AlphaGo

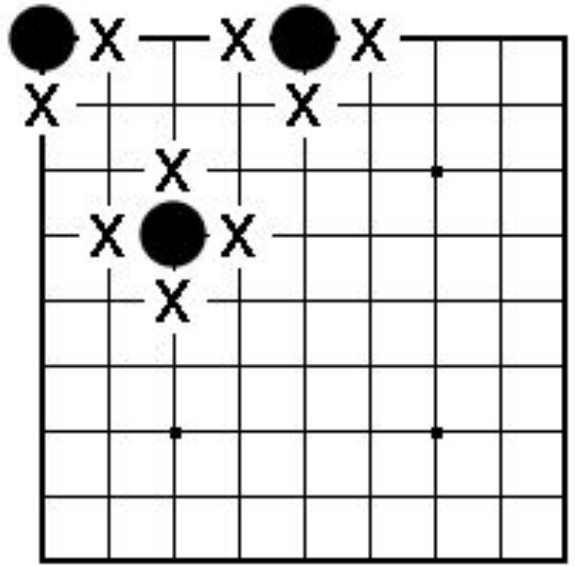
Шабалин Евгений
21.02.2020

Что такое ГО

Что такое ГО

- Логическая настольная игра родом из Китая
- Поле 19x19
- Конечное число позиций

Что такое ГО

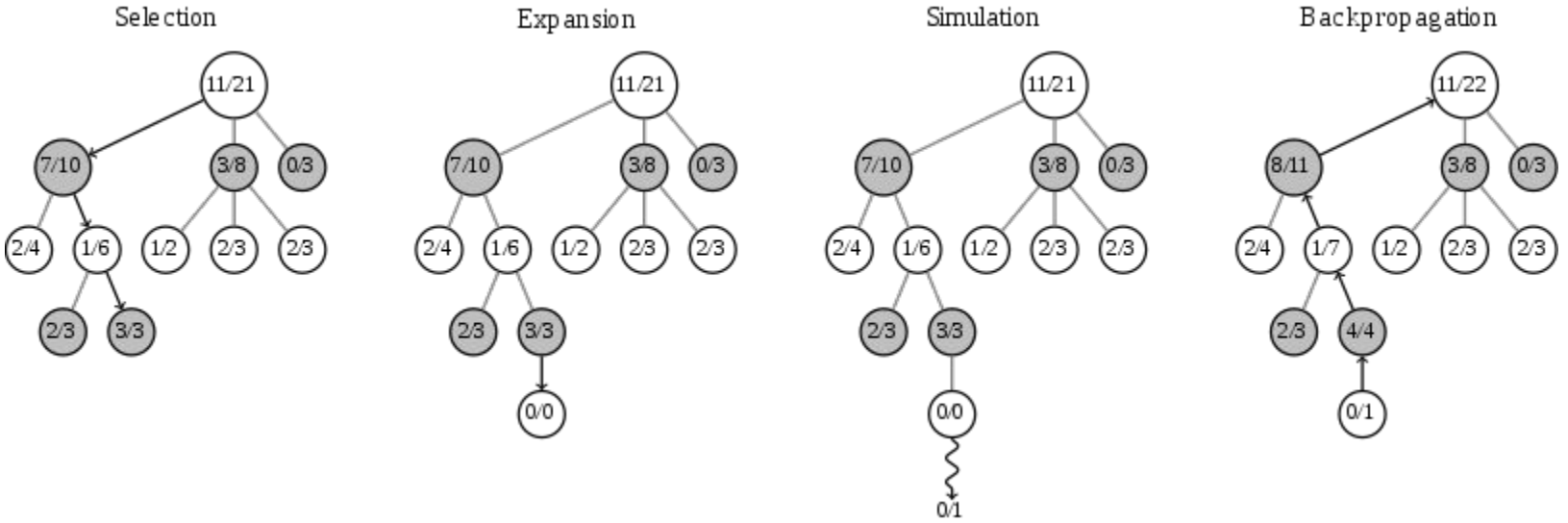


Почему Го настолько сложная?

