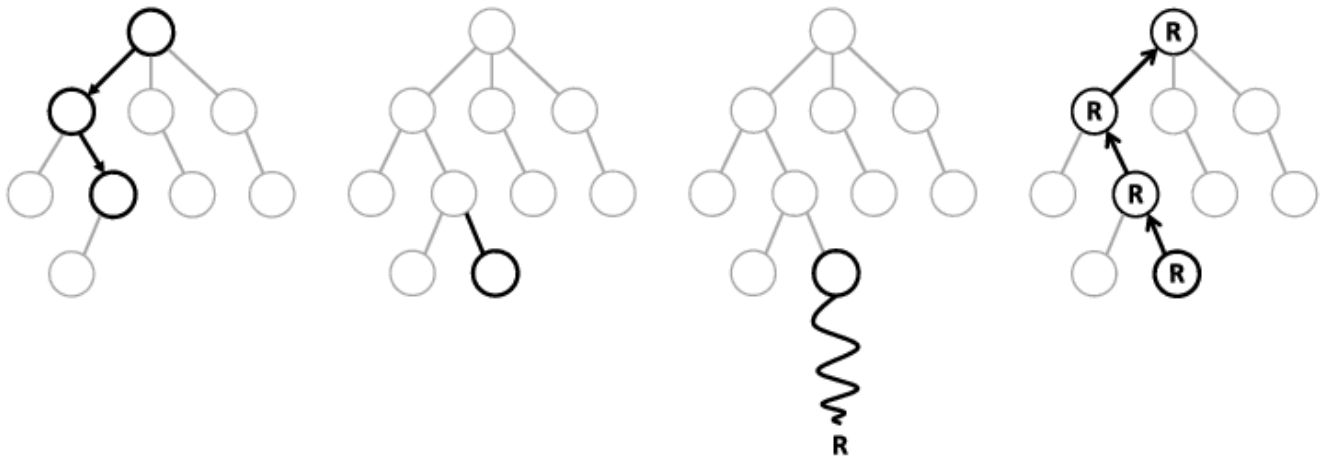
AlphaZero и MuZero

Романов Владимир БПМИ192

Национальный исследовательский университет
«Высшая школа экономики» (Москва)

14 мая 2022 г.

Monte Carlo Tree Search



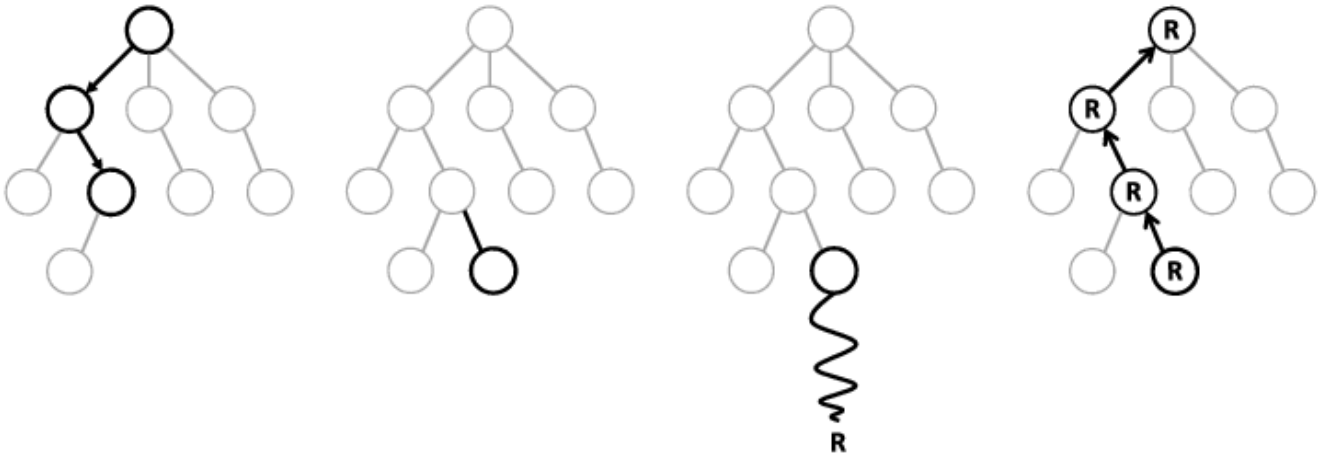
(a) Selection (b) Expansion (c) Simulation (d) Backpropagation

