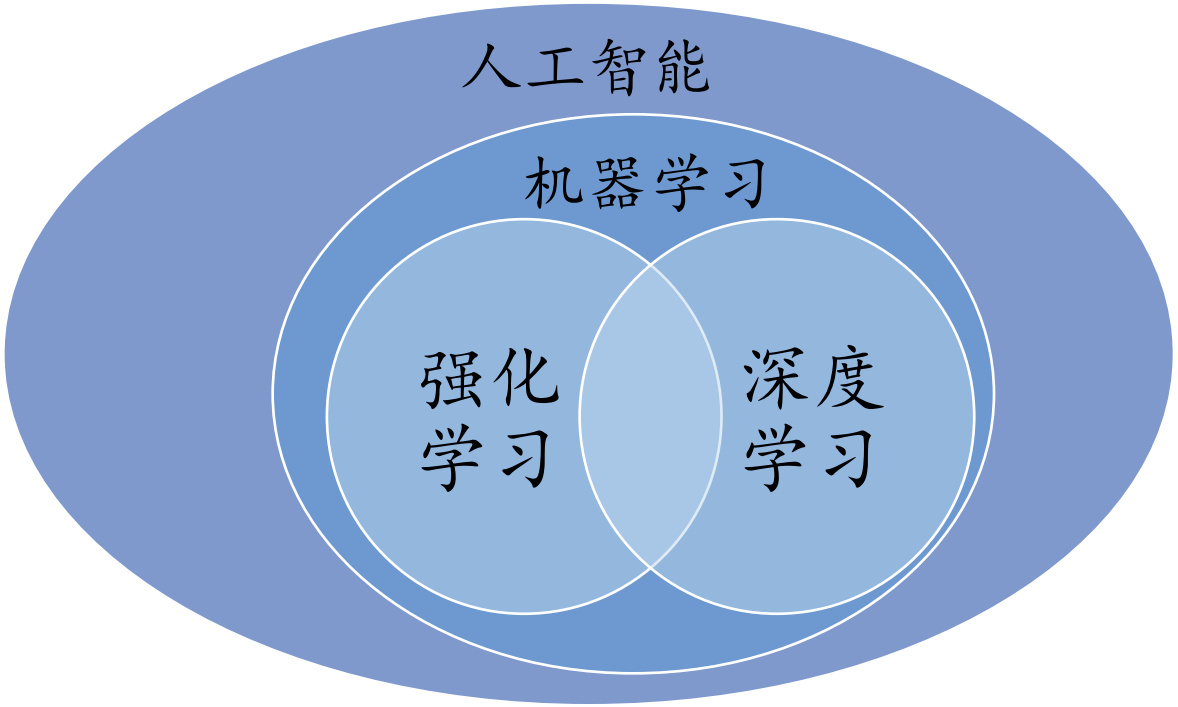
Reinforcement Learning & Optimal Control

强化学习与最优控制

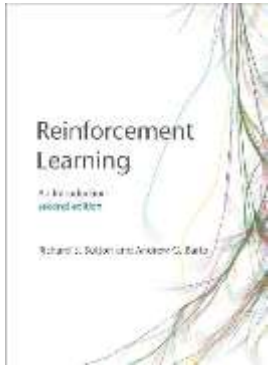
引言

关系与层次

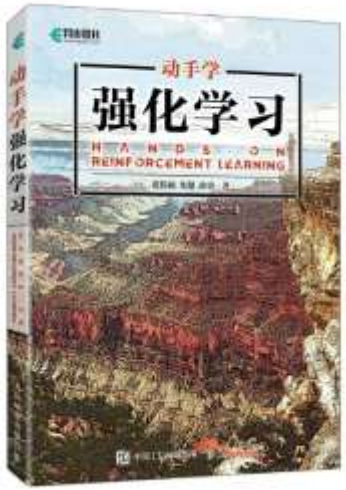


推荐教材

- Sutton & Barto, 《**Reinforcement Learning: An Introduction**》,
  - http://incompleteideas.net/book/RLbook2020.pdf



