Advance Game Playing

Research Review

As part of the Advance Game Playing project, the Research Review of the "Mastering the game of Go with deep neural networks and tree search" article is presented in this one-page summary.

The main goals and new techniques will be described first followed by the results presented on the article itself.

Original Article:
Mastering the game of Go with deep neural networks and tree search
Author of Research Review:
Carlos Conde Pereda

Goal

The game of Go is considered one of the most challenging games of classic game for artificial intelligence mostly due to its high branching factor b^d where $b \approx 250$ and $d \approx 150$ making exhaustive search infeasible.

The main goal of AlphaGo is to obtain a search algorithm that overcomes this limitation by reducing the search tree by:

- Reducing the depth of search with position evaluation using value networks $v\theta(s)$
- Reducing the breadth of search with a probability distribution of the legal moves obtained by sampling actions from a policy network $p(a|s)$

Combining efficient value and policy networks with existing **MCTS**, the goal is to find an efficient search algorithm capable of beating consistently current state-of-the-art algorithms

Pipeline

First Stage ($p\theta$)

The first stage consist on generating a probability distribution of legal moves that can be used to choose the next move. Two policy networks are used for this purpose:

1. $p\theta$ policy obtained by training a supervised learning (SL) policy network directly from expert human moves alternating between convolutional layers with weights θ and rectifier nonlinearities
2. $p\pi$ policy obtained using a linear softmax of small pattern features with weights π

By doing this, the $p\theta$ is capable of predicting moves with an accuracy between 55~57% (compared to the 44.4% of other research groups) in 3ms and $p\pi$ obtains an accuracy of 24.2% in 2μs.

Second Stage ($p\rho$)

The goal of this stage is to improve the policy network by gradient reinforcement leaning (RL) to obtain $p\rho$. This is accomplished by playing games between the current $p\rho$ and randomly selected previous iteration of the policy network.

By doing this, the RL policy network is capable of winning more than 80% of the games against the SL policy network, and 85% of the games against *Pachi*, a sophisticated Monte Carlo search program ranked at 2 amateur dan on KGS.

Third Stage ($v\theta$)

The focus in the last stage is the position evaluation estimating a value function $v^p(s)$ that predicts the outcome from position s by using the policy network p. For this purpose the strongest policy from previous stages is used, in this case the RL based $p\rho$.

The structure is similar to that of the policy networks, but outputs a single prediction instead of a probability distribution.

Searching

The proposed search algorithm combines **MCTS** with all the policy and value networks discussed:

1. Tree edges store action value, visit count and prior probability ($Q(s,a)$, $N(s,a)$ and $P(s,a)$)
2. At each step t, an action a_t is selected for state s_t based on the action value plus a bonus proportional to the prior probability and that decays with repeated visits
3. At the leaf node at level L, $p\theta$ is processed once and new probabilities are computed
4. Leaf node is evaluated with a combination of $v\theta$ and $p\pi$:

$$V(s_L) = (1-\lambda)v_\theta(s_L) + \lambda z_L$$

5. At the end of simulation the algorithm chooses the most visited move

Results

- Note that $p\theta$ performs better on *AlphaGo* than $p\rho$ presumably because humans select a diverse beam of promising moves, whereas RL optimizes for the single best move
- Value function $v\theta$ performs better with $p\rho$
- On a tournament against the most used Go programs (*Crazy Stone, Zen, Pachi, …*), *AlphaGo* wins 494 out of 495 games (99.8%)
- With 4 handicapped stones, *AlphaGo* wins between 77% and 99% of the games
- AlphaGo was also evaluated with variants of the position evaluation showing that the best performance is obtained with the mixed valuation ($\lambda=0.5$), winning ≥95% against other variants