- В каждой вершине мы храним пару $(\Sigma, n(s))$
- Σ — суммарная награда в поддереве, $n(s)$ — число посещений
- $V(s) = \dfrac{\Sigma}{n(s)}$ — оценка награды для вершины
- $W(s) = V(s) + c\sqrt{\dfrac{\log n^{parent}(s)}{n(s)}}, \quad \pi(s) = \arg\max\limits_{a} W(f(s,a))$

Minimax MCTS

- В случае двух игроков будем использовать self-play
- Введем отдельную переменную o, равную 1 если ходит первый, иначе -1
- Теперь $W(s) = oV(s) + c\sqrt{\frac{\log n^{parent}(s)}{n(s)}}$

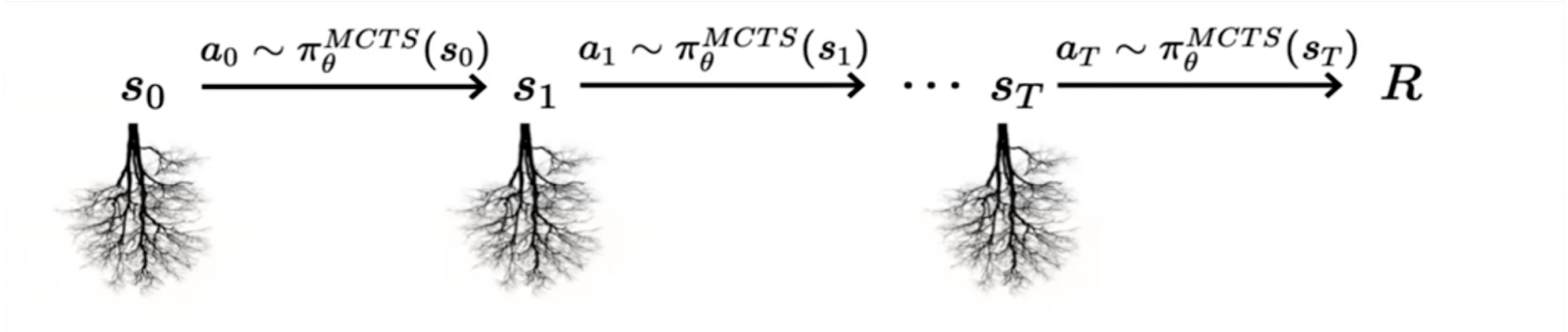
MCTS в AlphaZero



(a) Selection (b) Expansion (c) Simulation (d) Backpropagation

- Пусть у нас есть π_θ и V_θ
- При оценке значения $V(s) = \lambda V_\theta(s) + (1-\lambda)\hat{V}(s)$, где $\hat{V}(s)$ мы получаем с помощью симуляции политики π_θ
- $W(s) = V(s) + c\dfrac{\pi_\theta(a|s_{parent})}{1+n(s)}$

Обучение нейросети в AlphaZero



- Теперь $\pi_\theta^{MCTS} \propto n(f(s,a))^{\frac{1}{\tau}}$
- Сохраняем тройки $(s_t, \pi_\theta^{MCTS}(s_t), R)$
- Минимизируем $(R - V_\theta(s))^2 - (\pi_\theta^{MCTS}(s))^T \log \pi_\theta(s) + \kappa\|\theta\|_2^2$

