Research Review: AlphaGo

The DeepMind team built a Go program that was able to defeat a human professional player (European Go champion Fan Hui) by 5 games to 0. Previous state-of-the-art were able to play at amateur level. This achievement was considered to be at least a decade away.

The previous strongest Go program was based Monte Carlo tree search using a policy, trained to predict human expert moves, to narrow the search tree and to select the move during simulation. The issue with that program is that the policy was based only on a linear combination of input features.

The DeepMind team built a search algorithm that is a combination of Monte Carlo simulation, a “value network” to evaluate game positions and a “policy network” to sample game moves. These deep neural networks are used to reduce the breadth and depth of the search tree.

In the first stage of the training, supervised learning is used to train the “policy network”, observing human experts. Then, the “policy network” is optimised for winning games (rather than accurately predict the outcome of a game) using reinforcement learning and self-play. In the second stage, the “value network” is trained to evaluate board positions, also using reinforcement learning and self-play. Finally, because evaluations of value and policy networks are much more computationally intense than traditional heuristics, AlphaGo uses a distributed multi-threaded search, running the simulations on 1,202 CPUs and evaluating the value and policy networks in parallel on 176 GPUs.

While playing the European Go champion, AlphaGo used the policy network to more intelligently select which position to expand, and the value network to more precisely evaluate those positions, compared to Deep Blue in it’s chess match against Garry Kasparov. The result was that AlphaGo evaluated thousands of times less positions than Deep Blue, closer mimicking the way in which humans play. Another significant distinction between the two game agents was that Deep Blue relied on evaluation functions that were handcrafted and optimised for it, while AlphaGo used only generic supervised and reinforcement learning method to independently learn the best policy and value networks.