Mastering the game of Go without human knowledge

David Silver, et al.[1]

Early in 2016, the game of Go, which (since Chess had fallen in the 1990's) had been long held as the last bastion of human superiority in the realm of game playing, succumbed to the onslaught of Artificial Intelligence when the player regarded as the best in the world, Lee Sedol, was defeated 4-1 in a match against Alpha Go. Alpha Go was a system designed by Google's AI research subsidiary Deep Mind, with the singular goal of achieving superhuman performance at the game of Go. To this end, the engineers provided Alpha Go with hundreds of thousands of example games from which to learn. Opening and endgame books were encoded into the system. Adjustments were made based on expert human advice. Enormous amount of time and processing power was utilized in running Alpha Go's learning algorithms. Ultimately, Alpha Go was able to convincingly defeat the human champion, but it was not without its flaws, and the training phase required vast amounts special handling to perform the supervised learning algorithms.

Now, less than two years later, Deep Mind has developed a system that, using Reinforcement Learning, is able to trounce Alpha Go and its successor, Alpha Go Master. This system is called Alpha Go Zero. If Alpha Go Zero had been fed exponentially more data, and had run on substantially more hardware, and had had many more months of training time, and had dozens of experts providing feedback and input, this would have been an impressive feat. This was not the case, however. Alpha Go Zero was provided with nothing more than the rules of the game, from which it learned enough about the game to defeat the version of Alpha Go that beat Lee Sedol (Alpha Go Lee) 100 games to 0. Alpha Go Zero ran on one machine with four Tensor Processing Units (TPUs), as opposed to many machines and 48 TPUs for Alpha Go Lee.

Alpha Go Zero started with random data, and used self-play with only black and white stones as input features into a single neural network (vis separate policy and value networks in Alpha Go Lee) and a tree search algorithm as part of its reinforcement learning system. As input, the neural network received the board state and its history. The output is a vector of move probabilities (the probability of selecting each move) and a scalar value representing the probability that the current player wins from the current position. At each of the positions, a Monte Carlo tree search is performed and the output probability used to improve the network's policy. The neural network's parameters are then updated so that the output matches the self-play prediction more closely, then the network's output is fed into the next round of self-play. It took only 72 hours of this training before Alpha Go Zero was able to win it's match against Alpha Go Lee 100-0. Interestingly, when Alpha Go Lee's gameplay was evaluated, the researchers found that the system learned many common strategies that are used by human players, and even developed new approaches that were previously unknown or unused.

The success of Alpha Go Zero is a milestone for reinforcement learning. The generality of the approach should lend itself to applications beyond just playing board games.

[1] https://www.nature.com/articles/nature24270.epdf?author_access_token=VJXbVjaSHxFoctQQ4p2k4tRgN0jAjWel9jnR3ZoTv0PVW4gB86EEpGqTRDtpIz-2rmo8-KG06gqVobU5NSCFeHILHcVFUeMsbvwS-lxjqQGg98faovwjxeTUgZAUMnRQ