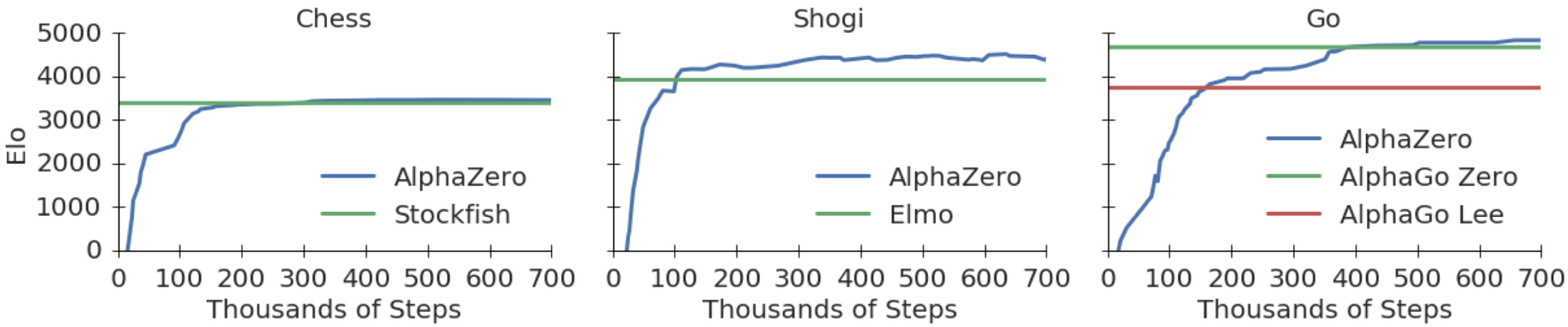
Как кодировать состояние

Go		Chess		Shogi	
Feature	Planes	Feature	Planes	Feature	Planes
P1 stone	1	P1 piece	6	P1 piece	14
P2 stone	1	P2 piece	6	P2 piece	14
		Repetitions	2	Repetitions	3
				P1 prisoner count	7
				P2 prisoner count	7
Colour	1	Colour	1	Colour	1
		Total move count	1	Total move count	1
		P1 castling	2		
		P2 castling	2		
		No-progress count	1		
Total	17	Total	119	Total	362

Table S1: Input features used by *AlphaZero* in Go, Chess and Shogi respectively. The first set of features are repeated for each position in a $T = 8$-step history. Counts are represented by a single real-valued input; other input features are represented by a one-hot encoding using the specified number of binary input planes. The current player is denoted by P1 and the opponent by P2.

На вход нейросети подается тензон размера $N \times N \times (MT + L)$, представляющий из себя набор из $MT + L$ игровых полей.

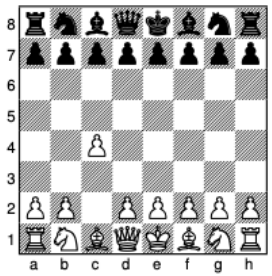
Сравнение

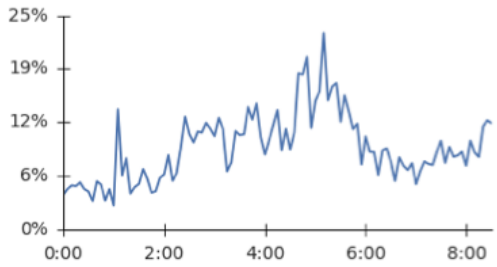


Game	White	Black	Win	Draw	Loss
Chess	*AlphaZero*	*Stockfish*	25	25	0
	Stockfish	*AlphaZero*	3	47	0
Shogi	*AlphaZero*	*Elmo*	43	2	5
	Elmo	*AlphaZero*	47	0	3
Go	*AlphaZero*	*AG0 3-day*	31	–	19
	AG0 3-day	*AlphaZero*	29	–	21