- 张伟楠，沈健，余勇. 《**动手学强化学习**》



推荐课程（英文）

- UCL David Silver RL Course: https://www.davidsilver.uk/teaching/
  - 课程视频：Introduction to Reinforcement Learning （10节课）
    https://www.bilibili.com/video/BV17x411Z7Zo?zw
- 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/

推荐课程（中文）

- 李宏毅“机器学习（强化学习部分）”
  - https://www.bilibili.com/video/av94519857/
- 张伟楠，SJTU, RL course
  - 课程主页：https://wnzhang.net/teaching/sjtu-rl-2024/
- 俞扬，南京大学， 强化学习课程：
  - 课程主页：https://www.lamda.nju.edu.cn/introrl/
- 赵世钰， 西湖大学，强化学习的数学原理
  - https://www.bilibili.com/video/BV1sd4y167NS?spm_id_from=333.788.videopod.episodes

课程成绩

- 平时作业（1-2次）：
  - 占30%
- 期末大作业：
  - 占40%
- 课堂表现：
  - 占30%

课程大纲

- 绪论
- MDP与动态规划
- 值函数估计
- 无模型方法
- 规划与学习
- 参数化值函数与策略
- 深度强化学习价值方法
- 深度强化学习策略方法
- 探索与利用
- 基于模型的深度强化学习
- 模仿学习
- 离线强化学习
- 多智能体强化学习
- 前沿：LLM+RL、DM+RL

