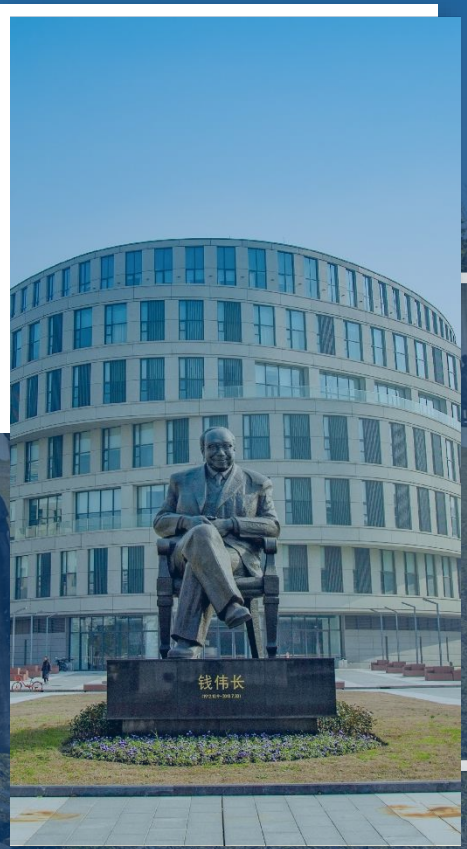
上海大学未来技术学院 SCHOOL OF FUTURE TECHNOLOGY, SHANGHAI UNIVERSITY
上海大学人工智能研究院 INSTITUTE OF ARTIFICIAL INTELLIGENCE, SHANGHAI UNIVERSITY

