AlphaGo by DeepMind Team

The article presents the research aiming for building a program, AlphaGo, designed to play the game of Go at the level of the strongest human players, thus achieving one of artificial intelligence's 'grand challenges'. The game of Go has long been regarded as the most challenging of classic games to solve for AI as it embodies many of the problems or impediments faced by the domain, such as an intractable search space, complex decision making task and rather infeasible evaluation of board positions and moves. For the first time effective move selection and position evaluation function have been developed as a result of unprecedented combination of reinforcement and supervised learning.

Supervised Learning has been applied to build a policy network which aims at predicting expert moves. A 13-layer policy network, denoted as the SL policy network has been trained on randomly sampled state-action pairs from 30 million positions from KGS Go Server. The policy network receives a simple representation of the board state as input and outputs probability distribution over all legal moves which is the result of softmax applied in the final layer. The network predicted expert moves on held out test set with 57% of accuracy using all input attributes, while 55.7% only with raw board position and move history as input, thereby, it exceeded all the other state-of-the art solutions developed so far.

Further the policy network has been trained by policy gradient reinforcement learning what has come down to n games played in parallel, between the current policy network and a randomly sampled opponent using parameters from previous iterations. This step aimed at decreasing the network variance, thus preventing overfitting.

The next stage was creation of value network which evaluates positions by estimating a value function v(s) predicting the outcome for each player from a given position s. The value network has been trained on a newly generated self-play data set consisting of 30 million distinct positions individually sampled from separate games. The training resulted in MSEs of 0.226 and 0.234 on the training and test set respectively, indicating minimal overfitting.

The final phase was combining the policy and value networks in the Monte Carlo tree search algorithm selecting actions by lookahead search. Such combination of MCTS with deep neural network requires enormous amount of computation what has been addressed by the final version of AlphaGo which exploited 40 search threads, 48 CPUs, and 8 GPUs. In the final evaluation it has been shown that the SL policy network performed better in AlphaGo than the stronger the RL (reinforcement learning) policy network as the SL one appears to mimic more humans players as they choose a diverse beam of promising moves, whereas the RL optimizes for a single move. However, the RL network outperformed the SL one in the value function of AlphaGo.

Performance of AlpaGo has been tested against the existing state-of-the-art Go programs. AlphaGo surpassed the other Go programs by winning 494 out of 495 games (99.8% of win rate). The program has been also competing with a human expert in the field (European Go champion), defeating him by 5 games to 0. This was a first ever victory of a computer program over a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.