Дебюты

A10: English Opening

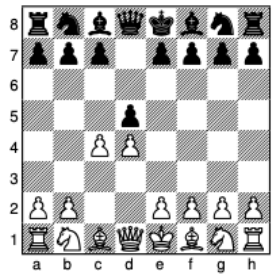


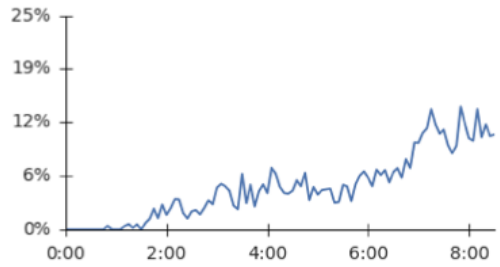


w 20/30/0, b 8/40/2

1...e5 g3 d5 cxd5 ♘f6 ♗g2 ♘xd5 ♘f3

D06: Queens Gambit

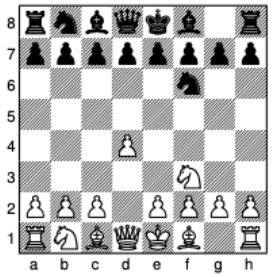


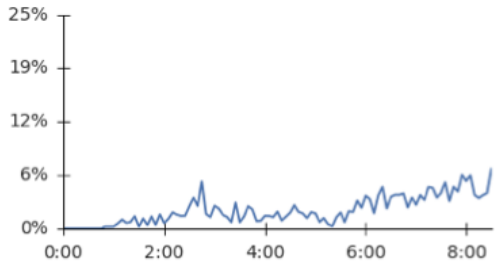


w 16/34/0, b 1/47/2

2...c6 ♘c3 ♘f6 ♘f3 a6 g3 c4 a4

A46: Queens Pawn Game

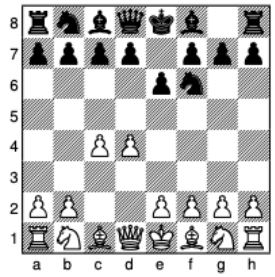


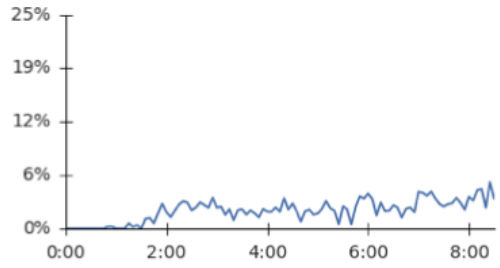


w 24/26/0, b 3/47/0

2...d5 c4 e6 ♘c3 ♗e7 ♗f4 O-O e3

E00: Queens Pawn Game

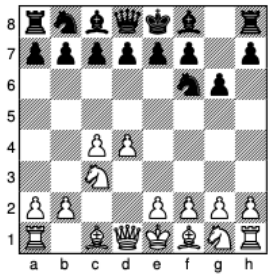


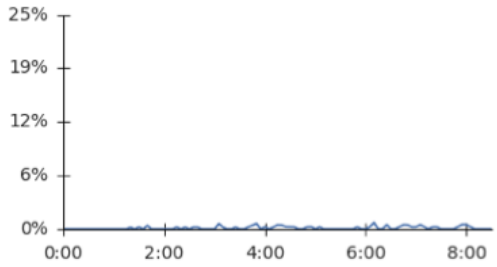


w 17/33/0, b 5/44/1

3.♘f3 d5 ♘c3 ♗b4 ♗g5 h6 ♕a4 ♘c6

E61: Kings Indian Defence

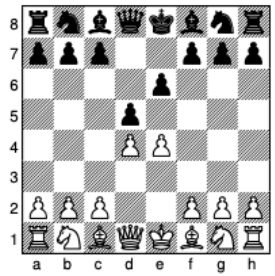


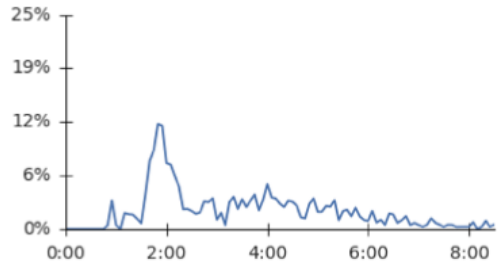


w 16/34/0, b 0/48/2

3...d5 cxd5 ♘xd5 e4 ♘xc3 bxc3 ♗g7 ♗e3

C00: French Defence





w 39/11/0, b 4/46/0

3.♘c3 ♘f6 e5 ♘d7 f4 c5 ♘f3 ♗e7

Что если неизвестна динамика?

