A summary on
Mastering the game of Go without human knowledge

Early this year (2017), AlphaGo Master defeated the strongest human players in Go 60-0. In a few months time, Deepmind created another version of AlphaGo Zero which is more powerful than AlphaGo Master. This is a summary of the publication.

Unlike the predecessors, AlphaGo Zero learns without human knowledge. Instead of using two deep neural networks: a policy network that outputs move probabilities and a value network was trained initially by supervised learning to accurately predict the human expert moves, AlphaGo Zero uses a single neural network. AlphaGo Zero is trained solely by self-play reinforcement learning and started with random moves. It uses a new reinforcement learning algorithm that incorporates lookahead search inside training loop.

The new method combines the role of both policy network and value network. It is trained from self-play games by a novel reinforcement algorithm. In each position s, an MCTS search is executed and guided by the neural network. The MCTS can be viewed as a powerful policy improvement operator here because it usually selects much stronger moves than the raw moves of the neural network. When AlphaGo Zero finds a game winner, it would be sampled and can be view as a powerful policy evaluation operator. The main concept is to use these operators repeatedly in a policy procedure, the neural network's parameters are updated and the next iteration of search become even stronger.

Applying the reinforcement learning pipeline to train AlphaGo Zero, after 24h of training, it is able to beat a human trained computer player. After 36 h, AlphaGo Zero outperformed AlphaGo Lee which defeated Lee Sedol, the winner of 18 international titles. In this case, the AlphaGo Zero used a single machine with 4 tenso process units (TPUs), where AlphaGo Lee was distributed over many machines with 48 TPUs.

Deepmind later created a second instance of AlphaGo Zero using a larger neural network and over a longer period. After approximately 40 days, AlphaGo Zero was able to beat the strongest existing program AlphaGo Master and achieved an Elo (professional Go player rating) rating of 5,185, compared to 4,858 for AlphaGo Master which beat the strongest human player earlier this year 5-0.

AlphaGo Zero demonstrate that a pure reinforcement learning approach is fully feasible for the most challenging domains. Without human knowledge, it takes a few extra hours to train initially, but it exceeds human expert level eventually. Despite discovering the existing human knowledge of Go, Alpha Go Zero was able to find new variants that were previously unknown.