前导课程/预备知识

- 线性代数
- 概率论
- 机器学习、深度学习

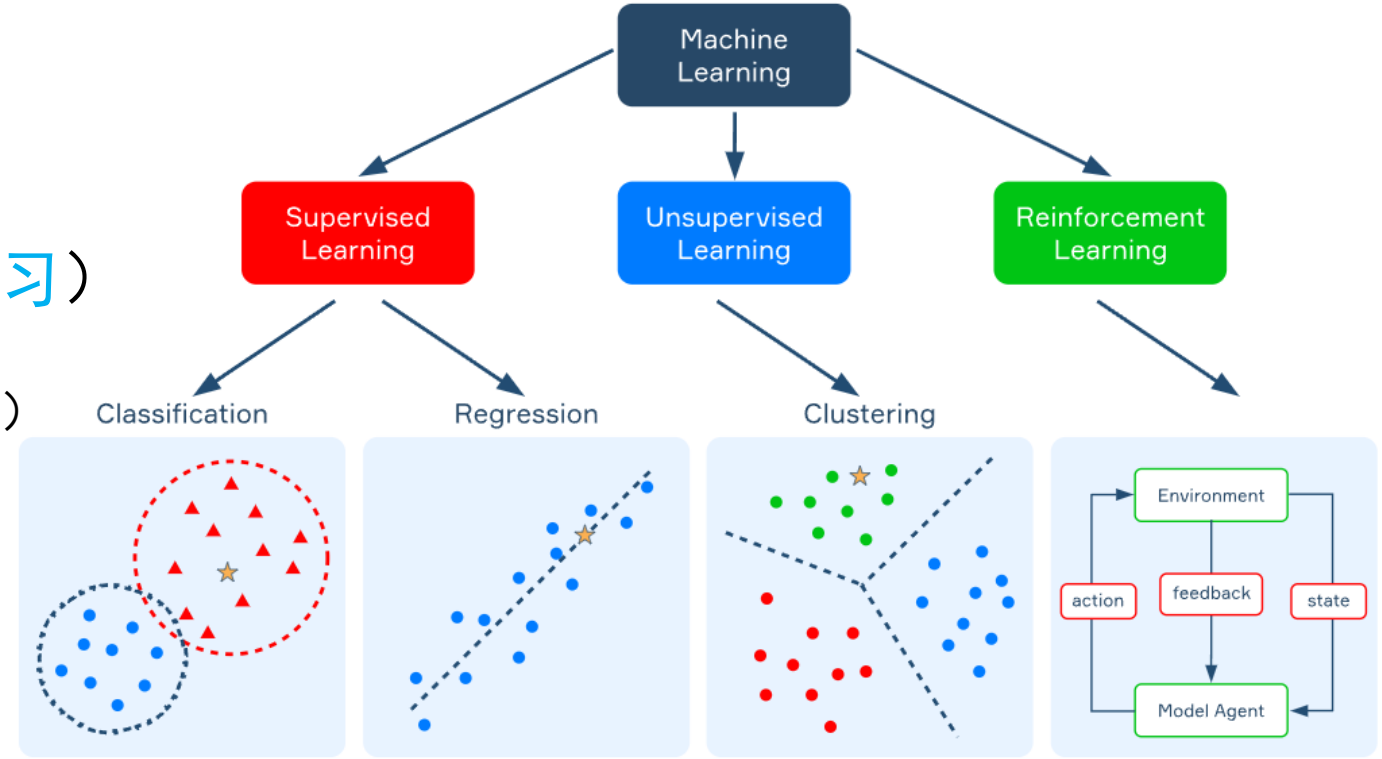
- 凸优化
- 信息论

第1章 绪论

目录

两种人工智能任务类型

- 预测型任务
  - 根据数据预测所需输出（监督学习）
  - 聚合/生成数据实例（无监督学习）
- 决策性任务
  - 在环境交互式动作（强化学习）
    - 转变到新的状态
    - 获得即时奖励
    - 随着时间的推移最大化累计奖励
    - Learning from interaction in a trial-and-error manner



https://hyperskill.org/learn/step/10403

决策智能的任务和技术分类

- 根据决策环境的动态性和透明性，决策任务大致分为以下四个部分

环境特性	白盒环境 ● 变量和目标之间的关系可以用具体公式表示	黑盒环境 ● 变量和目标之间的关系无法用具体公式表示
静态环境 ● 无状态转移 ● 单步决策	运筹优化 • （混合整数）线性规划 • 非线形优化	黑盒优化 • 神经网络替代模型优化 • 贝叶斯优化
动态环境 ● 有状态转移 ● 多步决策	动态规划 • MDP直接求解 • 树、图搜索	强化学习 • 策略优化 • Bandits、序贯黑盒

序贯决策（Sequential Decision Making）

- 序贯决策中，智能体序贯地做出一个个决策，并接续看到新的观测，直到最终任务结束

机器狗例子：操作轮足和地形持续交互，完成越过障碍物的任务

绝大多数序贯决策问题，可以用强化学习来解

应用案例

- 自动驾驶

Alex Kendall et.al, Learning to Drive in a Day. ICRA 2019: 8248-8254
https://www.youtube.com/watch?v=eRwTbRtnT1I

