeated the strongest previous versions of AlphaGo, which were trained from human data using handcrafted features, by a large margin. Figure 1: Performance of AlphaGo Zero shows that if we select move only based on the fully train neural network from AlphaGo zero without using any lookahead, achieved Elo rating of 3055. AlphaGo zero achieved Elo rating of 5185.

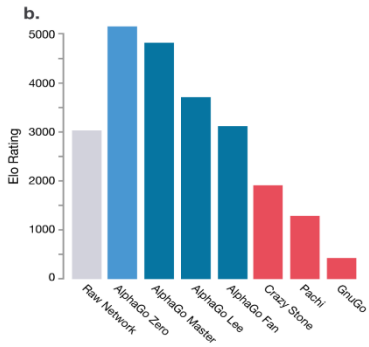

Figure 1: Performance of AlphaGo Zero

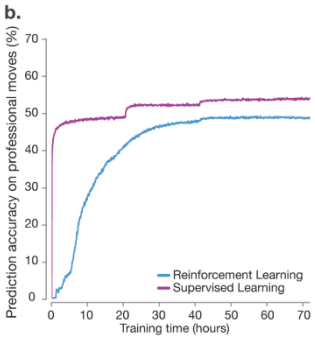

Figure 2: Prediction accuracy on human professional moves

As Figure 2: Prediction accuracy on human professional moves shows, two neural networks, one is trained by self-play (reinforcement learning), another is trained by supervised learning. The accuracy of reinforcement learning neural network is not as good as supervised learning.