Alpha Go - from Fan to Zero

This review is a summarisation of the development of Alpha Go, developed initially as a Go-playing program and later generalised to Shogi and Chess.

Although multiple iterations of the program exist, we will focus on three distinct version which has associated papers. Namely, AlphaGo Fan, AlphaGo Zero and Alpha Zero.

Alpha Go Fan - Mastering the Game of Go with Deep Neural Networks and Tree Search

The first iteration AlphaGo Fan was the first computer program capable of beating a professional Go player on a full-size game. The key innovation behind the program lies in the successful tuning of deep neural networks.

In game tree search environment, the computational complexity of a game can be characterised by the breadth and depth of the tree and cn be characterised as O(bˆd). In most settings, an exhaustive search is impossible, and heuristics have been devised to search efficiently.

The search problem was attacked from two angles in Alpha Go Fan through the implementation of two separate deep neural networks. First, a policy network identifies moves likely to result in a win; while a second value network provides an estimate of the terminal state and winning probability.

The policy network was first trained on expert human player games using supervised learning; the model achieved a state of art status in the accuracy of predicting future human moves. The same weights are then used as initial values in a similar network but trained using reinforcement learning; this step redirects the weights from predicting human moves to identifying the best possible action.

A value network is then trained based on the games generated by the above policy network. The value network has a similar structure to the policy network but instead of returning a vector of probability for each legal move, a single scalar value estimating the probability of winning given the current position is obtained. The non-linear function approximation provided a much more powerful representation than the standard linear approximation used in other programs and contributed significantly to the success of the program.

The tree search method implemented by Alpha Go Fan is Monte Carlo Tree Search, which also produces the strongest competitor in the game of Go. During the Monte Carlo Tree Search, the policy network provides the prior probability for the selection step but the value decays with number of visit to encourage exploration. A leaf is expanded when the number of visits to this leaf has surpassed a critical threshold. The value of the leaf is then evaluated by weighting

the value network and a fast Monte Carlo Rollout. Finally, the values are rolled all the back to the current root.

The two neural network provides Alpha Go Fan with the ability to search efficiently across breadth by focusing on moves with high winning probability but at the same time reduce the depth requirement with an accurate value estimation. Both innovations are critical to the astounding superhuman feat.

Alpha Go Zero - Mastering the Game of Go without Human Knowledge

This paper introduces several improvements on top of the already world champion, AlphaGo Master. But the most notable change was the removal of several handcrafted features and the exclusive use of reinforcement learning.

The elimination of supervised learning demonstrated that learning from pure self-playing was practically feasible and that human knowledge was no longer essential in super-human performance. Further, the agent along with numerous improvements was able to defeat the strongest predecessor Alpha go Lee in just 36 hours of training.

Two additional improvements were the unification of the policy and value network and the implementation of the residual network. The empirical analysis demonstrated the single policy value network raised the strength of the program through improved computational efficiency, while the residual network lifted the accuracy of the overall neural networks. The Monte Carlo rollouts were eliminated as a result of improved accuracy.

One notable point on the superior performance of Alpha Go Zero may be related to the optimisation of Deep Neural Networks (DNN). It is widely accepted that the solution of DNNs does not correspond to the global maximum. However, the random initialisation of Alpha Go Zero along with reinforcement learning may have provided the optimisation with greater search space to land in a superior local maximum.

Alpha Zero - Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

The latest iteration of Alpha Go is no longer confined to the game Go. The new program is based on a similar architecture to the Alpha Go Zero; however, game-specific adaptation such as symmetry was eliminated and incorporated outcome such as draw to play Shogi and chess.

The paper showed the same architecture was able to perform superior against all state of the art programs with no adaptation including hyperparameters.