应用案例

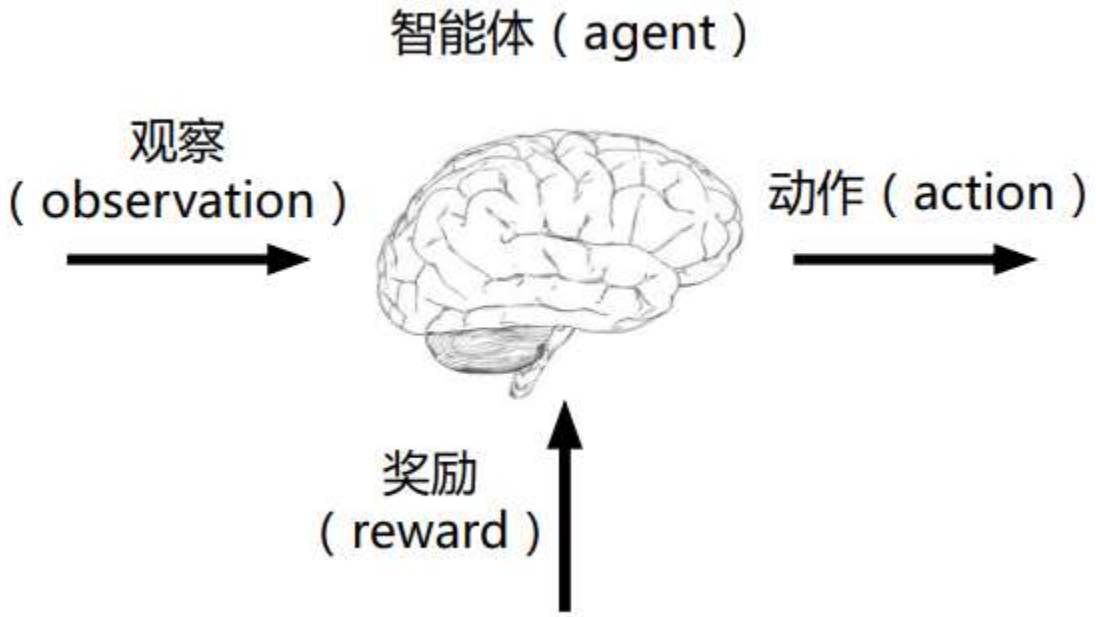
- 机械臂操控

CoRL 2023: Finetuning Offline World Models in the Real World

目录

强化学习定义

- 通过从交互中学习来实现目标的计算方法



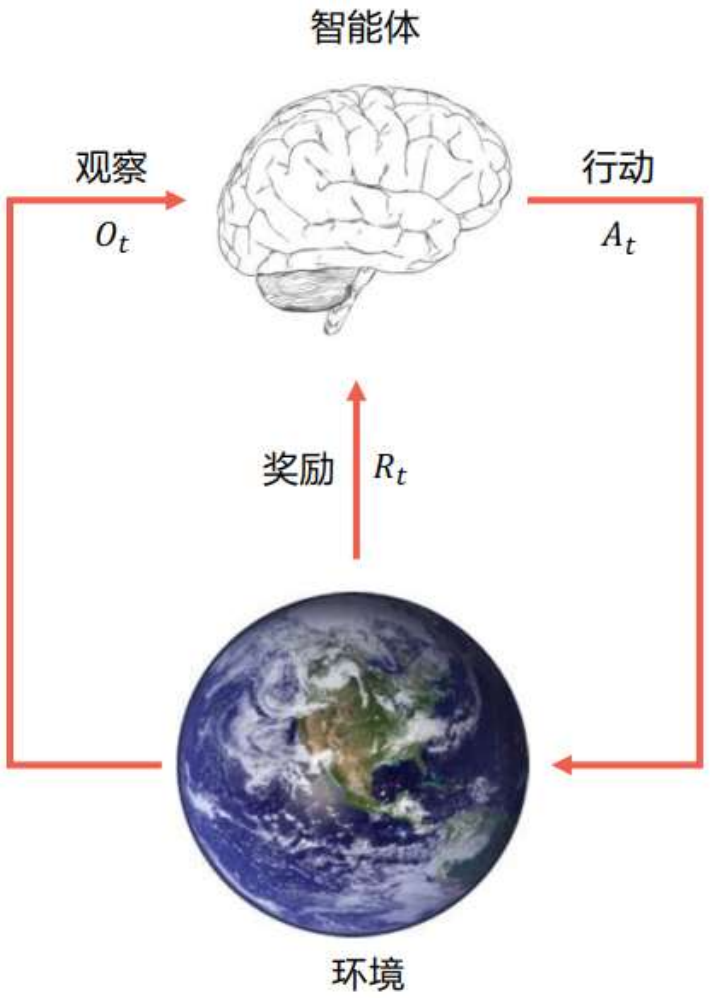
三个方面：

感知：在某种程度上感知环境的状态（state, reward）

动作（action）：可以采取动作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 （total reward）

强化学习交互过程



- 在每一步 t，智能体：
  - 获得观察 O_t/S_t
  - 执行动作 A_t
  - 获得奖励 R_t
- 环境：
  - 获得动作 A_t
  - 给出奖励 R_t
  - 转移、给出状态 O_t/S_t
- 交互步长： $t=t+1$

