Research Review: AlphaGo

Summary:

The game of Go has long been reviewed as the most challenging of classic games for AI due to its enormous search space (branching factor $\approx 250^{150}$) and the difficulty of evaluating players' moves. In this case, the DeepMind[1] team introduced a new approach that uses a '**value network**' to evaluate board positions and a '**policy network**' to select 'optimal' moves. The foundation of both networks are based on the **convolutional neural network**. Particularly, the policy network is trained by both supervised learning from human expert games and **reinforcement learning** from games of self-play. A new search algorithm that combines **Monte Carlo Tree Search (MCTS) with value and policy networks** is also introduced. As a result, AlphaGo is the first computer program has defeated human professional players in the full-sized game of Go which is a historical breakthrough in AI domain.

Training the networks:

In training stage, there are 4 neural networks need to be trained. The first one is the **supervised learning policy network** which is a 13-layer convolutional neural network trained on 160 thousand human expert games. A final soft-max layer outputs a probability distribution over all legal moves a that indicates the possibilities of each move that a human player might take. In addition to SL policy network, a faster and less accurate **rollout policy network** is trained for faster moves prediction in searching stage. The third network is a **reinforcement learning policy network** that aim to improve the performance of SL policy network by self-play. The RL policy network has the same architecture as SL policy network and is initialized by copying the weights from the trained SL policy network. After playing against itself 1.2 million times, it won more than 80% of games against the SL policy network. The last neural network is the **Reinforcement learning of value network** which trained on 1.5 billion self-play moves. It predicts the wining rate of each move at any given state s.

Game-play stage:

At game-play stage, the MCTS plays a very important role here. The basic idea of MCTS is to running Monte Carlo simulation for many times and choose the most frequent next move. During game-tree expansion stage, the trained SL policy network is used. So, only the nodes with a large possibility figure would be selected and others will be ignored in most cases. Then, at every leaf nodes, the trained value network and the rollout network are engaged for evaluation purpose. The evaluation is based on both networks equally. After evaluation, the result is used to update the Q-value of current branch down the root of the game tree. Repeating the simulation until the time is running out. Finally, return the most frequent next move.

Results:

- AlphaGo achieved a 99.8% winning rate against other Go Programs.
- Defeated the human European Go champion by 5 games to 0.
- Defeated the human World Go champion Lee Sedol by 4 games to 1. (March 2016)
- Proved that neural network is so strong and it has implicitly understood many sophisticated aspects of Go.

References:

[1] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), pp.484-489.