人工智能导论

——第2课：强化学习



叶林奇

未来技术学院（人工智能研究院）

2024秋季学期

提纲

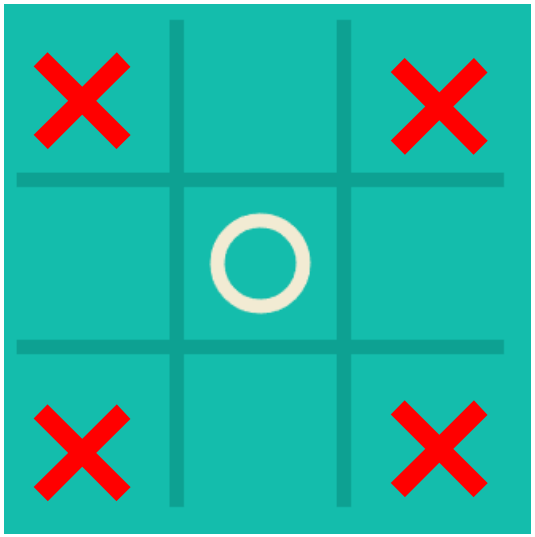
一、从AlphaGo讲起

二、强化学习经典算法

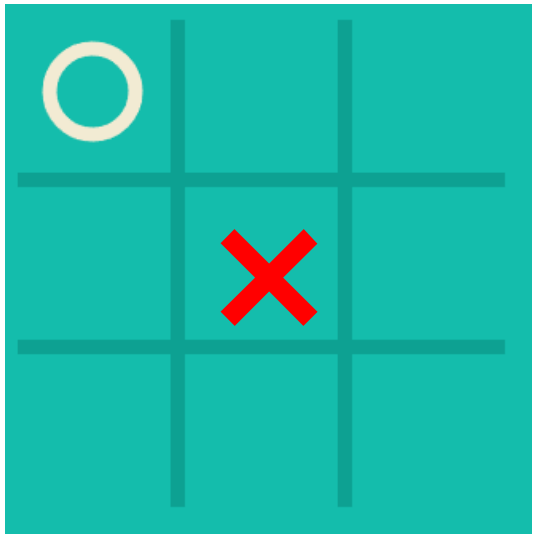
本课要点

1. MiniMax Tree Search
2. Monte Carlo Tree Search
3. AlphaGo

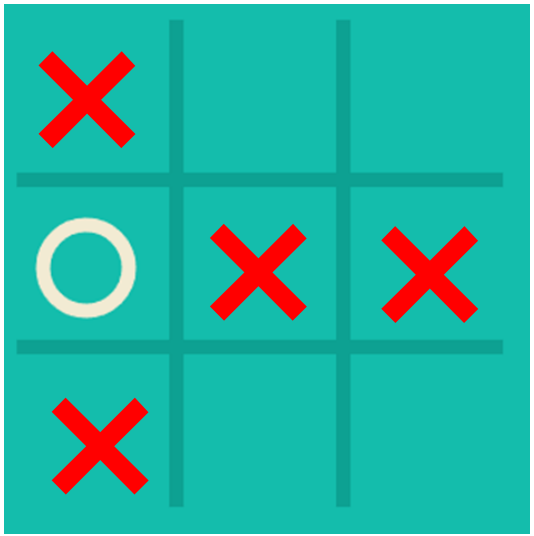
MiniMax Tree Search



下中必下角



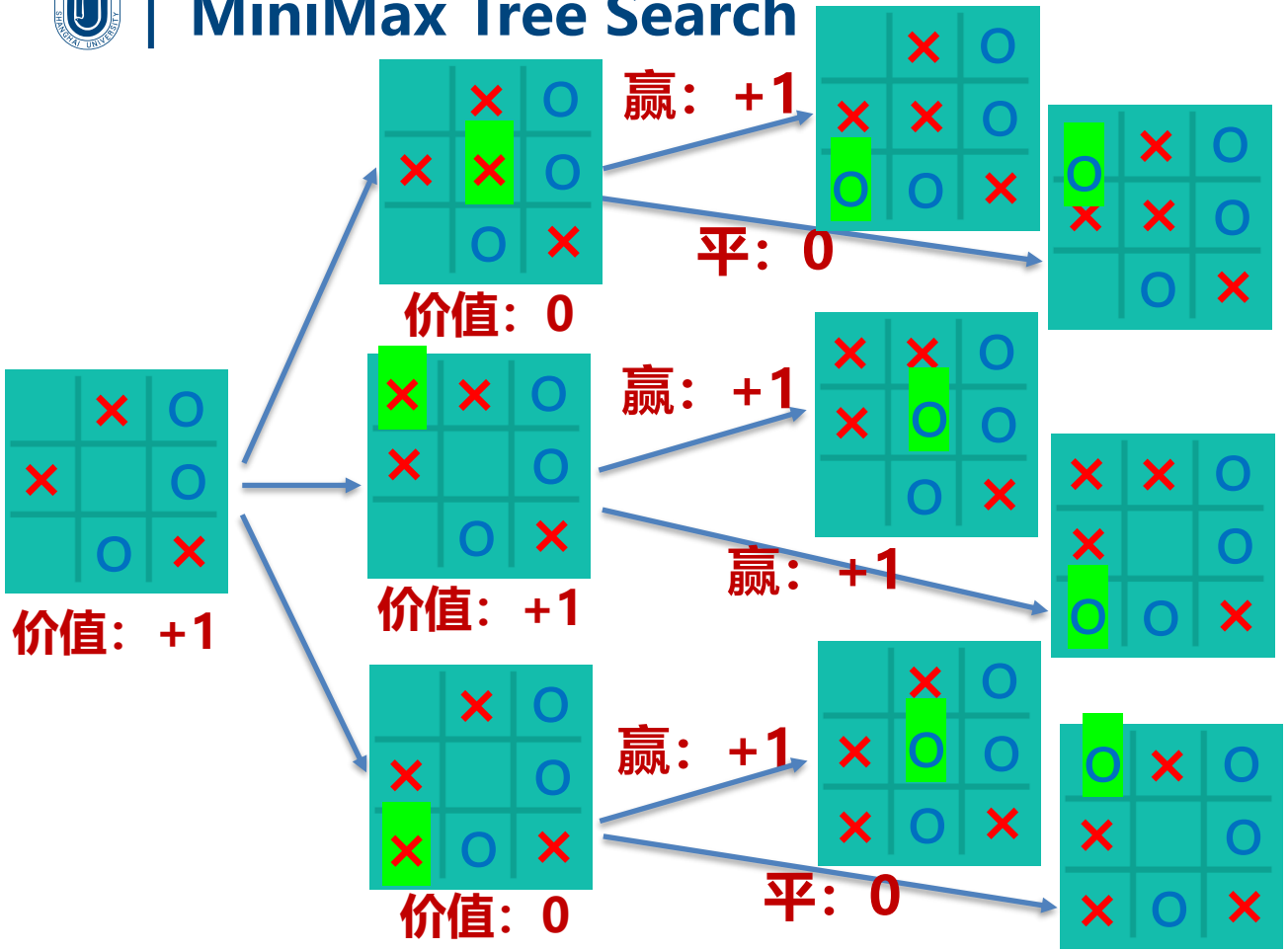
下角必下中



下边占旁边

位置的价值高

MiniMax Tree Search



MiniMax Tree Search

Deep Blue，1997

围棋

围棋规则

- 提子
- 围地多的一方获得胜利

围棋复杂度

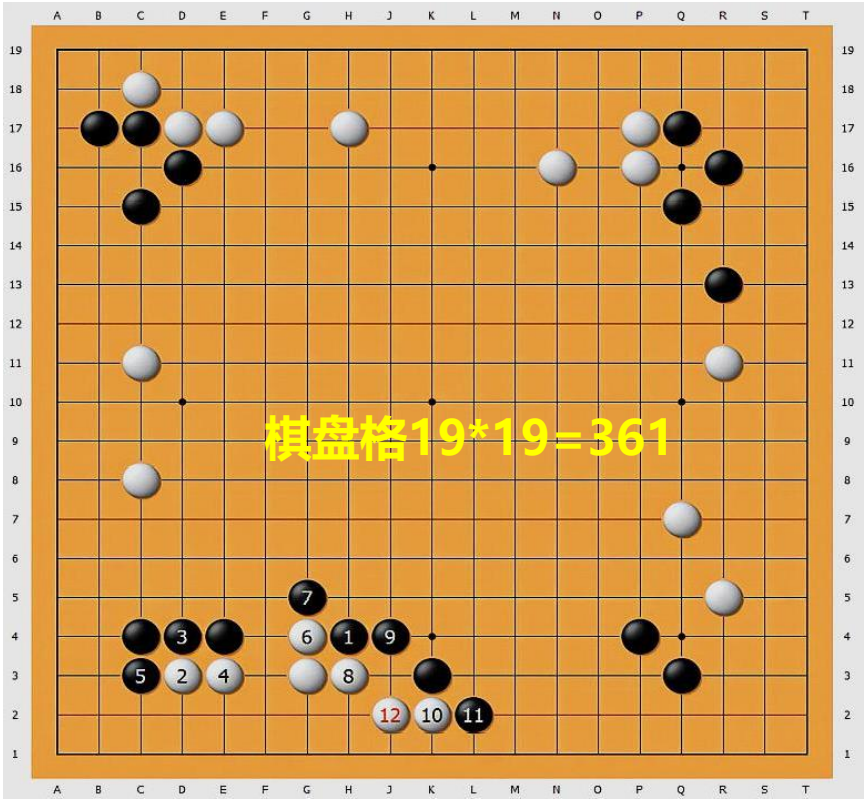
围棋游戏树大小：$361! \approx 10^{768}$

围棋游戏树复杂度：10^{360}

井字棋复杂度：10^{5}

国际象棋复杂度：10^{123}

宇宙原子总数：10^{80}

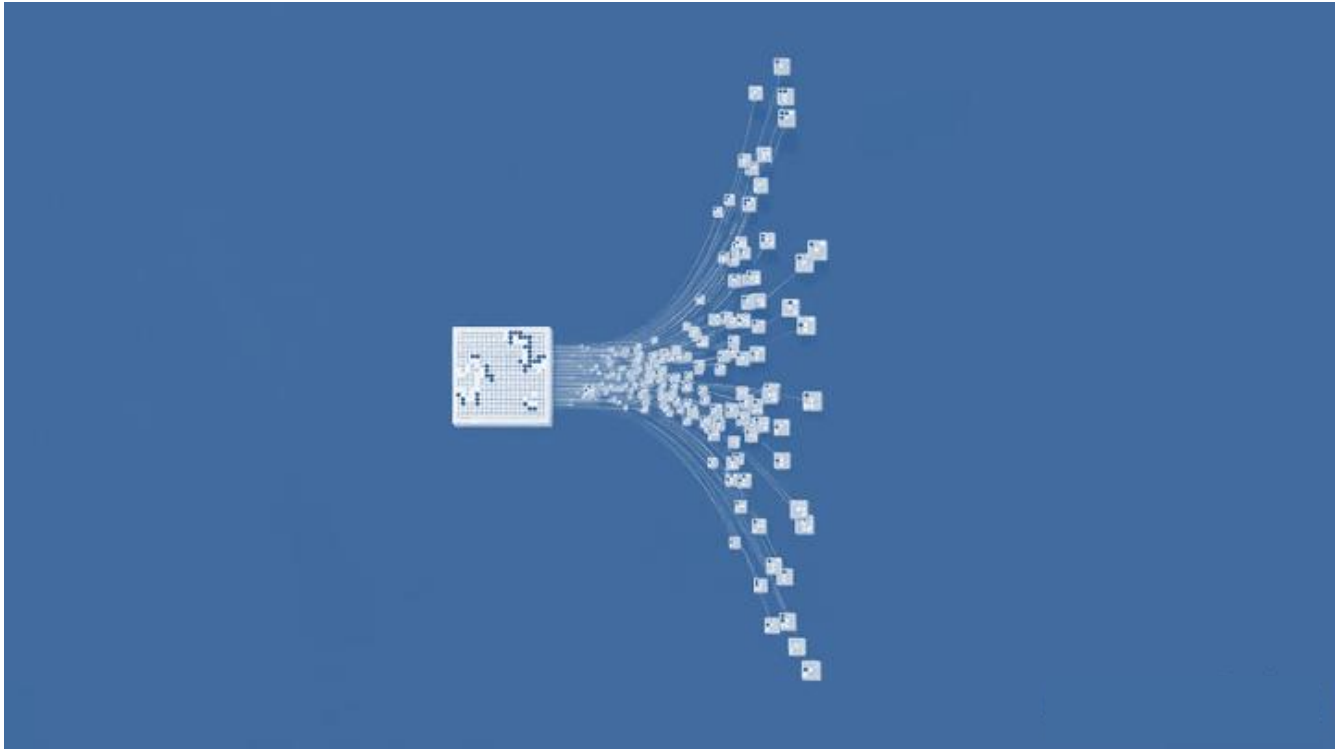


Monte Carlo Tree Search

Monte Carlo Tree Search

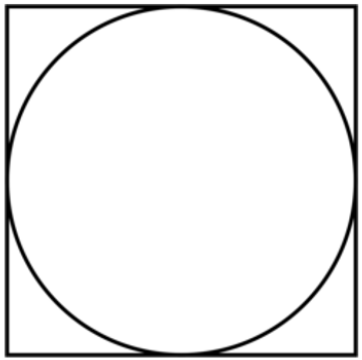
II

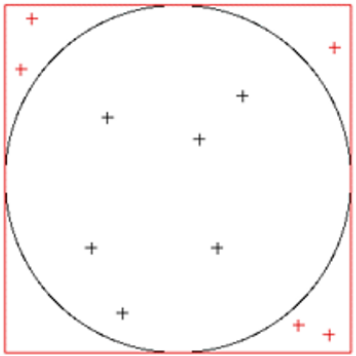
MiniMax Tree Search + Monte Carlo Rollout