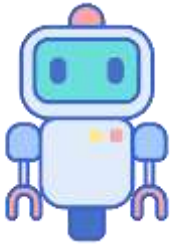
在与动态环境的交互中学习

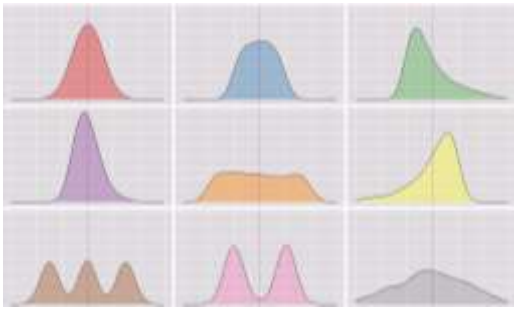
- 有监督、无监督学习

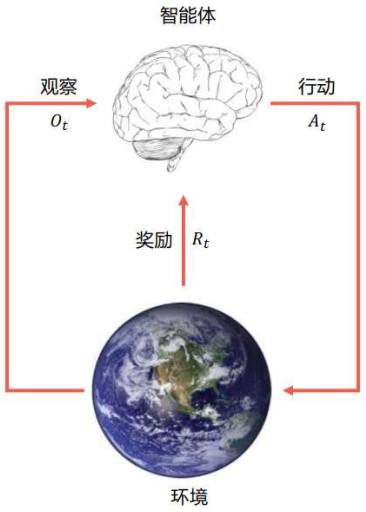
泛化模型

固定数据分布（IID假设）

- 强化学习







Agent不同，数据分布不同

强化学习关键要素

- 历史/轨迹（History/Trajectory）是观察、动作和奖励的序列

$$H_t = < S_1, A_1, R_1, S_2, A_2, R_2, \cdots, S_{t-1}, A_{t-1}, R_{t-1}, S_t, A_t, R_t >$$

