22 April 2019

ZUBKOV MAKSIM, DCAM 777

MASTERING THE GAME OF GO WITH DEEP NEURAL NETWORKS AND TREE SEARCH

AGENDA

- Problem Setting
- MCTS
- Learning process
- Classification policy network
- Reinforcement learning policy network
- Regression value network
- MCTS_Tic-Tac-Toe implementation

PROBLEM SETTING

- Go is a Markov game – a game of perfect information
- States space S
- Action space $\mathscr{A}(s)$
- Value function v^*
- Reward $r^i(s) : r^1(s) = -r^2(s)$
- Transaction function $f(s, a)$
- Outcome $z_t = \pm r(s_T)$

PROBLEM SETTING

- Policy – probability distribution over legal actions $p(a \mid s)$
- Expected outcome if all actions for both players are selected according to policy $v^p(s)$
- Most of games are to

$v^*(s)$ – the outcome of the game

from every state s

under perfect play by all players

$$v^*(s) = \begin{cases} z_T, \text{ if } s = s_T \\ \max_a - v^*(f(s,a)), \text{ othervise} \end{cases}$$

$$v^p(s) = \mathbb{E}\left[z_t \; s_t = s, a_{t\ldots T} \sim p\right]$$

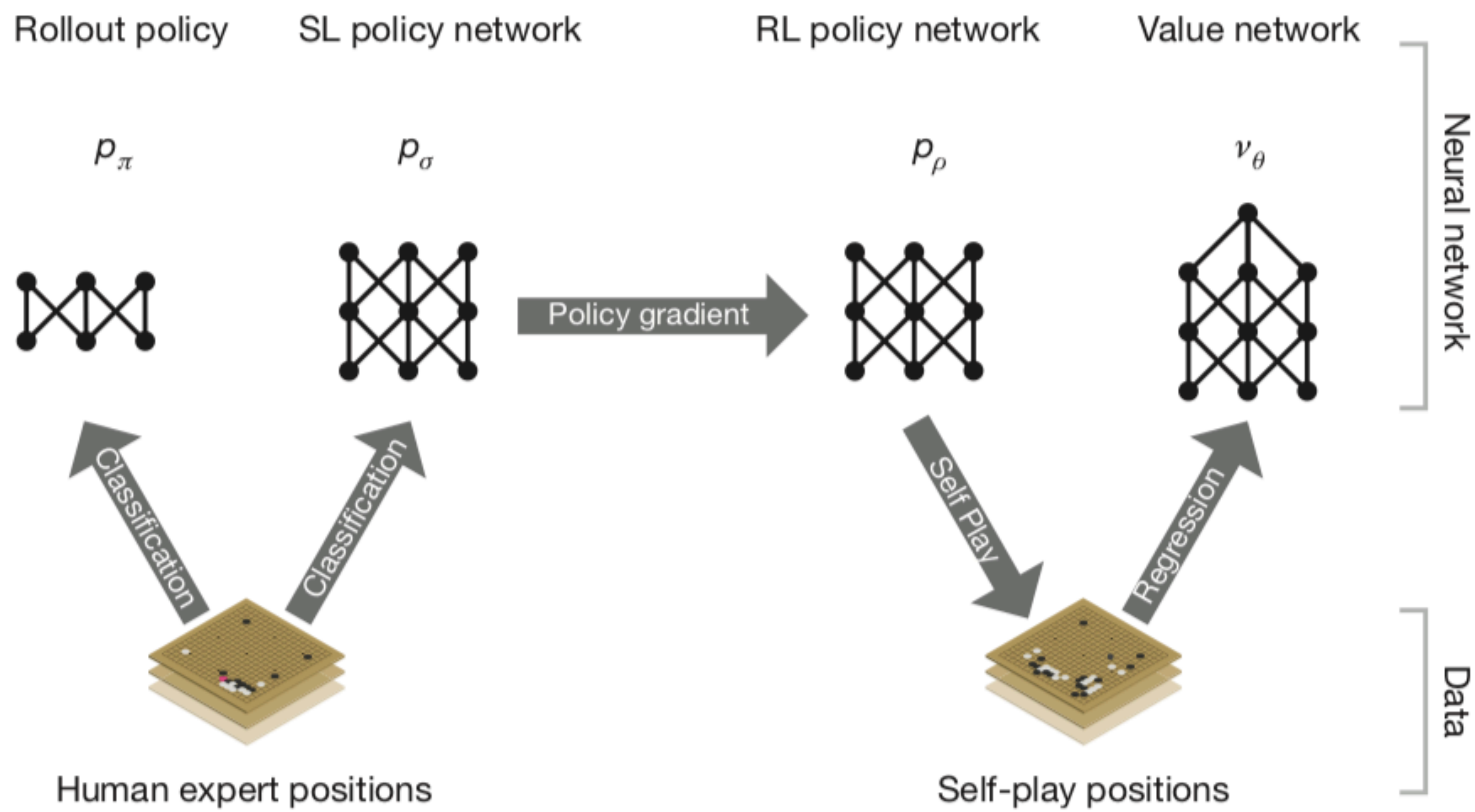
MCTS

- Prior probability $P(s, a)$
- MC estimates of total action value $W_v(s, a); W_r(s, a)$
- Number of evaluations and rollout rewards $N_v(s, a); N_r(s, a)$
- Combined mean action value for edge $Q(s, a)$

- Selection
- Evaluation
- Backup
- Expansion

LEARNING PROCESS

- SL policy network p_σ
- Fast policy that can rapidly p_π
- RL policy network p_ρ



