Mastering the game of Go with deep neural networks and tree search

Name: Yung-Chun Lu

Date: 2017/02/14

Go has been considered in many ways of the difficulties by artificial intelligence, like a challenging decision-making task, an intractable search space, and an optimal solution so complex it appears infeasible to directly approximate using a policy or value function. In order to tackle down the challenges of enormous search space and the difficulty of evaluating board positions and moves, DeepMind introduces a new approach that uses “value networks” to evaluate board positions and “policy networks” to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play. The structure of this article could be breakdown into three pieces, how to training pipeline works, how to search the best move by combining the policy and value network and how to evaluate the result.

The first stage of training pipeline is using supervised learning to learn the thinking policy of human expert moves. AlphaGo trained a 13-layer policy network, which is called the SL policy network, from 30 million positions from the KGS Go Server. The network predicted expert moves on a held out test set with an accuracy of 57.0% compared to the state-of-the-art from other research groups of 44.4%. The second stage is aiming at improving the policy network by policy gradient reinforcement learning(RL) in order to maximize the outcome against previous versions of the policy network. When played head-to-head, the RL policy network won more than 80% of games against the SL policy network. RL policy is also tested against the strongest open-source Go program, Pachi, a sophisticated Monte Carlo search program. The RL policy network won 85% of games against Pachi. The final stage of training pipeline is focusing on value network which predict the outcome from position of game played by both same policy. The mean square errors between the predicted value and the corresponding outcome are 0.226 and 0.234 on the training and test set respectively, indicating minimal overfitting.

AlphaGo combines the policy and value networks in an Monte Carlo tree search algorithm that selects actions by lookahead search. At each state, the selected move will maximize the action value and prior probability, minimize visit count. It is worth noting that the SL policy network performed better in AlphaGo than the stronger RL policy network, presumably because humans select a diverse beam of promising moves, whereas RL optimizes for the single best move.

To evaluate AlphaGo, DeepMind ran an internal tournament among variants of AlphaGo and several other Go programs, including the strongest commercial programs Crazy Stone and Zen, and the strongest open source programs Pachi and Fuego. The results of the tournament suggest that single machine AlphaGo is many dan ranks stronger than any previous Go program, winning 494 out of 495 games (99.8%) against other Go programs. The result also suggests that the two position-evaluation mechanisms are complementary: the value network approximates the outcome of games played by the strong but impractically slow, while the rollouts can precisely score and evaluate the outcome of games played by the weaker but faster rollout policy.

Finally, DeepMind evaluated the distributed version of AlphaGo against Fan Hui, a professional 2 dan, and the winner of the 2013, 2014 and 2015 European Go championships. AlphaGo won the match 5 games to 0. This is the first time that a computer Go program has defeated a human professional player, without handicap, in the full game of Go—a feat that was previously believed to be at least a decade away.