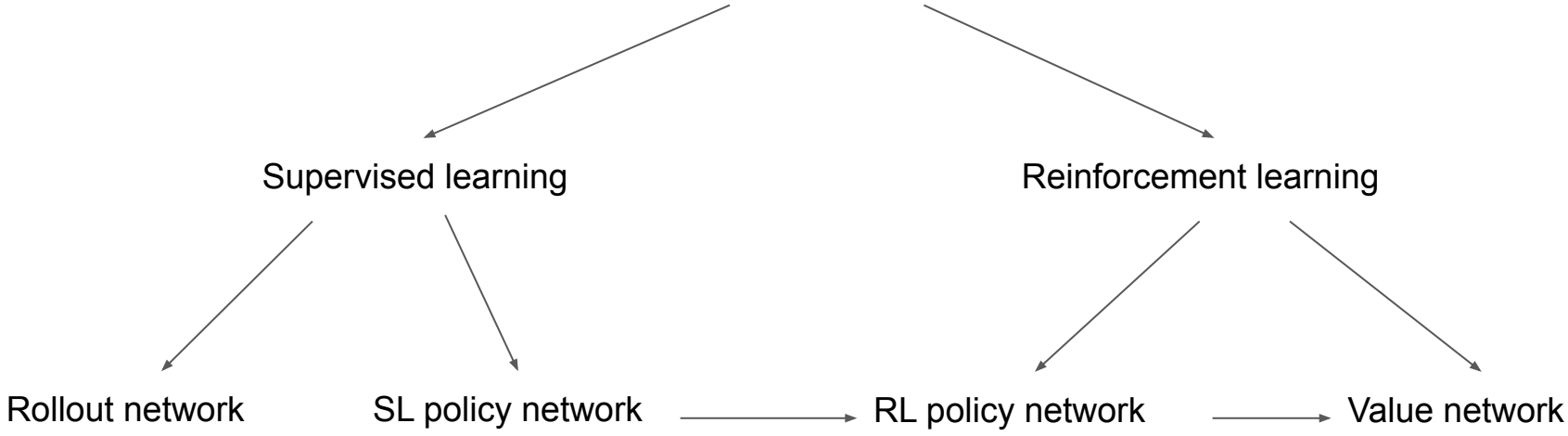
Почему Го настолько сложная?

- 19 * 19 = 361 различных позиций
- Около 150 адекватных ходов на каждой стадии
- Очень сложно понять, хороший ход или нет
- Продолжительность игры 150-250 ходов

Monte Carlo tree search



AlphaGo


Supervised learning
Reinforcement learning
Rollout network
SL policy network
RL policy network
Value network

Supervised learning

Возьмем за основу игры профессионалов и будем пытаться предугадать их ход. Обучим 2 сетки: Rollout policy и SL policy network.

Rollout policy: очень быстрая сеть, которая берет линейную комбинацию большого количества признаков и софтмаксом превращает все в вероятности.

SL policy network: медленная, но более точная (57%) сеть, состоящая из 13 сверточных слоев по 192 фильтра в каждом