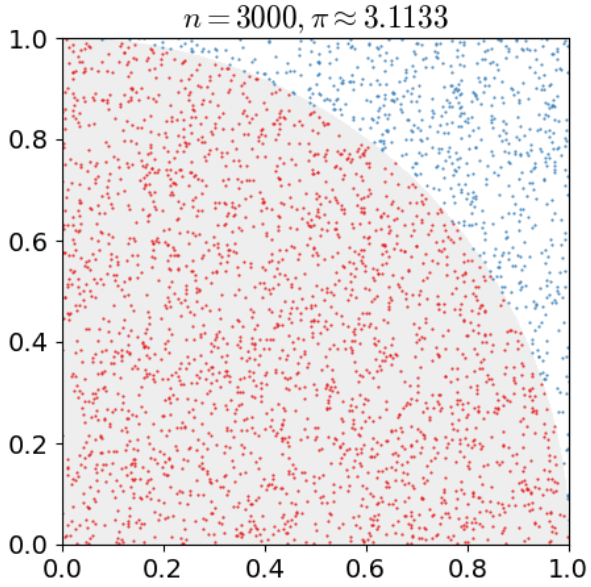


Monte Carlo Tree Search

- 蒙特卡洛方法（Monte-Carlo methods）是一类广泛的计算算法。生活中处处都是MC方法。
  - 依赖于重复随机抽样来获得数值结果
- 例如，计算圆的面积



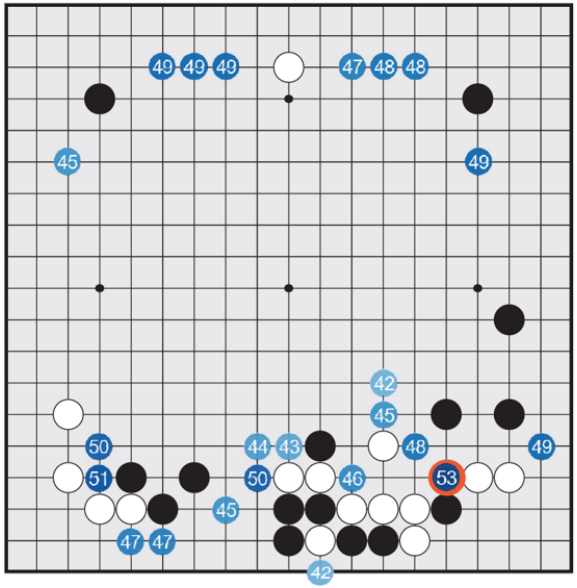


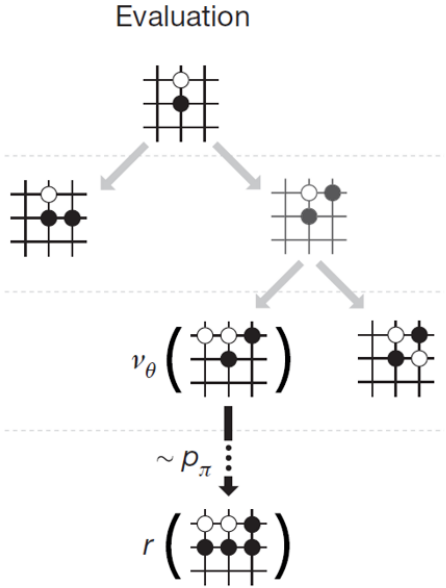


$$\text{Circle Surface} = \text{Square Surface} \times \frac{\#\text{points in circle}}{\#\text{points in total}}$$

Monte Carlo Tree Search

Monte Carlo Rollout：通过随机抽样估计当前状态下的胜率

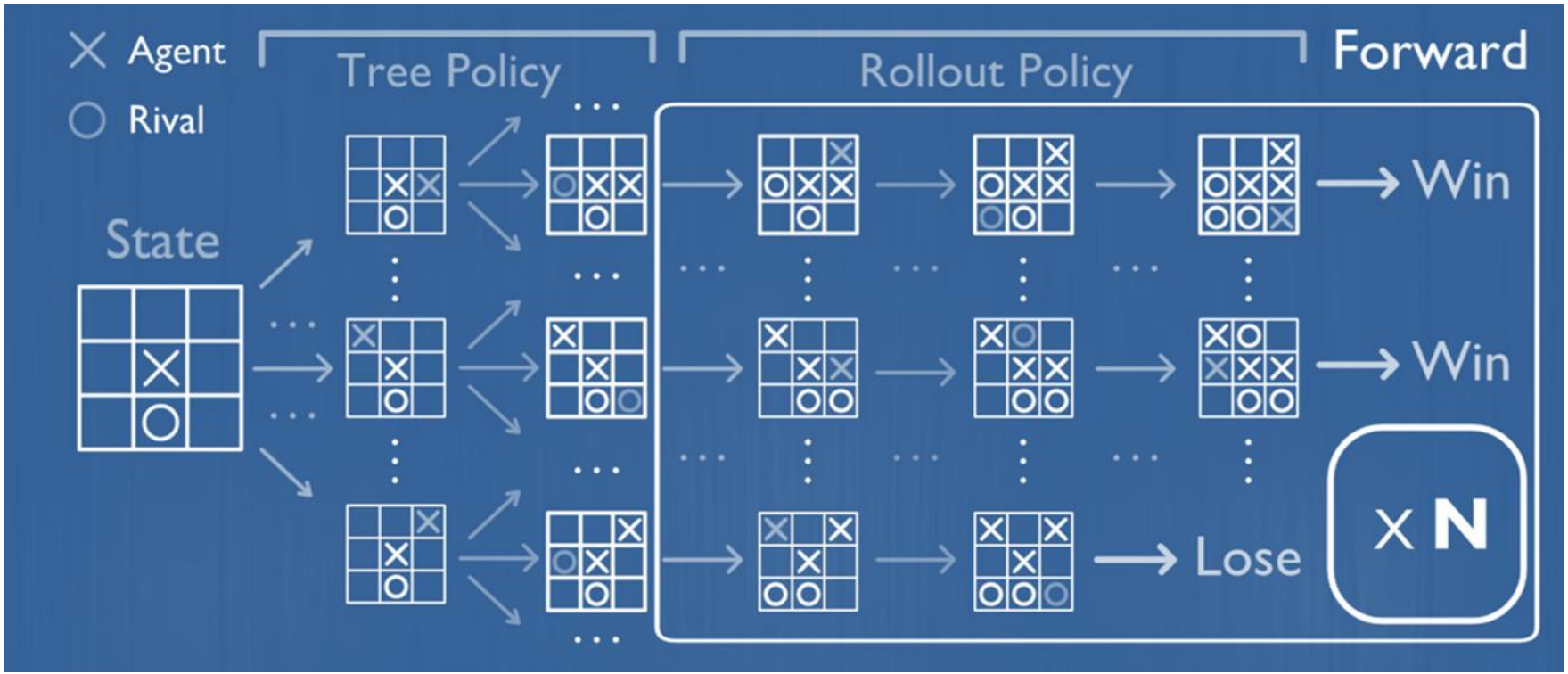




$$\text{Win Rate}(s) = \frac{\#\text{win simulation cases started from } s}{\#\text{simulation cases started from } s \text{ in total}}$$