CLASSIFICATION POLICY NETWORK

- Random selected mini-batch $\{s^k, a^k\}_{k=1}^{m}$
- Predict the winner of games played by the RL policy network against itself v_θ

$$\Delta\sigma = \frac{\alpha}{m}\sum_{k=1}^{m}\frac{\partial \log p_\sigma(a^k \mid s^k)}{\partial \sigma}$$

REINFORCEMENT LEARNING POLICY NETWORK

- n games
- Playing until termination on T^i step
- Predict the winner of games played by the RL policy network against itself

$$z_t^i = \pm r(s^{T^i})$$

$$\Delta\rho = \frac{\alpha}{n} \sum_{i=1}^{n} \sum_{t=1}^{T^i} \frac{\partial \log p_\rho(a_t^i \mid s_t^i)}{\partial \rho} (z_t^i - v(s_t^i))$$

VALUE NETWORK REGRESSION

Random sampling U

Sampling $t = 1...U - 1$ moves from SL policy network

$a_t \sim p_\sigma(\cdot \mid s_t)$

Semple one move random from all avaliable moves a_U

Then sampling moves until the end, $t = U + 1..T$

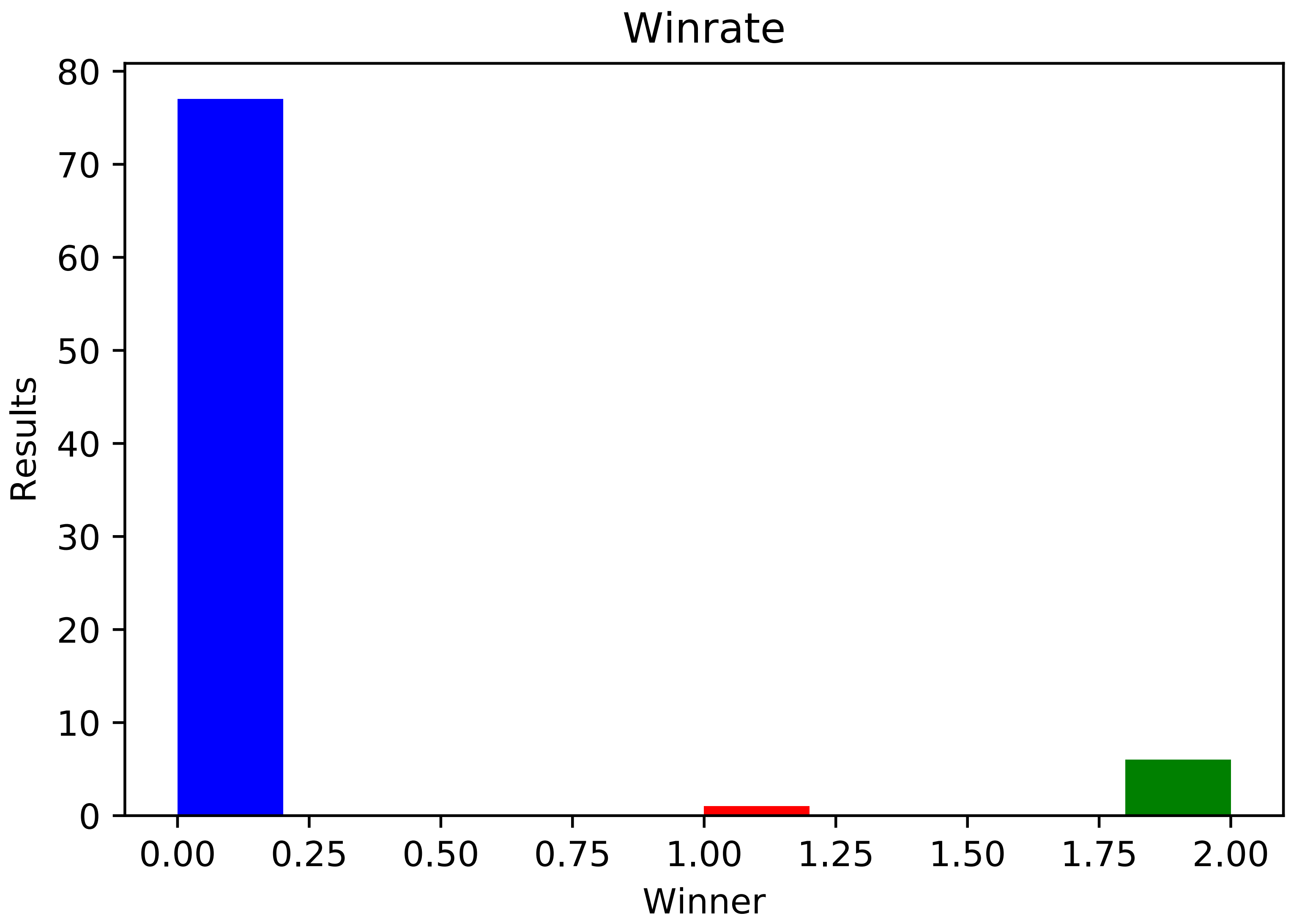
Finally get $z_t = \pm r(s_T)$

Then semple $v^{p_\rho} = \mathbb{E}\left[z_{U+1}, a_{U+1...T} \sim p_\rho\right]$

$$\Delta\theta = \frac{\alpha}{m} \sum_{k=1}^{m} \frac{\partial v_\theta(s^k)}{\partial \theta}(z^k - v_\theta(s^k))$$

TIC-TAC-TOE

- Blue - draw
- Red - player win
- Green - computer win



THANK YOU FOR LISTENING!