Master the Game of Go without Human Knowledge

Jcykcai

Key technical Contributions

- It uses a single neural network, rather than separate policy and value network.
- A new reinforcement learning algorithm that incorporates lookahead search inside the training loop.

Policy&Value Network

- $(\mathbf{p}, v) = f_\theta(s)$
- $\mathbf{p}$ is the vector of move probabilities
- v is a scalar evaluation, estimating the probability of the current player winning from position s

Monte Carlo Tree Search (MCTS)

- In each position s, a MCTS is executed, guided by the neural network.

- The MCTS outputs probabilities of π of playing each move, which select much stronger moves than raw move probabilities $\mathbf{p}$

Monte Carlo Tree Search (MCTS)

- Each node s in search tree contains edges (s, a) for all legal actions $a \in A(s)$
- Each edge stores a set of statistics $\{N(s, a), W(s, a), Q(s, a), P(s, a)\}$
- $N(s, a)$ is the visit count, $W(s, a)$ is the total action value, $Q(s, a)$ is mean actions value and $P(s, a)$ is the prior probability.

Monte Carlo Tree Search (MCTS)

- At each non-leaf node, an action is selected according to

$$a_t = \underset{a}{\operatorname{argmax}}(Q(s_t, a) + U(s_t, a))$$

$$U(s, a) = c_{\text{puct}} P(s, a) \frac{\sqrt{\sum_b N(s, b)}}{1 + N(s, a)}$$

- where c_{puct} is a constant determining the level of exploration; this search control strategy initially prefers actions with high prior probability and low visit count, but asymptotically prefers actions with high action value.

Monte Carlo Tree Search (MCTS)

- For each leaf node s_L
- Evaluate $(d_i(\mathbf{p}), v) = f_\theta(d_i(s_L))$,where d_i is a dihedral reflection or rotation selected uniformly at random from *i* in [1..8]
- Backup, through each step $t < L$

$$N(s_t, a_t) += 1, W(s_t, a_t) = W(s_t, a_t) + v, Q(s_t, a_t) = \frac{W(s_t, a_t)}{N(s_t, a_t)}$$

Monte Carlo Tree Search (MCTS)

- After MCTS, the AlphaGo Zero selects a move *a* to play by a policy π , where $\pi_a \propto N(s,a)^{\frac{1}{\tau}}$

Training

- The neural network is trained by self-play game that uses MCTS to play each move

- A game terminates at step *T* when both players pass, when the search value drops below a resignation threshold or when the game exceeds a maximum length.

- The game is scored a final reward of $r_T \in \{-1, +1\}$

Training

- Collect triples $(s_t, \pi_t, z_t), z_t = \{ + / - \} r_T$
- Loss function $(\boldsymbol{p}, v) = f_\theta(s)$ and $l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$