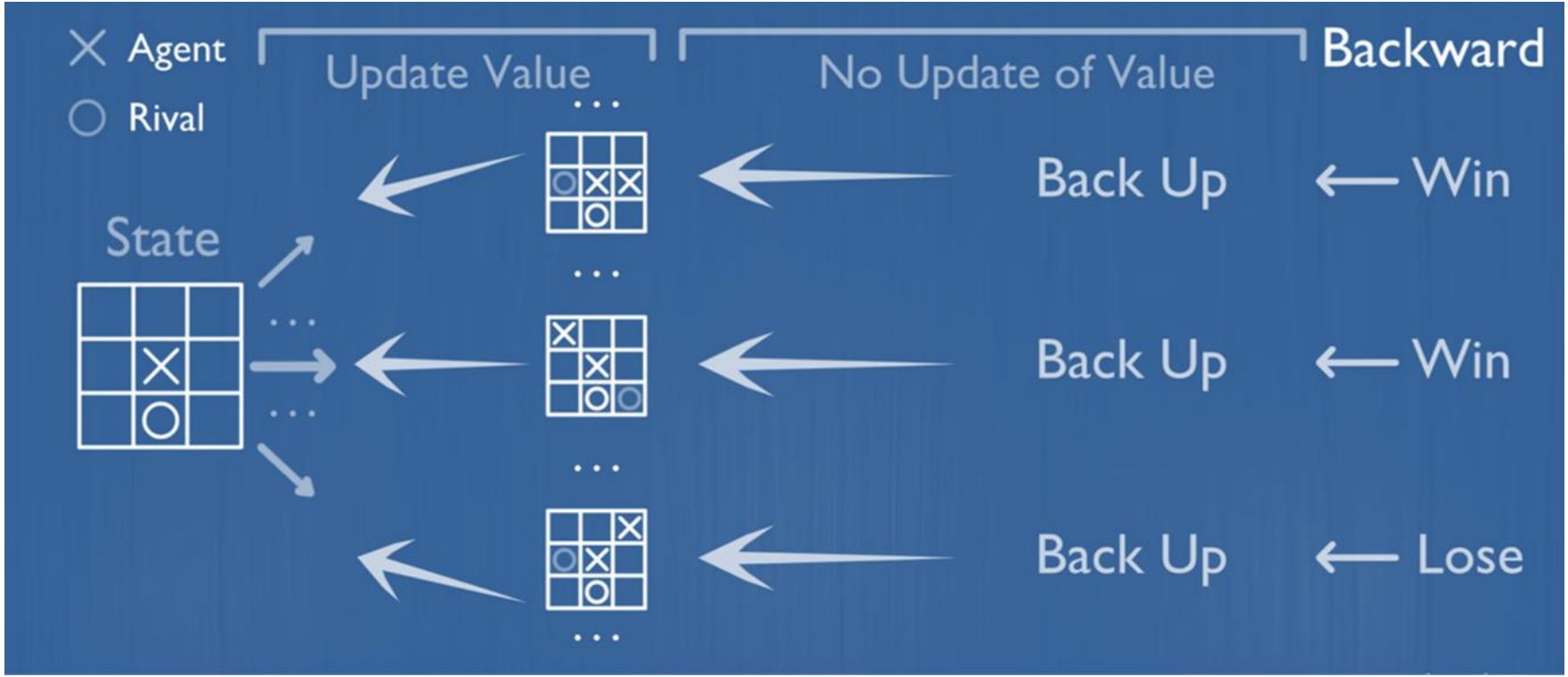
Monte Carlo Tree Search

MiniMax Tree Search + Monte Carlo Rollout→ Monte Carlo Tree Search



Monte Carlo Tree Search

MiniMax Tree Search + Monte Carlo Rollout→ Monte Carlo Tree Search



AlphaGo

MCTS + Deep NN → AlphaGo

人类棋谱

监督学习
策略网络

p_σ

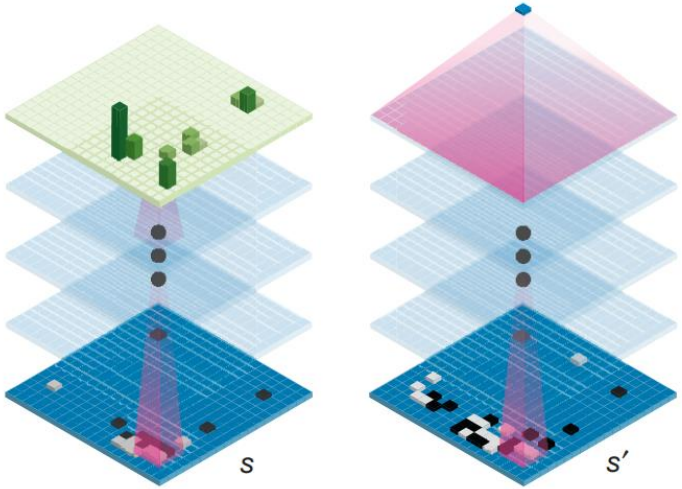
自我对弈

强化学习
策略网络

$p_{\sigma/\rho}(a|s)$

强化学习
价值网络

$v_\theta(s')$



监督学习
16万局人类棋谱

强化学习
3000万局自博弈

本课小结

1997，深蓝战胜国际象棋王卡斯帕罗夫

Minimax Tree Search + Monte Carlo Rollout

通过穷举找到价值　　通过随机抽样估计价值

Deep Neural Network + Monte Carlo Tree Search

用深度神经网络拟合价值　　通过 先穷举+后抽样 估计价值

AlphaGo

通过深度神经网络进行树搜索

2016，AlphaGo战胜围棋冠军李世石

参考资料

谷歌Deepmind

2016 Nature论文

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

BiliBili

UP主PenicillinLP视频

手机扫码观看/分享

提纲

一、从AlphaGo讲起

二、强化学习经典算法

How I Teach My Mice To Walk Backwards

The Backwards Brain Bicycle

在与动态环境的交互中学习

有监督、无监督学习

Model ⟵

Fixed Data

强化学习

Agent ⟷

Dynamic Environment

Agent不同，交互出的数据也不同！

和动态环境交互产生的数据分布

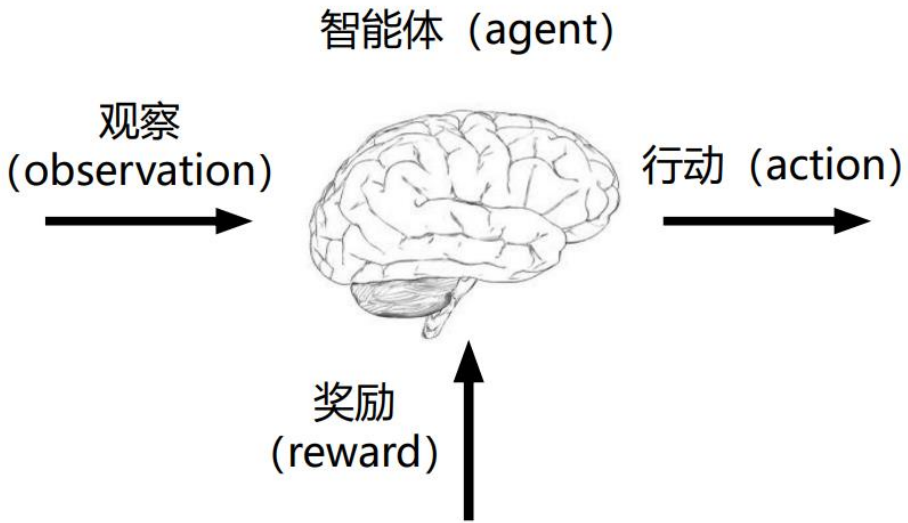
- 给定同一个动态环境（即MDP），不同的策略采样出来的(状态-行动)对的分布是不同的

- 占用度量（Occupancy Measure）

$$\rho^{\pi}(s,a) = \mathbb{E}_{a\sim\pi(s),\, s'\sim p(s,a)}\left[\sum_{t=0}^{T}\gamma^{t}\, p(s_t = s, a_t = a)\right]$$

