AlphaGo research paper

The AlphaGo program was built to beat human professional players in the game of Go.

Traditionally exhaustive search of optimal value function v*(s) which determines the outcome of the game was infeasible in Go as it contains approximately 250^{150} possible sequences of moves (250 – game's breadth which is the number of legal moves per position, 150 – game's depth which is the game length). Deep convolutional neural networks are used by Alpha Go to reduce the effective depth and breadth of the search tree: evaluating positions using a value network, and sampling actions using a policy network to select moves. The neural networks were trained using a pipeline consisting of several stages of machine learning. It's begun by training a supervised learning (SL) policy network p_σ directly from expert human moves. This provides fast, efficient learning updates with immediate feedback and high-quality gradients. Similar to prior work, a fast policy p_π was also trained that can rapidly sample actions during rollouts. Next, a reinforcement learning (RL) policy network p_ρ was trained that improves the SL policy network by optimizing the final outcome of games of self-play. This adjusts the policy towards the correct goal of winning games, rather than maximizing predictive accuracy. The program efficiently combines policy and value networks with Monte Carlo Tree Search (MCTS).

The playing strength of AlphGo was evaluated by running an internal tournament among variants of AlphaGo and several other Go programs. All programs were allowed 5s of computation time per move. 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. To provide a greater challenge to AlphaGo, we also played games with four handicap stones (that is, free moves for the opponent); AlphaGo won 77%, 86%, and 99% of handicap games against Crazy Stone, Zen and Pachi, respectively. The distributed version of AlphaGo was significantly stronger, winning 77% of games against single-machine AlphaGo and 100% of its games against other programs. Variants of AlphaGo that evaluated positions using just the value network ($\lambda = 0$) or just rollouts ($\lambda = 1$) were also assessed. Even without rollouts, AlphaGo exceeded the performance of all other Go programs, demonstrating that value networks provide a viable alternative to Monte Carlo evaluation in Go. However, the mixed evaluation ($\lambda = 0.5$) performed best, winning ≥95% of games against other variants. Finally, the distributed version of AlphaGo was evaluated against Fan Hui, a professional 2 dan, and the winner of the 2013, 2014 and 2015 European Go championships. Over 5–9 October 2015 AlphaGo and Fan Hui competed in a formal five-game match. AlphaGo won 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.