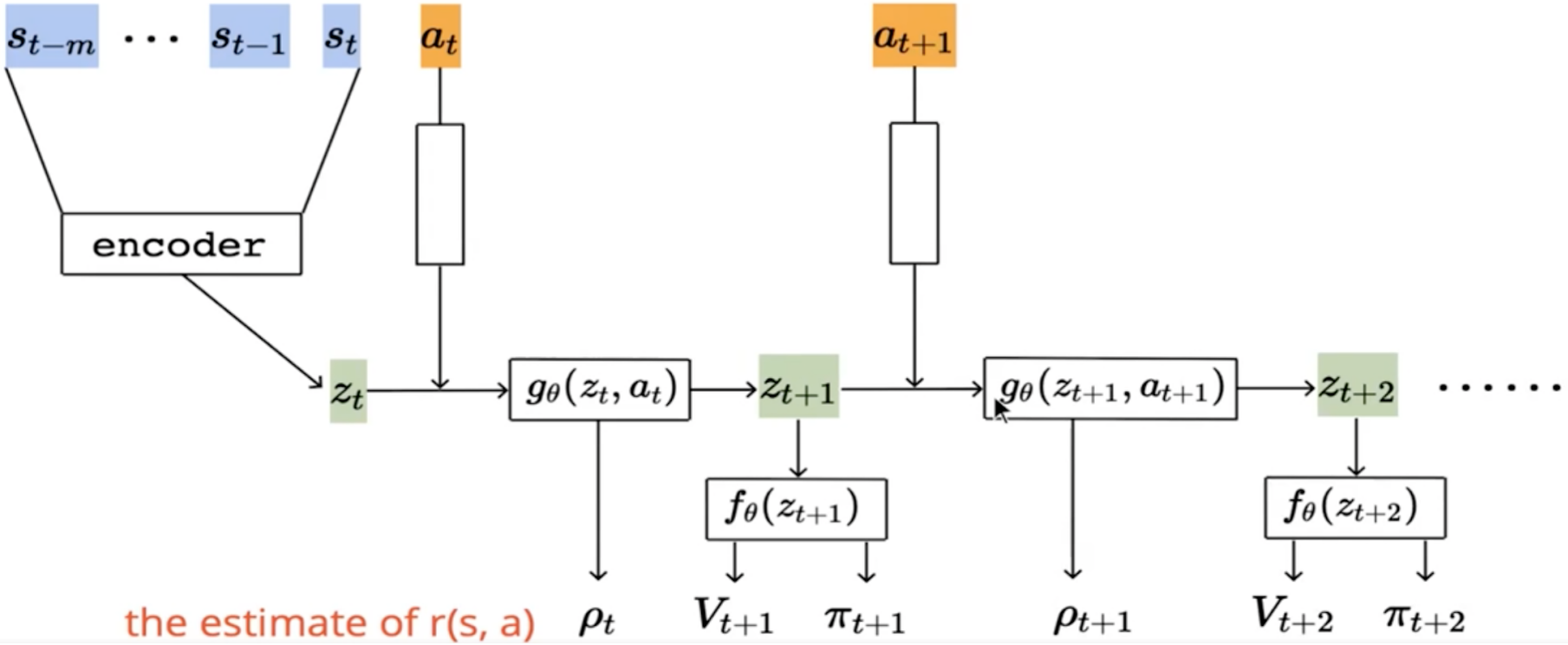
В AlphaZero у нас была динамика:

$$s_{root}, a_0, \ldots, a_L \xrightarrow{\text{dynamics}} s_L \xrightarrow[\text{and policy}]{\text{get value}} V_\theta(s_L), \pi_\theta(s_L) = f_\theta(s_L)$$

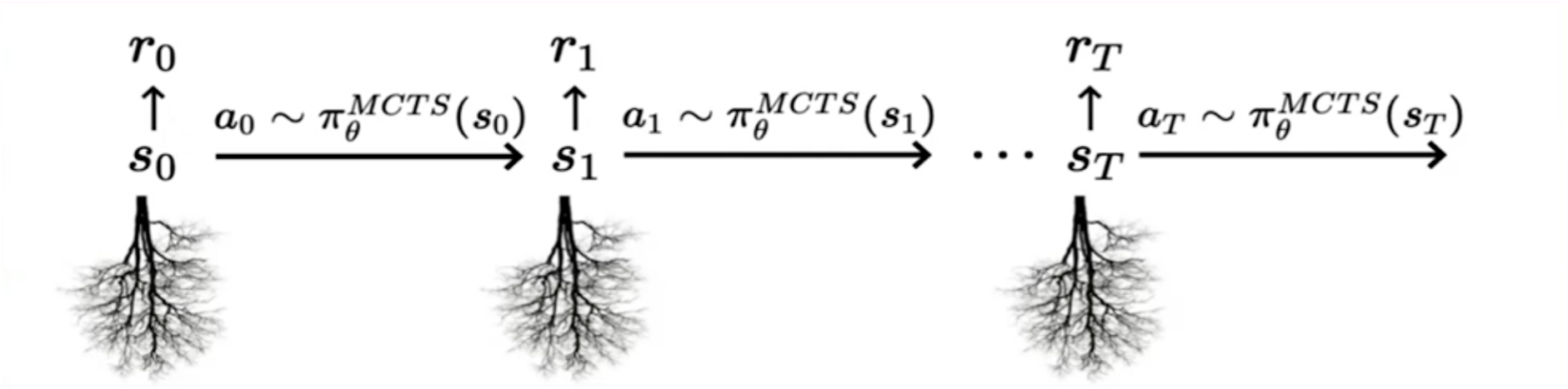
В MuZero предсказываются состояние, награды и политка:

$$s_{root}, a_0, \ldots, a_L \xrightarrow[\text{and policy}\\ \text{of future}\\ \text{states}]{\text{get value}} V_\theta(s_L), \pi_\theta(s_L) = f_\theta(s_L)$$