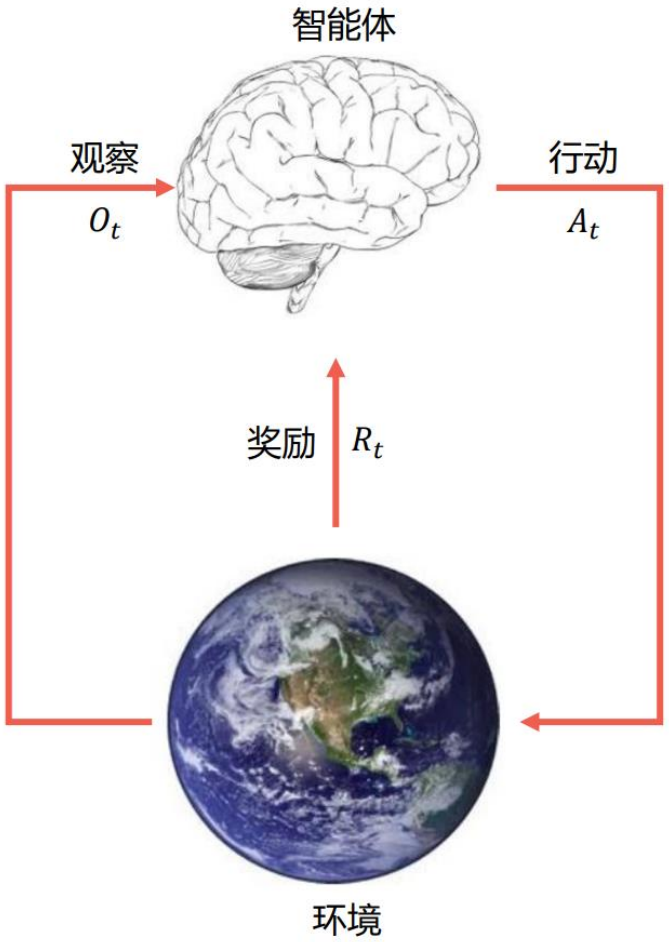
强化学习定义

- 通过从交互中学习来实现目标的计算方法



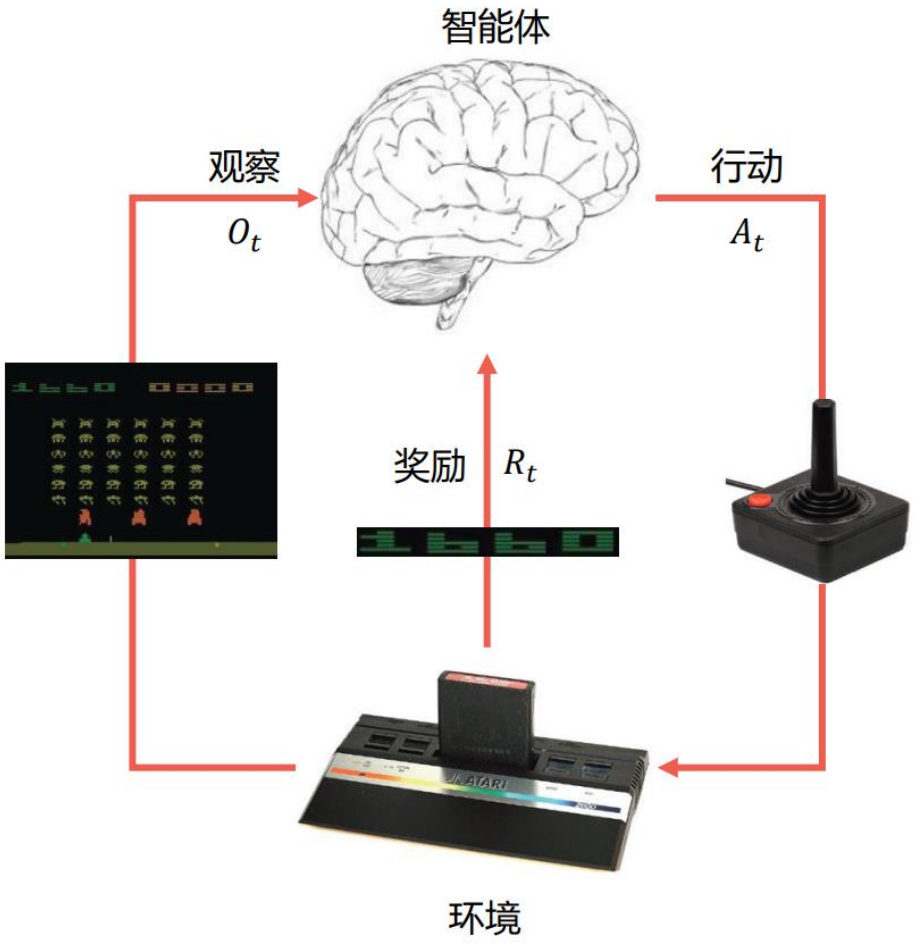
- 三个方面：
  - 感知：在某种程度上感知环境的状态
  - 行动：可以采取行动来影响状态或者达到目标
  - 目标：随着时间推移最大化累积奖励

强化学习交互过程



- 在每一步 t，智能体：
  - 获得观察 O_t
  - 获得奖励 R_t
  - 执行行动 A_t
- 环境：
  - 获得行动 A_t
  - 给出观察 O_{t+1}
  - 给出奖励 R_{t+1}
- t 在环境这一步增加

举例：Atari游戏



- 游戏规则未知
- 从交互游戏中进行学习
- 在操纵杆上选择行动并查看分数和像素画面

序列决策任务中的一个基本问题

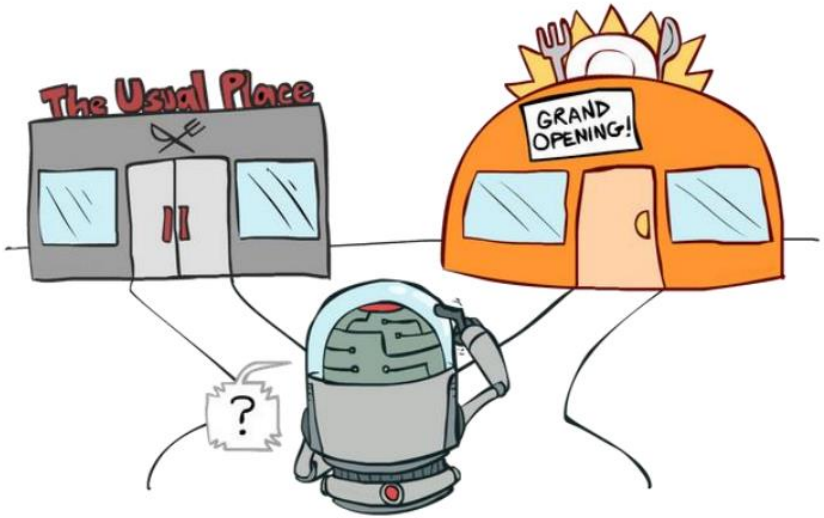
- 基于目前策略获取已知最优收益还是尝试不同的决策
  - **Exploitation** 执行能够获得已知最优收益的决策
  - **Exploration** 尝试更多可能的决策，不一定会是最优收益

$$\mathcal{E}_t = \{\pi_t^i \mid i = 1, \dots, n\} \xrightarrow{\text{探索}} \mathcal{E}_{t+1} = \{\pi_t^i \mid i = 1, \dots, n\} \cup \{\pi_e^j \mid j = 1, \dots, m\}$$

$$\exists\, V^\star(\cdot \mid \pi_t^i \sim \mathcal{E}_t) \le V^\star(\cdot \mid \pi_{t+1}^i \sim \mathcal{E}_{t+1}) \qquad \pi_{t+1}^i \sim \{\pi_e^i \mid i = 1, \dots, m\}$$