  - 即，一直到时间t 为止的所有可观测变量
  - 根据这个历史可以决定接下来会发生什么
    - 智能体选择动作
    - 环境选择观察和奖励

- 状态（state）是一种用于确定接下来会发生的事情（动作、观察、奖励）的信息

  - 状态是关于历史的函数

$$S_{t+1} = f(H_t)$$

强化学习关键要素

- 策略（Policy）是学习智能体在特定时间的行为方式
  - 是从状态到动作的映射
    - 确定性策略（Deterministic Policy）

$$a_t = \pi(s_t)$$

    - 随机策略（Stochastic Policy）

$$\pi(a_t|s_t) = P(A_t = a_t|S_t = s_t)$$

- 奖励（Reward）R_t 和 $r(s,a)$
  - 一个用于定义强化学习目标的标量，能立即感知到什么是“好”的（**即时奖励**）

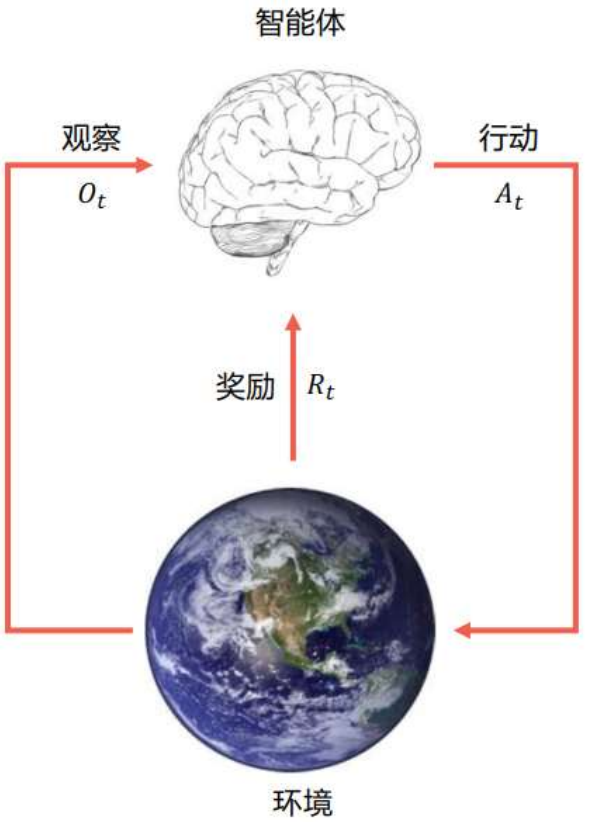
强化学习关键要素

- 环境的动态模型（model）刻画环境的“行为”
  - 转移到下一个状态

$$\mathcal{P}_{s,a}^{s'} = P[S_{t+1} = s' | S_t = s, A_t = a]$$

  - 提供一个即时奖励

$$\mathcal{R}_{s,a} = \mathbb{E}[R_t | S_t = s, A_t = a]$$



强化学习目标

- 强化学习中，智能体的学习目标为：
  - 在和环境持续交互的过程中，最大化期望累计奖励总和

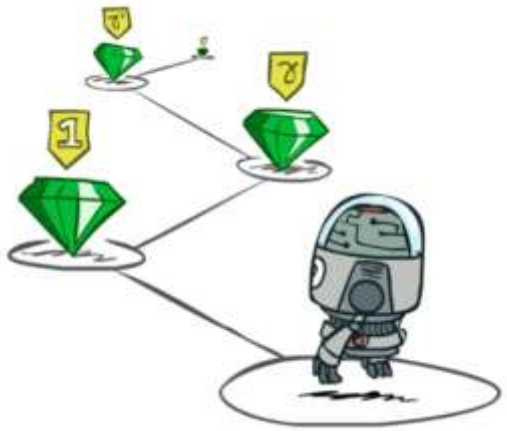
优化长期来看的“好”

$$\max_{\pi} \mathbb{E}_{H_T^{\pi}}[R_0 + \gamma R_1 + \cdots + \gamma^{T-1} R_{T-1}]$$

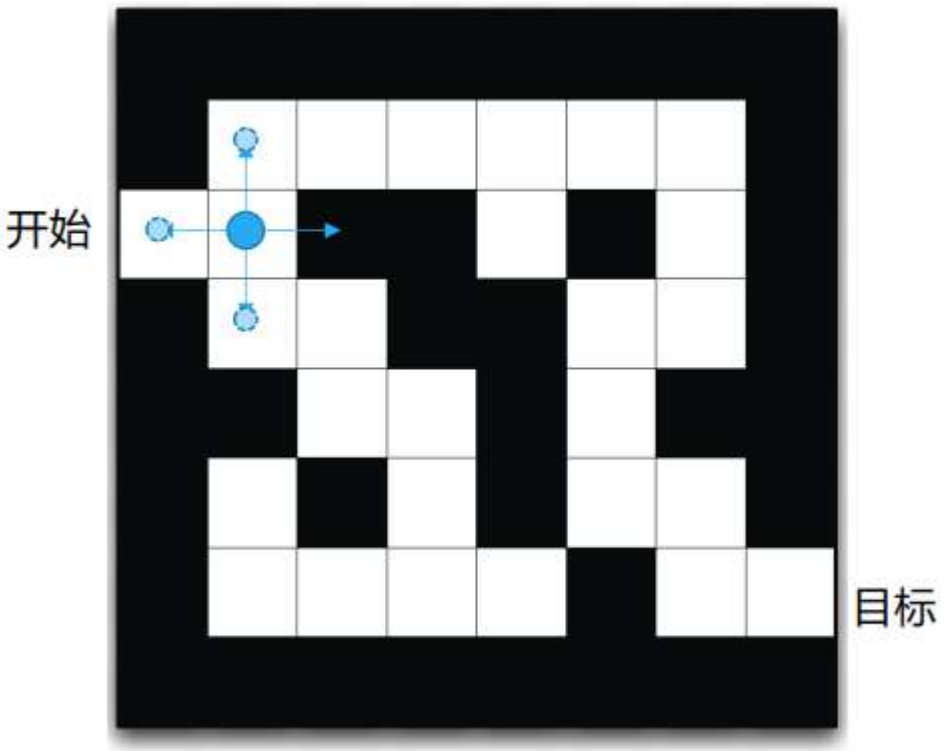
$$= \mathbb{E}_{H_T^{\pi}}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)\right]$$