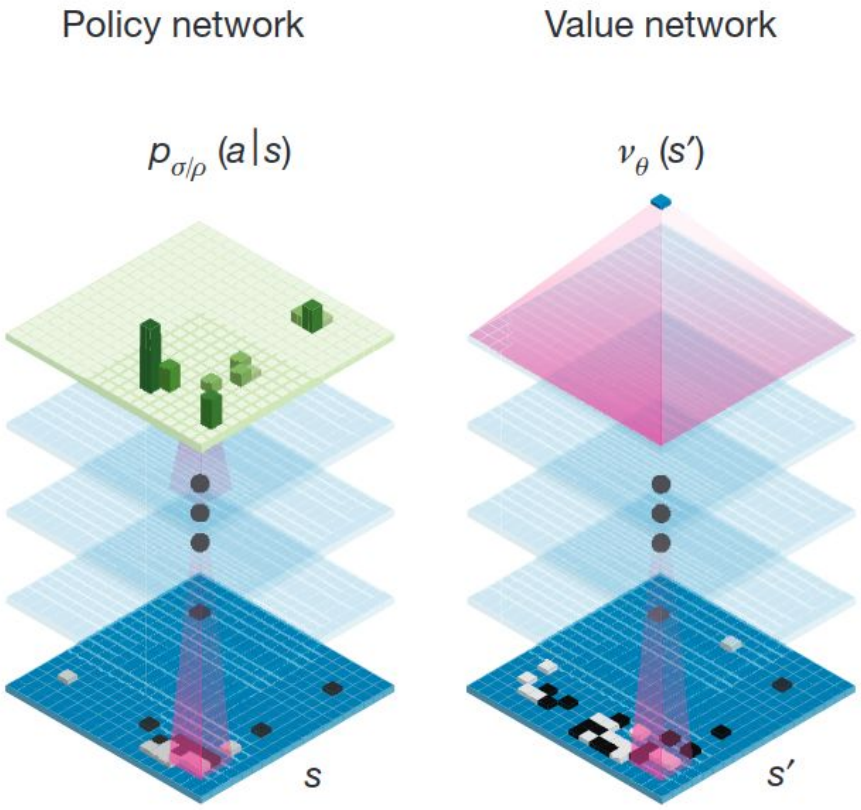
Используемые признаки

Feature	# of patterns	Description
Response	1	Whether move matches one or more response pattern features
Save atari	1	Move saves stone(s) from capture
Neighbour	8	Move is 8-connected to previous move
Nakade	8192	Move matches a *nakade* pattern at captured stone
Response pattern	32207	Move matches 12-point diamond pattern near previous move
Non-response pattern	69338	Move matches 3×3 pattern around move

Используемые признаки

Feature	# of planes	Description
Stone colour	3	Player stone / opponent stone / empty
Ones	1	A constant plane filled with 1
Turns since	8	How many turns since a move was played
Liberties	8	Number of liberties (empty adjacent points)
Capture size	8	How many opponent stones would be captured
Self-atari size	8	How many of own stones would be captured
Liberties after move	8	Number of liberties after this move is played
Ladder capture	1	Whether a move at this point is a successful ladder capture
Ladder escape	1	Whether a move at this point is a successful ladder escape
Sensibleness	1	Whether a move is legal and does not fill its own eyes
Zeros	1	A constant plane filled with 0
Player color	1	Whether current player is black

Reinforcement learning



RL policy network

Обучение происходит следующим образом:

- Текущая версия сети играет со случайно выбранной предыдущей итерацией сети
- Веса меняются по обычной формуле policy gradients

$$\Delta\rho \propto \frac{\partial \log p_\rho(a_t \mid s_t)}{\partial \rho} z_t$$

z_t — результат партии (± 1)

Value network

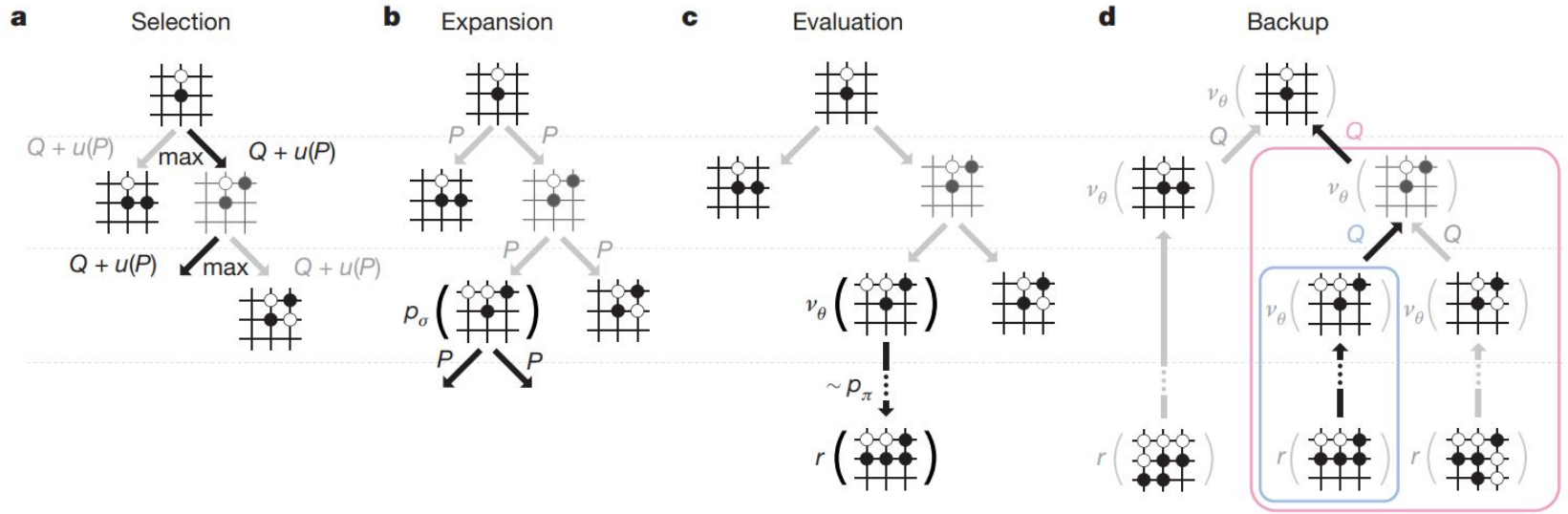
Вместо распределения вероятностей выдает одно число от -1 до 1 — результат партии и уверенность в нем