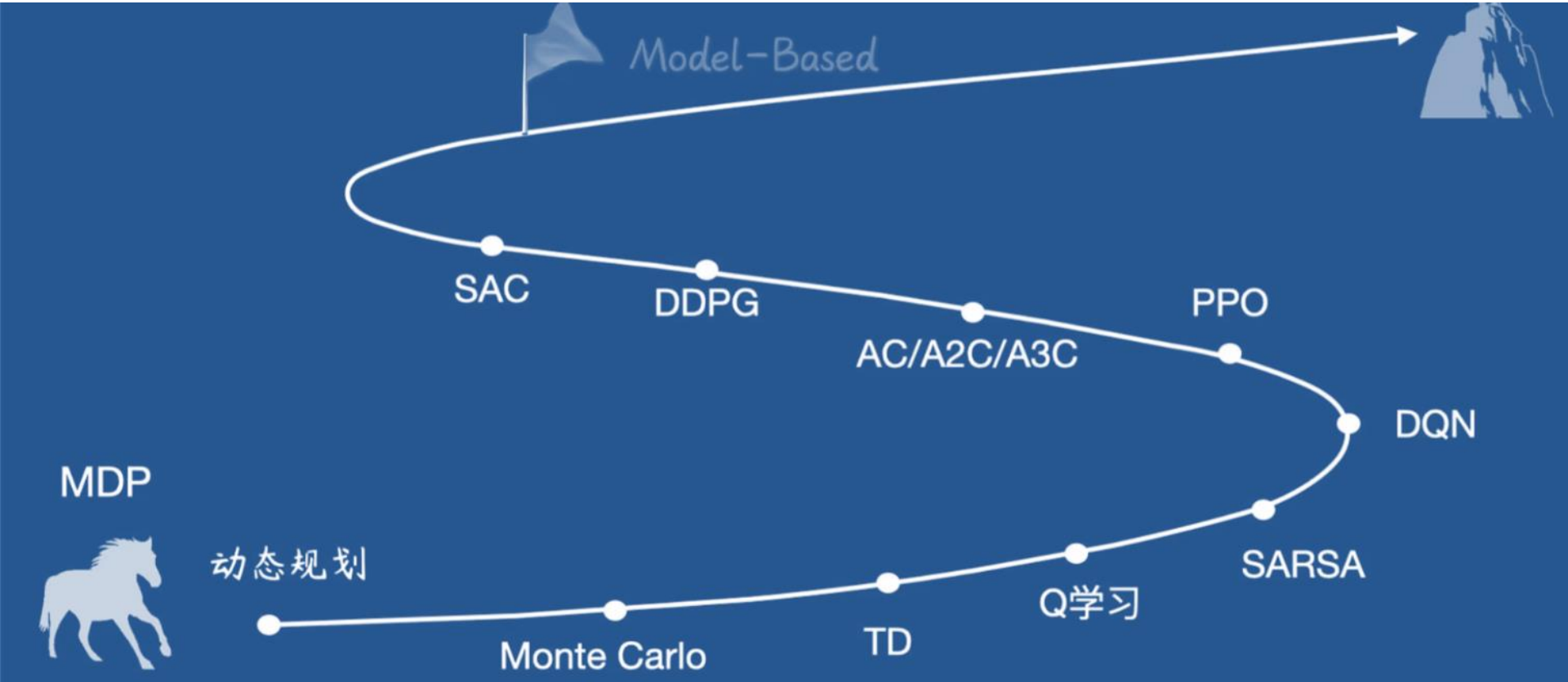
探索：可能发现更好的策略

一个例子


The Usual Place
GRAND OPENING!
?


10:20
搜索
美食
吃炸串
地方菜系
全球美食
轻食简餐
甜品饮
综合排序
距离
销量
筛选
守护联盟
津贴优惠
满减优惠
品质联盟
守护联盟
和番丼饭(东川路店)
4.8 月售5143
起送¥15 远距离配送¥5 ¥8
42分钟 4.7km
"不错哦,好待咖喱虾排饭很赞"
20减16 49减22 津贴1元 78减28 120减35
守护联盟
槿巷多料拌饭(闵行旗舰店)
4.6 月售1644
起送¥15 远距离配送¥3.5 ¥7.8
39分钟 4.0km
"第二次吃了,很香的豆腐" 闵行区韩国料理实惠第1名
28减15 45减20 津贴1元 65减25 90减32
蜜哆哆韩式炸鸡(颛桥店)
预订中 10:30配送
4.8 月售1145
起送¥15 远距离配送¥8
41分钟 5.0km
"点的无骨香酥鸡,整体很满意" 元气好店
28减12 48减24 78减33 118减40 6元会员红包
守护联盟
飯豊町·和風精致便当(东川路店)
4.9 月售3842
起送¥20 远距离配送¥3 ¥8.3
46分钟 4.7km
已检测体温 请放心食用
23减14 49减15 80减21 119减38 6元会员红包

强化学习


Model-Based
SAC
DDPG
AC/A2C/A3C
PPO
DQN
MDP
动态规划
Monte Carlo
TD
Q学习
SARSA

深度强化学习的崛起

- 2012年AlexNet在ImageNet比赛中大幅度领先对手获得冠军
- 2013年12月，第一篇深度强化学习论文出自NIPS 2013 Reinforcement Learning Workshop

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

DeepMind

“Human-level control through deep reinforcement learning”

letter

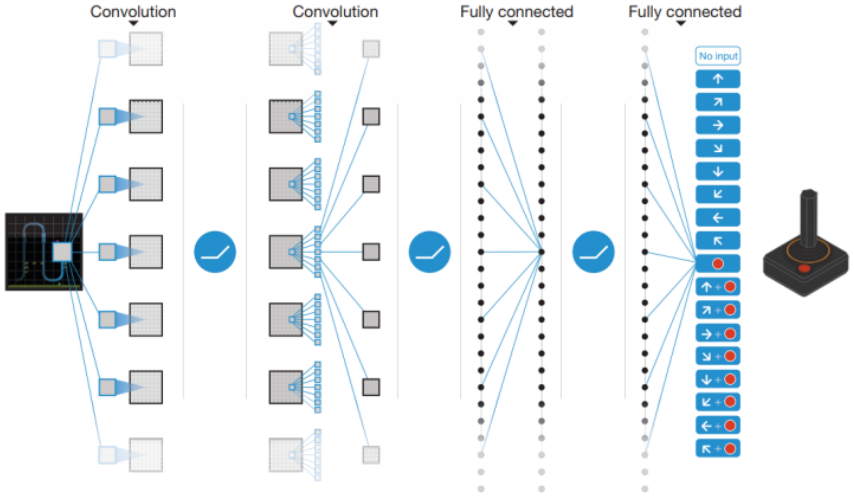
2015

Deep Q-Learning

深度强化学习

深度强化学习

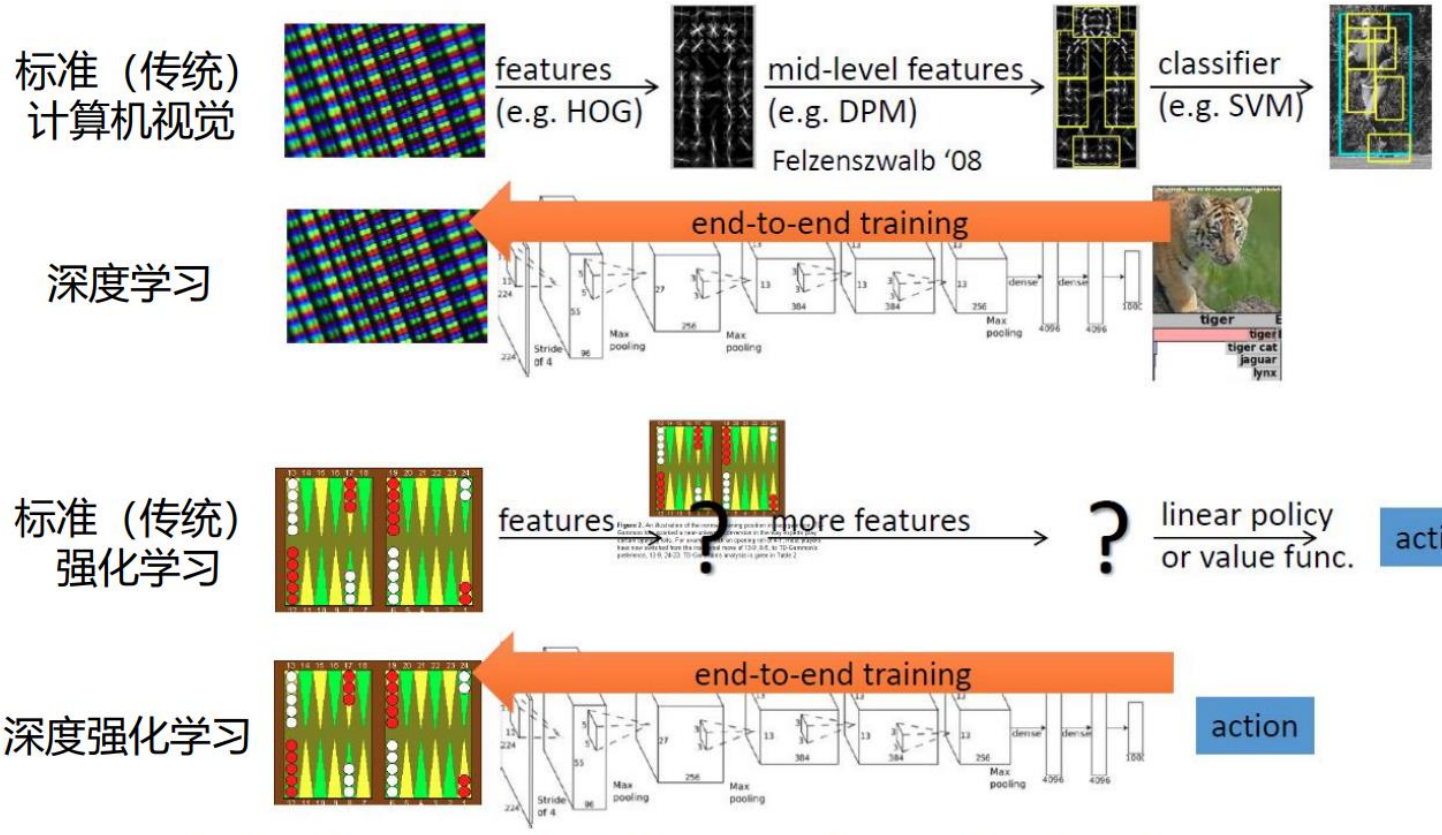
- 利用深度神经网络进行价值函数和策略近似
- 从而使强化学习算法能够以端到端的方式解决复杂问题



