A short summary of “Mastering the game of Go with deep neural networks and tree search”

by Rene Kretzschmar

Intro

The game of Go has long been perceived as the most challenging classic game for Artificial Intelligence. It has a huge game tree, so exhaustive search is impossible and board positions a very difficult to evaluate. Until now it was not possible for a machine to beat a professional human player. And Team AlphaGo said: “Challenge accepted!”

Goal

The goal was to implement a Go computer programm *AlphaGo*, who will play significantly better than all other Go programms. And eventually to beat a professional human player - a goal that was assument to be in reach 10 years from now - even with the rise of deep learning in the recent years.

Technics used

Playing

AlphaGo combines an MCTS algorithm with a policy neural network for move selection and a value neural network for board position evaluation. Each edge of the game tree accumulates an action value and a visit count when passed. The algorithm then selects actions with the highest action value, but indirect proportional to the visit count to encourage exploring.

Deep Neural Networks

The two neural networks have almost the same structure, 13 layers that alternate between convolutional layers with weights and rectifier nonlinearities. The policy network outputs a probability distribution over all legal moves of a cetain board position. The value network outputs a single prediction of the value of a certain board position.

Training

The team build a training pipeline consisting of three stages. Stage one is a supervised learning of the policy network. It learns from 30 mio pairs of board state / expert moves from the KGS Go Server and uses stochastic gradient ascent to maximize the likelyhood of a human move selected in a certain state. Stage two is a reinforcement learning of the policy network by playing games against the current version of the policy network and randomly selected previous iterations of itself. The final stage is a reinforcement learning of a value network from a generated self-play data set of 30 million distinct positions, each sampled from a different game.

Hardware

The standalone version of AlphaGo runs 40 search threads on 48 CPUs and the two neural networks on 8 GPUs. The distributed version also runs 40 search threads on 1202 CPUs and the neural networks on 176 GPUs.

Results

AlphaGo is many *dan* ranks stronger than any previous Go program with a winning rate of 99.8% even in handycapped games. The distributed version was significant stronger, winning 77% of the games against the standalone version and 100% against all other Go programs. Even without the policy network and only value network AlphaGo performed better than the other programms.

Finally the distributed version of AlphaGo won 5 games to 0 against Fan Hui, a professional 2 *dan* and the winner of the European Go championships in 2013, 2014 and 2015.

Goal reached. 10 years earlier than assumed.