

AlphaGo

Медведев Дмитрий Владимирович

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП



- Что такое GO
- Как решали задачу до AlphaGo
- Обучение AlphaGo
- Работа AlphaGo
- Результаты AlphaGo
- Обучение AlphaGo Zero
- Работа AlphaGo Zero
- Результаты AlphaGo Zero

Что такое Go

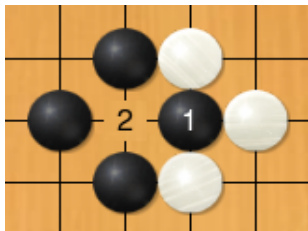
Го — детерминированная игра с полной информацией.



Для любой детерминированной игры с полной информацией, теоретически, можно просчитать всё дерево возможных ходов игроков и определить последовательность ходов, которая гарантированно приведёт по крайней мере одного из них к выигрышу или ничьей.

Правило ко

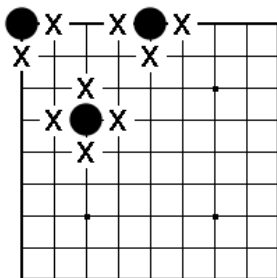
Запрещается делать ход, который приводит к повторению позиции, которая была на доске за один ход до этого

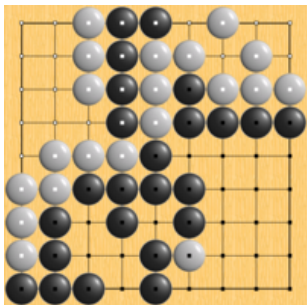


Чёрные только что сделали ход в пункт 1, взяв белый камень в пункте 2.

Правило ко запрещает белым ходить в пункт 2, так как после этого хода восстановится та же позиция, которая была до хода чёрных.

Дамэ камня (степени свободы, дыхания) отмечены «X»





Подсчёт очков на доске 9x9.



Пункты и камни, помеченные белыми метками — принадлежат белым, чёрными — чёрным.

Чёрные камни, помеченные белыми метками, и белые, помеченные чёрными — мертвы, они снимаются с доски и занимаемые ими пункты территории достаются противникам. При любом способе подсчёта белые победили с перевесом в 3,5 очка.

Шахматы и Го



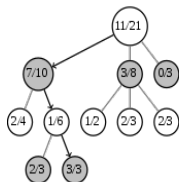
Рост количества возможных вариантов развития*

		
Ход 1	35	200
Ход 2	1 225	40 000
Ход 3	42 875	8 000 000
Ход 4	1 500 625	1 600 000 000

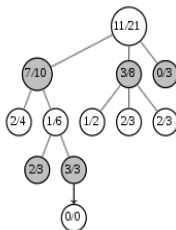
* начиная со среднестатистической игровой позиции

Monte Carlo Tree Search

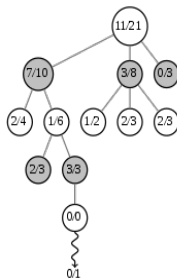
Selection



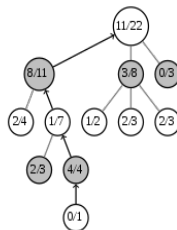
Expansion



Simulation



Backpropagation



проблема exploration и exploitation

Есть проблема как выбирать в Selection дочернее звено: можно начать заходить, всё время в одно и тоже, возможно наилучшее звено («эксплуатировать» его), но не проверив остальные.

Upper Confidence Bounds:

$$\frac{w_i}{n_i} + c \sqrt{\frac{\ln N_i}{n_i}}$$

w_i — число побед одержанных в симуляциях, после совершения хода рассматриваемого звена

N_i — суммарное число симуляций совершенных в поддеревьях рассматриваемого звена

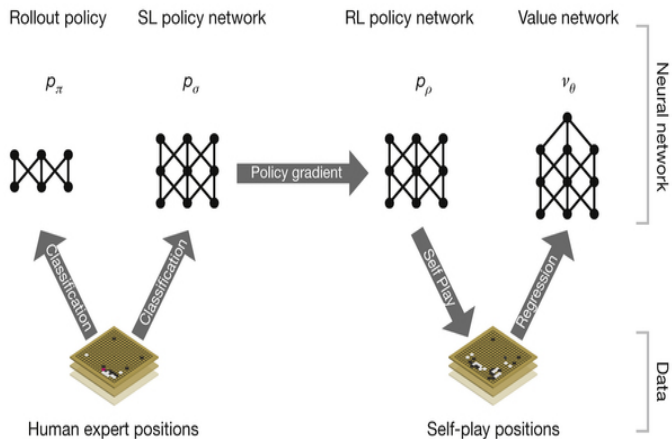
c — «exploration» параметр

$w_i/n_i = \text{«Value»}$

Решение состоит в использовании трёх «технологий»:

- Monte Carlo tree search
- reinforcement learning
- deep learning

основные компоненты AlphaGo



policy network(стратегическая сеть) — нейронные сети, которые помогают выбирать хороший ход. Всего их три вида:

- SL policy network
- RL policy network
- Rollout policy network (быстрая)

SL policy network

supervised learning policy network: обучение на играх людей

Градиент: $\Delta\sigma \propto \frac{\partial \log p_\sigma(a|s)}{\partial \sigma}$

σ — параметры сети,

a — action(ход),

s — state(позиция),

p — probability (выход сети)

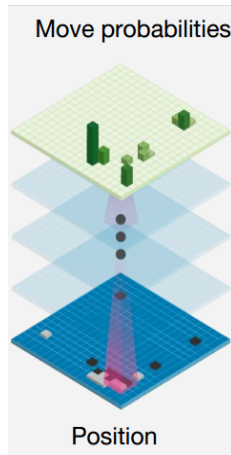
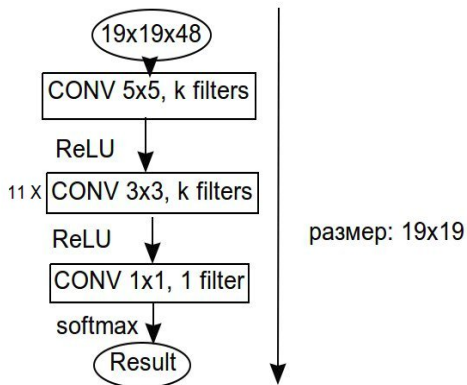
Признаки SL policy network

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

Extended Data Table 2: **Input features for neural networks.** Feature planes used by the policy network (all but last feature) and value network (all features).

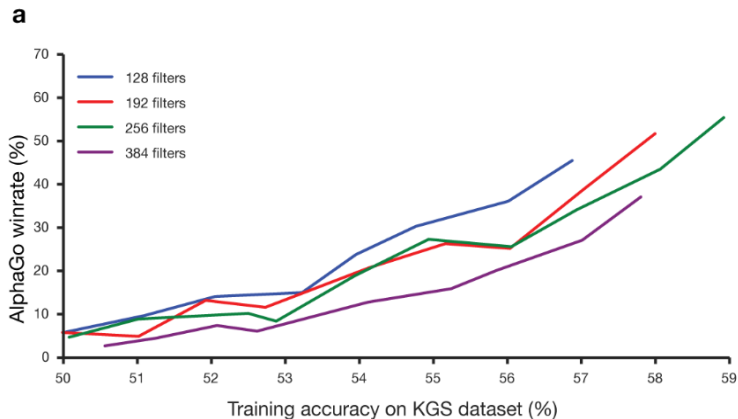
Архитектура SL policy network

На картинках процессы идут в разные стороны



Подбор числа фильтров SL policy network

Была выбрана модель с 128 фильтрами



RL policy network

reinforcement learning: обучение на играх с собой

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

ρ — параметры сети,

a — action(ход),

s — state(позиция),

p — probability (выход сети)

z — reward на последнем T ходе:

1 — победа

-1 — проигрыш.

Rollout policy network

Архитектура: один полносвязный слой, иначе говоря линейный классификатор.

Feature	# of patterns	Description
Response	1	Whether move matches one or more response features
Save atari	1	Move saves stone(s) from capture
Neighbour	8	Move is 8-connected to previous move
Nakade	8192	Move matches a <i>nakade</i> 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
Self-atari	1	Move allows stones to be captured
Last move distance	34	Manhattan distance to previous two moves
Non-response pattern	32207	Move matches 12-point diamond pattern centred around move

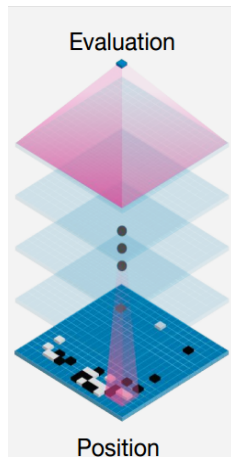
Extended Data Table 4: **Input features for rollout and tree policy.** Features used by the rollout policy (first set) and tree policy (first and second set). Patterns are based on stone colour (black/white/empty) and liberties ($1, 2, \geq 3$) at each intersection of the pattern.

Value network

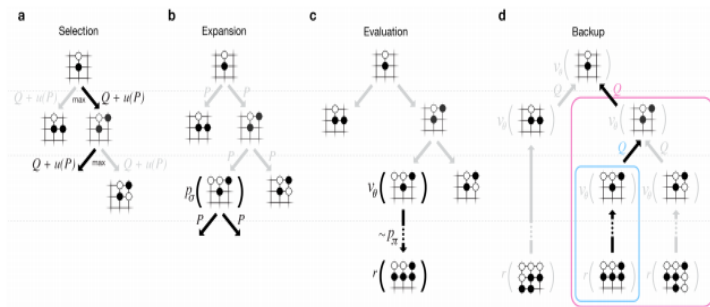
Value Network(Оценочная сеть) — сеть которая пытается оценить вероятность выигрыша в данной позиции, возвращает значение от -1 до 1.