Архитектура MuZero

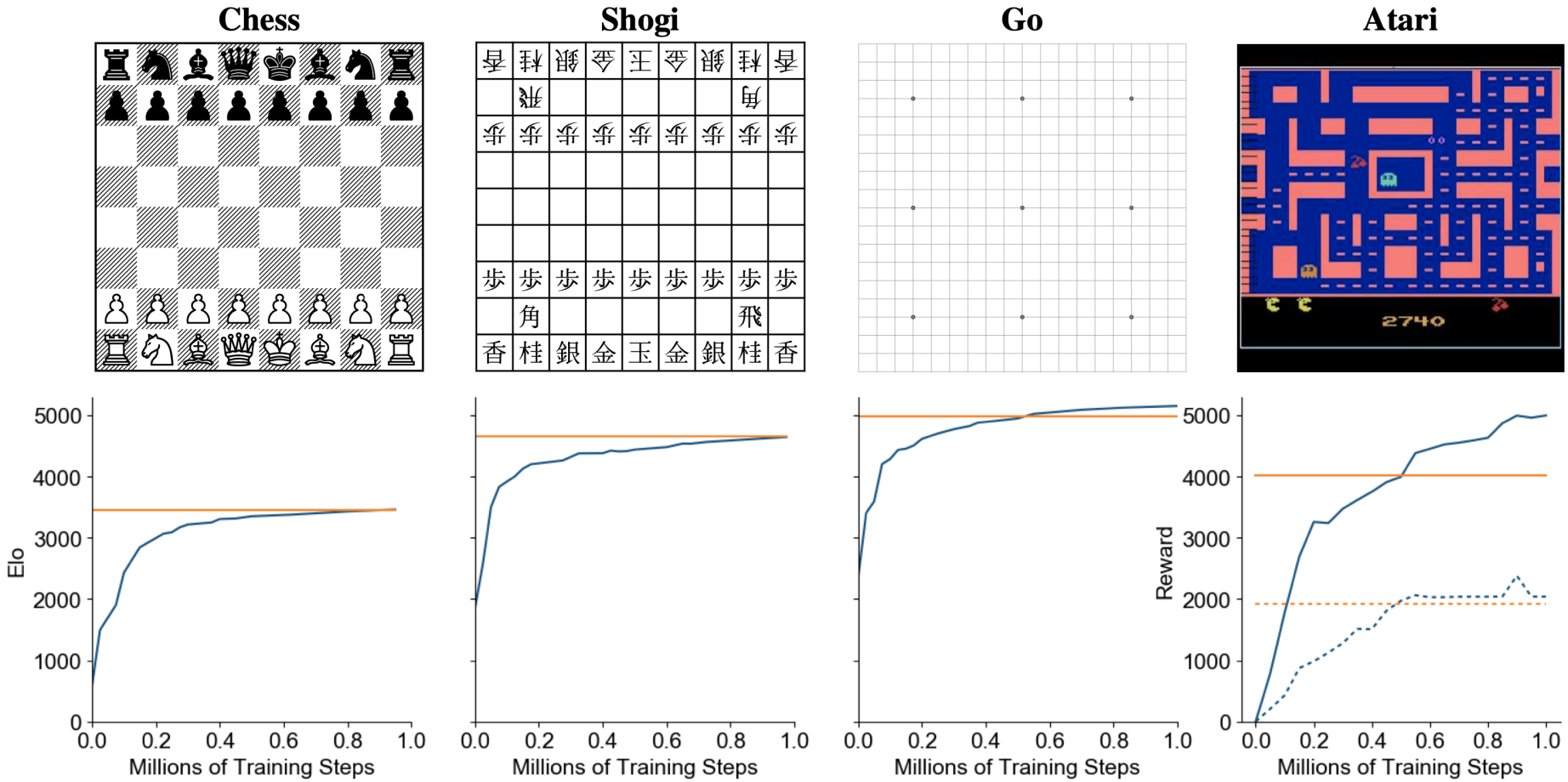


Обучение MuZero



- Теперь сохраняем не тройки, а всю игру:
  $(\ldots, s_t, a_t, \pi_t^{MCTS}, r_t, u_t, s_{t+1}, \ldots)$
- $v_k, \pi_k, \rho_k = NN(s_i, a_i, \ldots, a_{i+k})$
- Минимизируем
  $$\sum_{k=0}^{K} \left((u_{i+k} - v_k)^2 + (r_{i+k} - \rho_k)^2 - (\pi_{i+k}^{MCTS})^T \log \pi_k \right) + \kappa \|\theta\|_2^2$$