$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot);s'\sim\mathcal{E}}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i)\right) \nabla_{\theta_i} Q(s, a; \theta_i)\right]$$

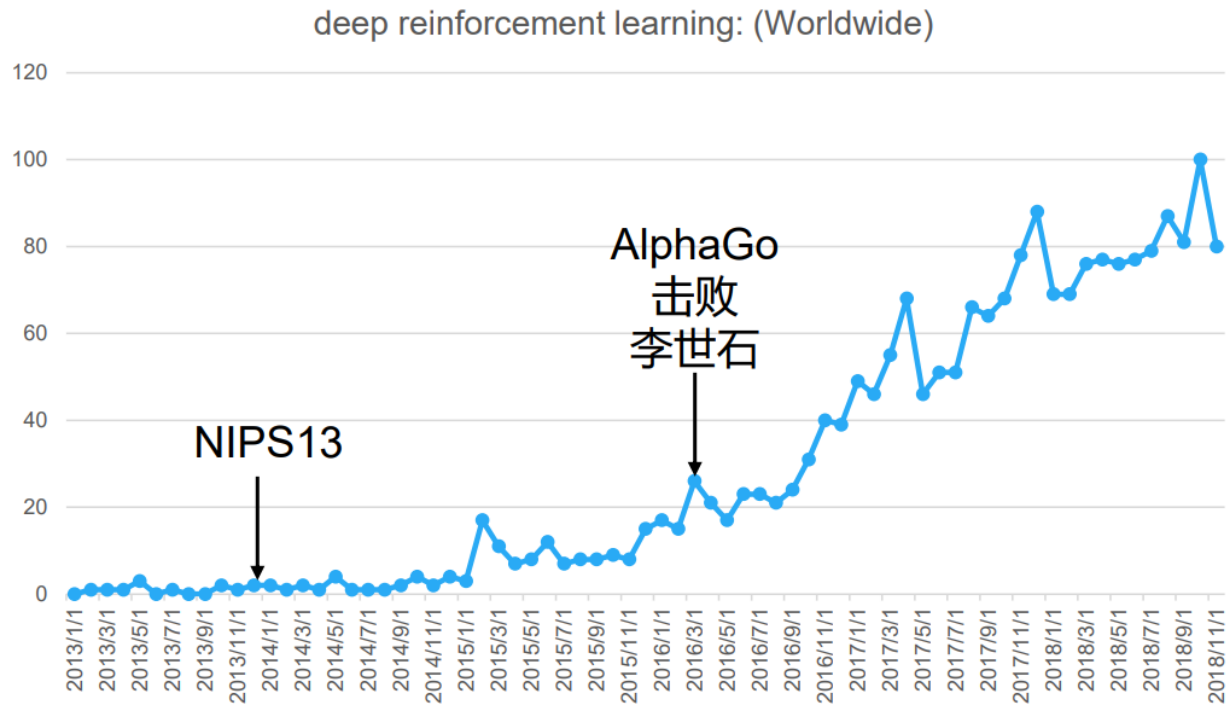
Q函数的参数通过神经网络反向传播学习

端到端强化学习



- 深度强化学习使强化学习算法能够以端到端的方式解决复杂问题
- 从一项实验室学术技术变成可以产生GDP的实际技术

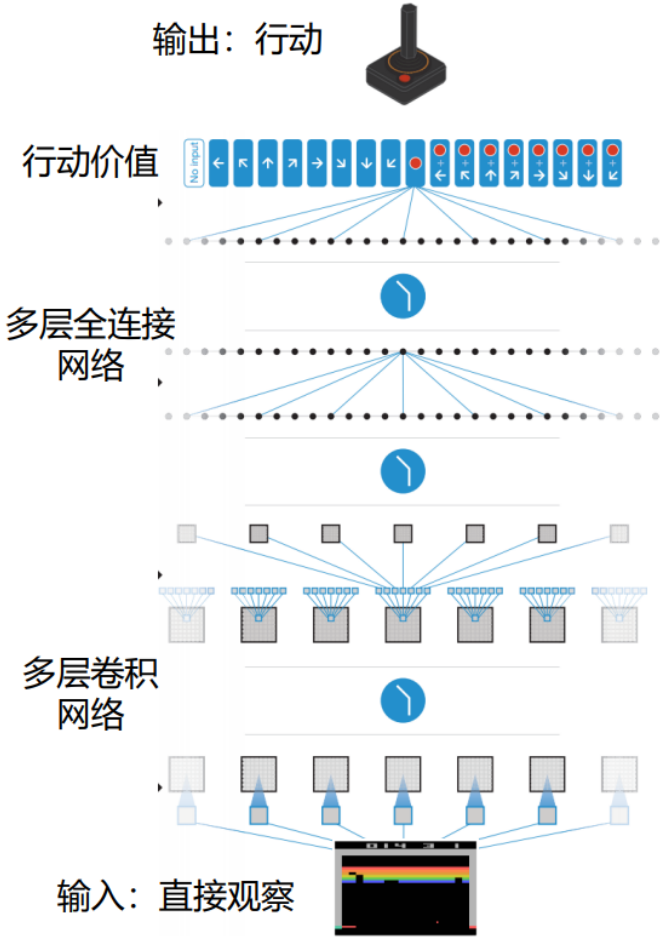
深度强化学习趋势



- Google搜索中词条“深度强化学习（deep reinforcement learning）”的趋势

深度强化学习带来的关键变化

- 将深度学习（DL）和强化学习（RL）结合在一起会发生什么？
  - 价值函数和策略变成了深度神经网络
  - 相当高维的参数空间
  - 难以稳定地训练
  - 容易过拟合
  - 需要大量的数据
  - 需要高性能计算
  - CPU（用于收集经验数据）和GPU（用于训练神经网络）之间的平衡
  - …
- 这些新的问题促进着深度强化学习算法的创新



深度Q网络（DQN）

- Q 学习算法学习一个由 θ 作为参数的函数 $Q_\theta(s,a)$
  - 更新方程 $Q_\theta(s_t,a_t) \leftarrow Q_\theta(s_t,a_t) + \alpha(r_t + \gamma \max_{a'} Q_\theta(s_{t+1},a') - Q_\theta(s_t,a_t))$

直观想法

- 使用神经网络来逼近 $Q_\theta(s,a)$，面对算法不稳定问题
  - 连续采样得到的 $\{(s_t,a_t,s_{t+1},r_t)\}$ 不满足独立分布
  - $\{(s_t,a_t,s_{t+1},r_t)\}$ 为状态-动作-下一状态-回报输入
  - $Q_\theta(s,a)$ 的频繁更新

