Paper to review

"Mastering the game of Go without human knowledge", D. Silver et.al., Nature, vol. 550 (2017) 354

Goal of Paper

In the previous paper (*Nature, vol. 529 (2016) 484*), the authors combined the supervised learning (SL) and reinforcement learning (RL) techniques for deep neural networks (DNN) to build a powerful algorithm for playing Go. *AlphaGo* which is based on this algorithm has defeated the world human champions of Go. The aim of this paper is to introduce a more powerful DNN-based algorithm for playing Go without using SL and then to analyze the performance of the model based on it, called *AlphaGo Zero.*

Algorithm for AlphaGo Zero

For later use, let us first define a state and an action as follows: state = configurations of stones at a given point and several steps before, action = to put a stone. Then the main points of the algorithm for Alpha Go Zero are as follows:

- A single DNN which takes states as inputs and then outputs both the policy (= with which percentage a possible action is taken next for a given state) and value function (= a function which return a number characterizing how much a given state is good for winning the game in future). This DNN is composed of multiple residual blocks of convolutional layers with batch normalization and Relu activation.
- Monte Carlo Tree Search (MCTS) based on the policy and value function determined by the DNN. For a given state, this determines within a limited time period the probabilities for taking possible next actions. By using MCTS at each step, self-play of a game is done. Once the game ends, the outcome is assigned to the states that appeared in the course of the game (± 1 for winner/loser).
- Update the weights of the DNN based on RL. This is carried out such that the following two quantities are decreased: the difference between the outcomes and values computed with the current value function and difference between the probabilities derived with the current policy and the probabilities computed with MCTS.
- The initial weights of the DNN is chosen randomly, and the weights are updated by repeating the above MCTS + RL process.

Summary

AlphaGo Zero turns out to be the strongest Go players in the world, even defeating AlphaGo in the previous paper with 100 - 0. On top of this, AlphaGo Zero can be trained much faster than AlphaGo. The followings are some comparison between AlphaGo Zero and AlphaGo:

- For AlphaGo, two DNNs are introduced, one for the policy and another for the value function, while AlphaGo Zero uses single DNN. This simplifies the architecture of the DNN drastically.
- AlphaGo uses a combination of SL and RL. Thus, for SL, data of the games by human experts are needed. On the other hand, AlphaGo uses RL only. In other words, no data of the games by human experts is needed to train AlphaGo Zero. AlphaGo Zero learns by itself.
- Input for AlphaGo Zero is configurations of the stones only, while the input for AlphaGo is configurations of the stones plus other human made features.