Сравнение


Chess
Shogi
Go
Atari
2740
Elo
Reward
Millions of Training Steps
Millions of Training Steps
Millions of Training Steps
Millions of Training Steps

Сравнение

Agent	Median	Mean	Env. Frames	Training Time	Training Steps
Ape-X [18]	434.1%	1695.6%	22.8B	5 days	8.64M
R2D2 [21]	1920.6%	4024.9%	37.5B	5 days	2.16M
MuZero	**2041.1%**	**4999.2%**	20.0B	12 hours	1M
IMPALA [9]	191.8%	957.6%	200M	–	–
Rainbow [17]	231.1%	–	200M	10 days	–
UNREAL[a] [19]	250%[a]	880%[a]	250M	–	–
LASER [36]	431%	–	200M	–	–
MuZero Reanalyze	**731.1%**	**2168.9%**	200M	12 hours	1M

Сравнение

Game	Random	Human	SimPLe [20]	Ape-X [18]	R2D2 [21]	*MuZero*	*MuZero* normalized
alien	227.75	7,127.80	616.90	40,805.00	229,496.90	**741,812.63**	10,747.5 %
amidar	5.77	1,719.53	74.30	8,659.00	**29,321.40**	28,634.39	1,670.5 %
assault	222.39	742.00	527.20	24,559.00	108,197.00	**143,972.03**	27,664.9 %
asterix	210.00	8,503.33	1,128.30	313,305.00	**999,153.30**	998,425.00	12,036.4 %
asteroids	719.10	47,388.67	793.60	155,495.00	357,867.70	**678,558.64**	1,452.4 %
atlantis	12,850.00	29,028.13	20,992.50	944,498.00	1,620,764.00	**1,674,767.20**	10,272.6 %
bank heist	14.20	753.13	34.20	1,716.00	**24,235.90**	1,278.98	171.2 %
battle zone	2,360.00	37,187.50	4,031.20	98,895.00	751,880.00	**848,623.00**	2,429.9 %
beam rider	363.88	16,926.53	621.60	63,305.00	188,257.40	**454,993.53**	2,744.9 %
berzerk	123.65	2,630.42	-	57,197.00	53,318.70	**85,932.60**	3,423.1 %
bowling	23.11	160.73	30.00	18.00	219.50	**260.13**	172.2 %
boxing	0.05	12.06	7.80	**100.00**	98.50	**100.00**	832.2 %
breakout	1.72	30.47	16.40	801.00	837.70	**864.00**	2,999.2 %
centipede	2,090.87	12,017.04	-	12,974.00	599,140.30	**1,159,049.27**	11,655.6 %
chopper command	811.00	7,387.80	979.40	721,851.00	986,652.00	**991,039.70**	15,056.4 %
crazy climber	10,780.50	35,829.41	62,583.60	320,426.00	366,690.70	**458,315.40**	1,786.6 %
defender	2,874.50	18,688.89	-	411,944.00	665,792.00	**839,642.95**	5,291.2 %
demon attack	152.07	1,971.00	208.10	133,086.00	140,002.30	**143,964.26**	7,906.4 %
double dunk	-18.55	-16.40	-	**24.00**	23.70	23.94	1,976.3 %
enduro	0.00	860.53	-	2,177.00	2,372.70	**2,382.44**	276.9 %
fishing derby	-91.71	-38.80	-90.70	44.00	85.80	**91.16**	345.6 %
freeway	0.01	29.60	16.70	**34.00**	32.50	33.03	111.6 %
frostbite	65.20	4,334.67	236.90	9,329.00	315,456.40	**631,378.53**	14,786.7 %
gopher	257.60	2,412.50	596.80	120,501.00	124,776.30	**130,345.58**	6,036.8 %
gravitar	173.00	3,351.43	173.40	1,599.00	**15,680.70**	6,682.70	204.8 %
hero	1,026.97	30,826.38	2,656.60	31,656.00	39,537.10	**49,244.11**	161.8 %

Список литературы

- AlphaZero: `https://arxiv.org/pdf/1712.01815.pdf`
- MuZero: `https://arxiv.org/pdf/1911.08265.pdf`