解决办法

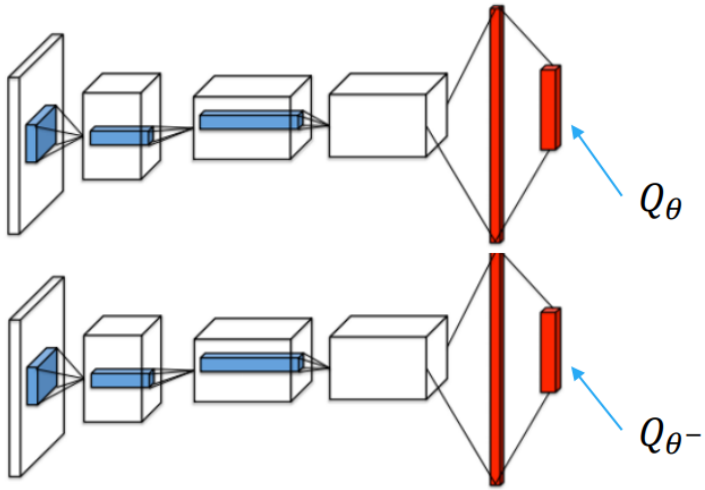
- 经验回放：均匀采样和优先经验回放
- 使用双网络结构：评估网络（evaluation network）和目标网络（target network）

"Human-Level Control Through Deep Reinforcement Learning"，Mnih, Kavukcuoglu, Silver et al. (2015)

目标网络

- 目标网络 $Q_{\theta^-}(s,a)$
  - 使用较旧的参数，记为 θ^-，每隔 C 步和训练网络的参数同步一次。
  - 第i次迭代的损失函数为

$$L_i(\theta_i) = \mathbb{E}_{s_t,a_t,s_{t+1},r_t,p_t\sim D}[\frac{1}{2}\omega_t(\underbrace{r_t + \gamma \max_{a'} Q_{\theta_i^-}(s_{t+1},a')}_{target} - Q_{\theta_i}(s_t,a_t))^2]$$

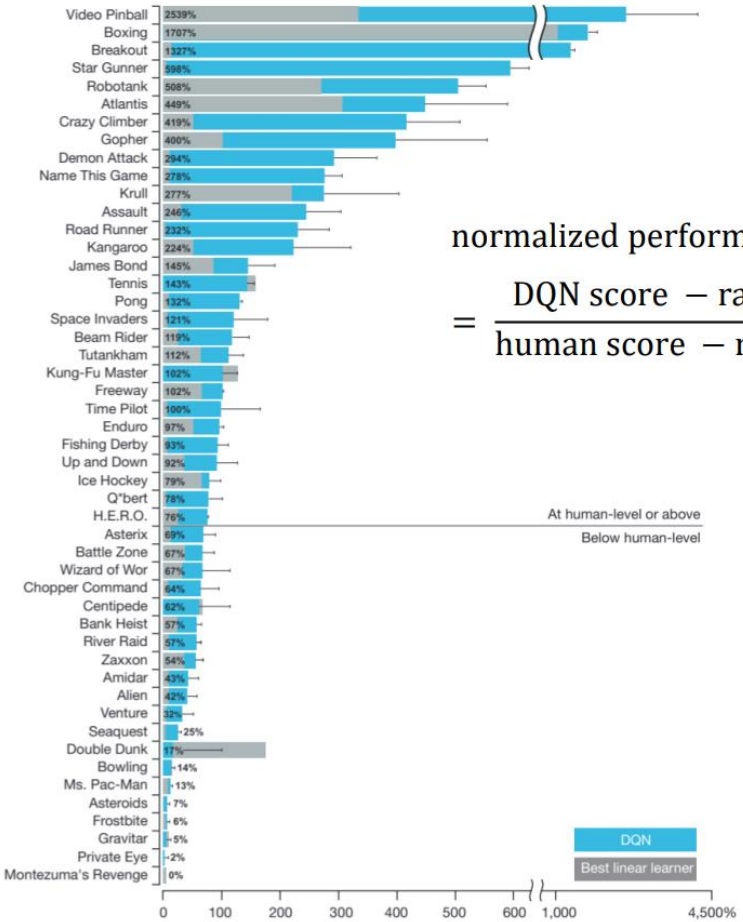


评估网络（evaluation network）

目标网络（target network）是评估网络几轮前的拷贝，不更新

"Human-Level Control Through Deep Reinforcement Learning"，Mnih, Kavukcuoglu, Silver et al. (2015)

在 Atari 环境中的实验结果



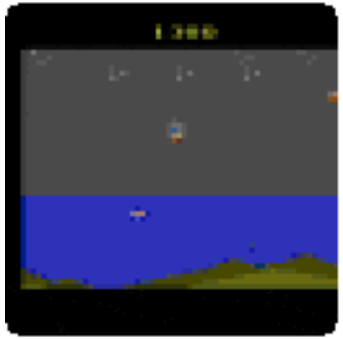
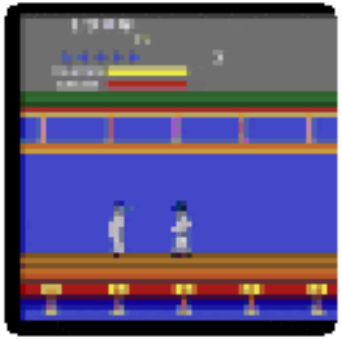
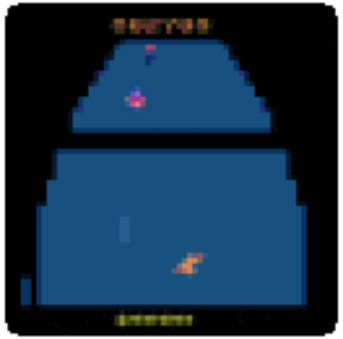
$$\text{normalized performance} = \frac{\text{DQN score} - \text{random play score}}{\text{human score} - \text{random play score}}$$

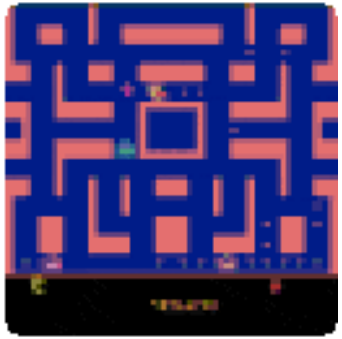
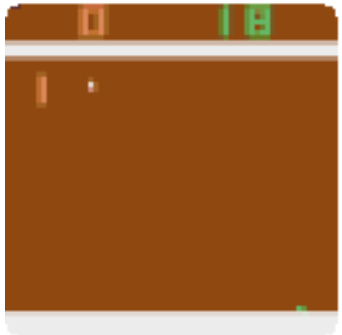
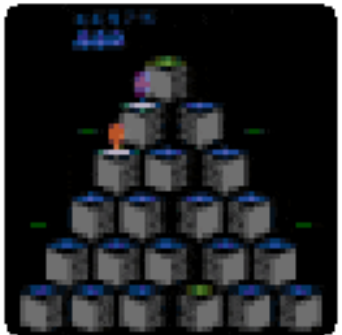
The performance of DQN is normalized with respect to a professional human games tester (that is, 100% level)

"Human-Level Control Through Deep Reinforcement Learning", Mnih, Kavukcuoglu, Silver et al. (2015)



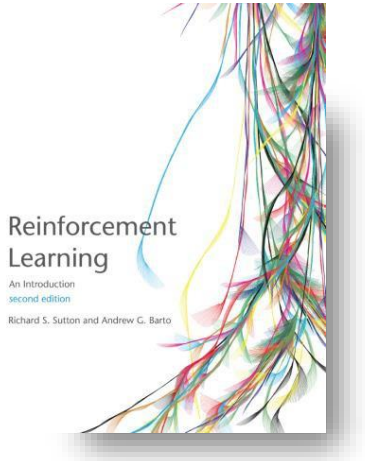



参考材料

- 参考书



Rich Sutton

Andew Barto

http://incompleteideas.net/book/RLbook2020.pdf

- 参考课程
  - UCL David Silver RL Course: https://www.davidsilver.uk/teaching/
  - Berkeley Sergey Levine Deep RL Course: http://rail.eecs.berkeley.edu/deeprlcourse/
  - OpenAI DRL Camp: https://sites.google.com/view/deep-rl-bootcamp/lectures
  - RL China Camp: http://rlchina.org/

谢谢大家