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

θ — параметры сети,
 s — state(позиция),
 v — value (выход сети)
 z — результат игр



Работа AlphaGo



Работа AlphaGo

1)**Selection:** $a_t = \underset{a}{\operatorname{argmax}}(Q(s_t, a) + u(s_t, a)), u(s, a) \propto \frac{P(s, a)}{1+N(s, a)}$

u - специальная добавка, которая стимулирует exploration

2)**Expansion:** $P(s, a) = p_\sigma(a|s)$

3)**Evaluation:** $V(s_L) = (1 - \lambda)v_\theta(s_L) + \lambda z_L$

4)**Backup:**

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

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

Фань Хуэй, счёт: 5—0



Результаты AlphaGo

Ли Седоль, счёт: 4—1



Результаты AlphaGo

Кэ Цзе, счёт: 3—0



Проблемы AlphaGo

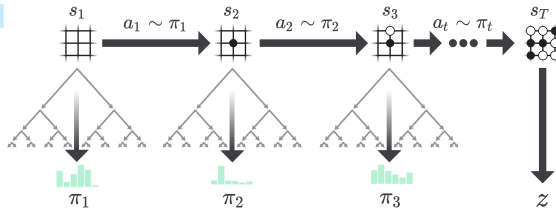
- Для стартового обучения используются игры людей
- Много «инженерных» признаков
- Нужны большие вычислительные мощности

Отличия:

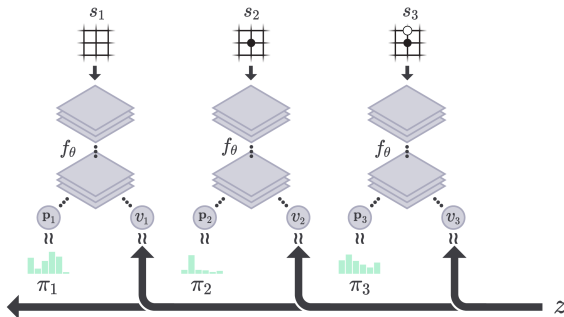
- Вместо value сети и policy сетей одна сеть с двумя выходами
- Использование MCTS при тренировке

обучение AlphaGo Zero

a. Self-Play



b. Neural Network Training



обучение AlphaGo Zero

Формула Loss:

$$L = (z - v)^2 - \pi^T \log p + c \|\theta\|^2$$

- z — результат партии 1 или -1
- v — value лежит от -1 до 1
- p — вектор распределений actions
- π — вектор распределений полученный с помощью раскрытия поддеревьев в MCTS
- c — коэффициент регуляризации

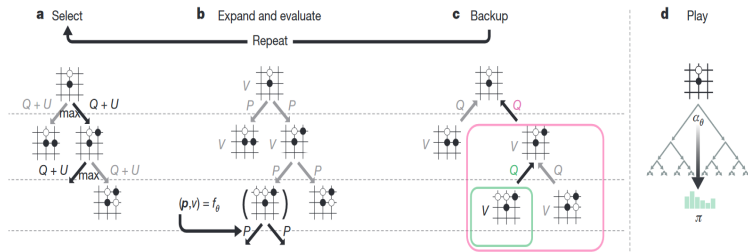
Изменения в признаковом пространстве:

- осталось 17 признаков
- 8 признаков — история ходов AlphaGo
- 8 признаков — история ходов оппонента
- 1 признак — цвет

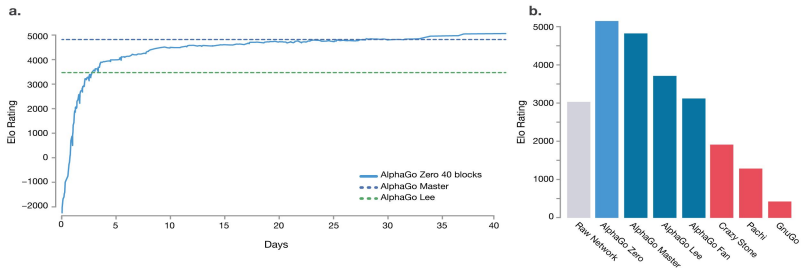
Изменения в архитектуре пространстве:

- появилась batch normalization
- стало 40 слоёв
- используются residual layers

работа AlphaGo Zero



результаты AlphaGo Zero



Оригинальная статья:

<https://gogameguru.com/i/2016/03/deepmind-mastering-go.pdf>

Хабрхабр:

<https://habrahabr.ru/post/343590/>

<https://habrahabr.ru/post/279071/>