Точно такая же архитектура сети, веса изменяются по формуле:

$$\Delta\theta \propto \frac{\partial v_\theta(s)}{\partial\theta}(z - v_\theta(s))$$

z — результат партии

MCST в AlphaGo

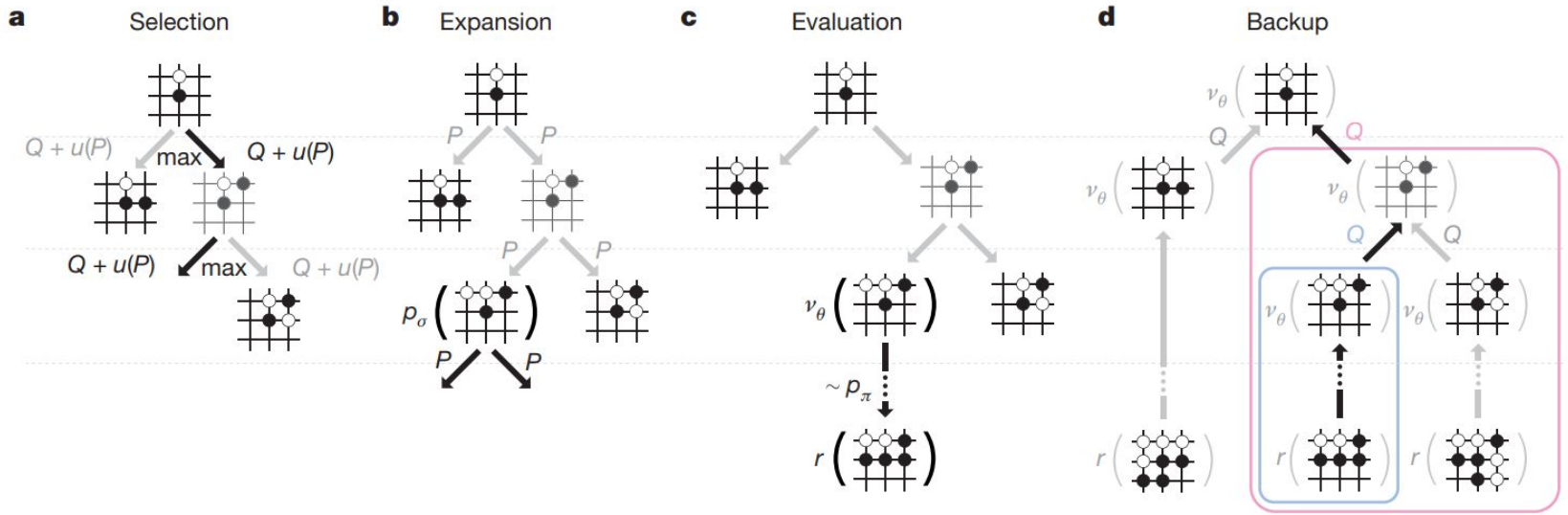


$$a_t = \text{argmax}(Q(s_t, a) + u(s_t, a))$$

$$u(s,a) \propto \frac{P(s,a)}{1+N(s,a)}$$

Supervised policy

MCST в AlphaGo



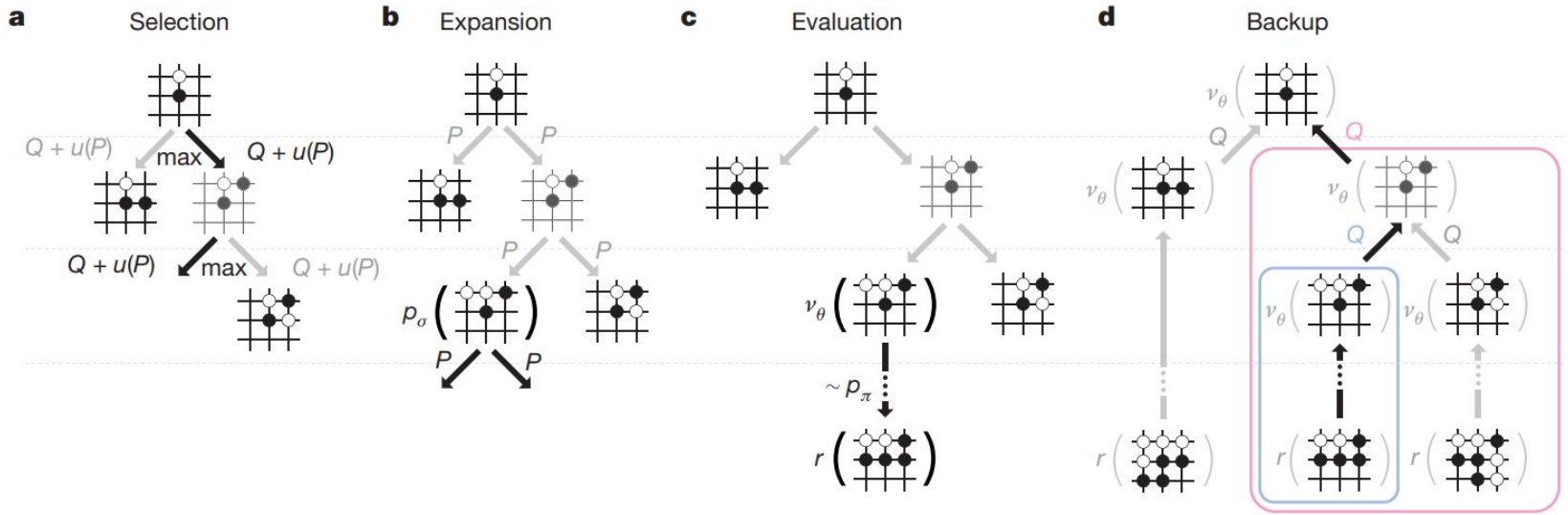
$$a_t = \text{argmax}(Q(s_t, a) + u(s_t, a))$$

$$u(s,a) \propto \frac{P(s,a)}{1+N(s,a)}$$

Supervised policy

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

MCST в AlphaGo



$$a_t = \text{argmax}(Q(s_t, a) + u(s_t, a))$$

$$u(s,a) \propto \frac{P(s,a)}{1+N(s,a)}$$

Supervised policy

$$N(s,a) = \sum_{i=1}^{n} 1(s,a,i)$$

$$Q(s,a) = \frac{1}{N(s,a)} \sum_{i=1}^{n} 1(s,a,i) V(s_L^i)$$

AlphaGo Zero

- Нет обучения на человеческих партиях
- Нет заинженеренных признаков, только позиция на поле (нет)
- Вместо value network и policy network используется одна сеть
- Отсутствие rollout

Вопросы:

- Как AlphaGo выбирает, куда пойти
- Как AlphaGo оценивает вероятность победить для заданного поля
- Какие недостатки были у AlphaGo

Список литературы

- Mastering the game of Go with deep neural networks and tree search http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15780-s16/www/AlphaGo.nature16961.pdf
- Mastering the game of Go without human knowledge https://www.nature.com/articles/nature24270.epdf