Build a game-playing agent project

Part 2: Research review

Michail Kovanis

Research paper: Silver D., Schrittwieser J., Simonyan K., et al., *Mastering the game of go without human knowledge* (2017). Nature. 550:354-359. doi: 10.1038/nature24270.

Objective: To achieve superhuman performance in the game of Go by using a single neural network, trained by self-play reinforcement-learning, starting from random moves and without prior human knowledge other than the rules of the game.

Methods: AlphaGo Zero uses a deep neural network (DNN), which takes as input the raw board representation *(s)* of a current position of a game of Go and its history, and outputs a vector of move probabilities and a value $(\boldsymbol{p},v) = f_\vartheta(s)$. The vector $\boldsymbol{p}$ represents the probability of a player to select a certain move and the value v represents the probability to win when starting from position s. The DNN is trained through many rounds of self-play and in the beginning, assigns random probabilities and values to each state s. At each time step *(t)* of a game of self-play, a Monte Carlo Tree Search (MCTS) is initiated (from the state s_t) and plays a whole game until the end. The MCTS selects to expand the leaf of the search tree with the highest Q value plus an upper confidence bound U, which depends on the computed prior probability P and the visit count for that edge N. When the leaf node is expanded, the DNN computes the probabilities $\boldsymbol{p}$ and v for the new position s and the Q values of the tree are updated. At the end of the search the probabilities π_t ($\pi_t \propto N^{1/\tau}$, where τ is a parameter controlling temperature) are returned and the DNN updates its parameters (ϑ) according to them. The next move of the self-play game is computed based on π_t until the end of the game, at which the game winner z is computed and the DNN is again updated to minimize the error between the predicted winner at t and z.

Results: AlphaGo Zero was evaluated against its previous versions, which all used prior human knowledge and achieved superhuman performance against Go champions Fan Hui (AlphaGo Fan) and Lee Sedol (AlphaGo Lee). It was also tested against AlphaGo Master, the strongest Go AI at that moment. All AlphaGos were allowed 5s to pick a move. AlphaGo Zero and Master operated on a single machine with 4 Tensor Processing Units (TPUs), while AlphaGo Fan and Lee were distributed over 176 GPUs and 48 TPUs respectively. On an Elo scale, they achieved 5,185 (Zero), 4,858 (Master), 3,739 (Lee) and 3,144 (Fan) ratings. Finally, in a 100-game match with 2-h time controls, AlphaGo Zero won over AlphaGo Master by 89 games to 11.

Discussion: This approach demonstrates that it is possible to train an AI, starting *tabula rasa*, to achieve superhuman performance in a very computationally intensive task such as playing the game of Go.