- 折扣系数

$$\gamma = 1 \ ? \ \gamma \in (0, 1)$$

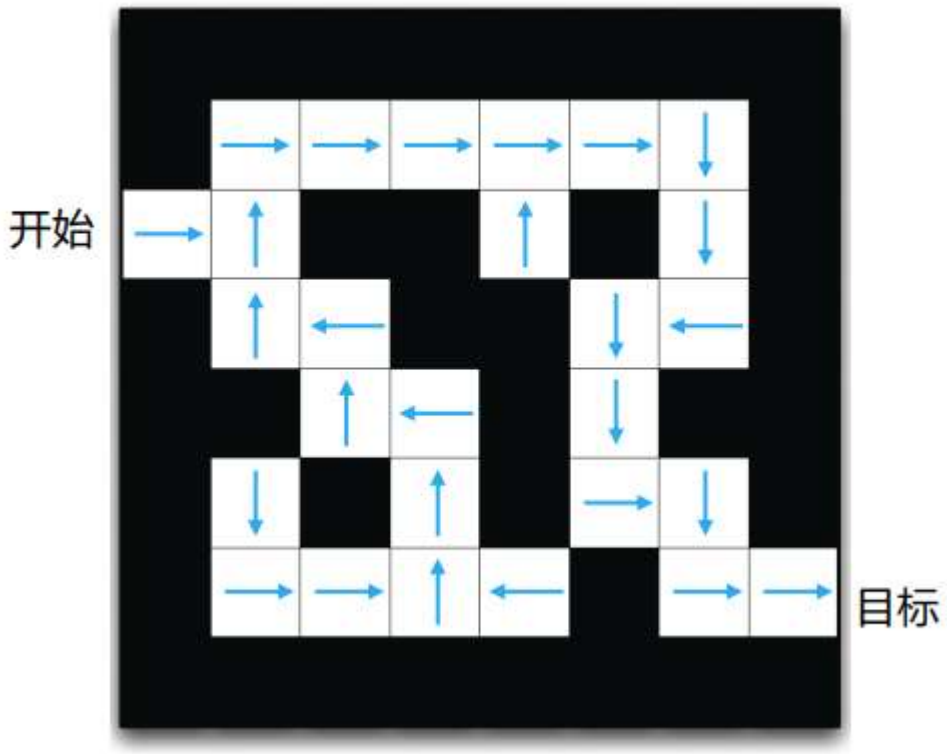


举例：迷宫



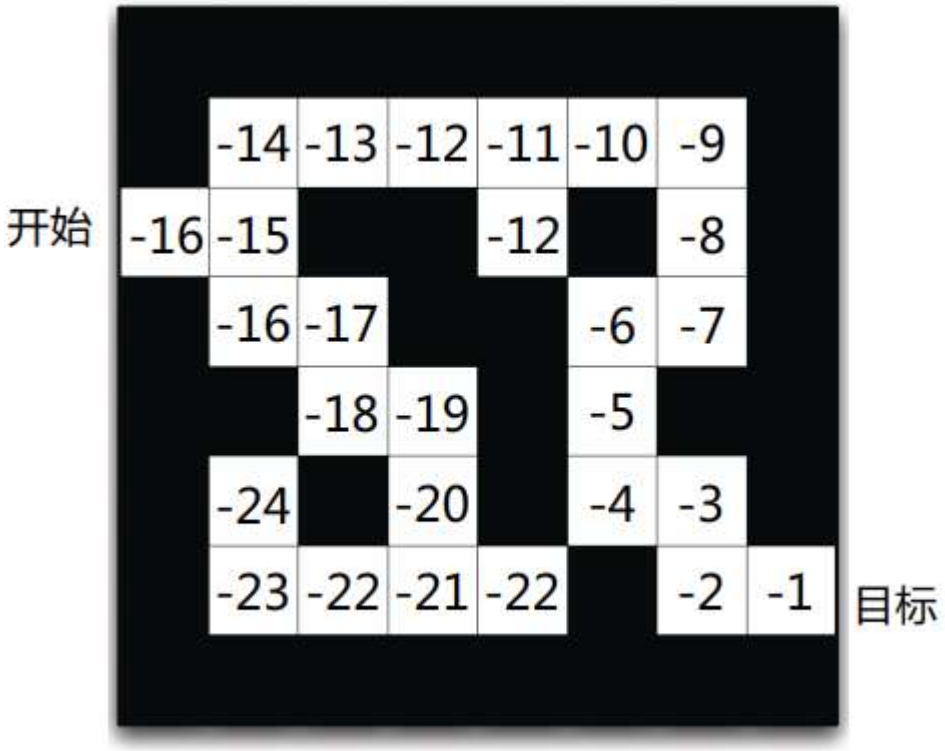
- 状态：智能体当前位置
- 动作：上、下、左、右
- 状态转移：根据动作方向移动一格
  - 如果下一格是墙则不动

举例：迷宫



- 状态：智能体当前位置
- 动作：上、下、左、右
- 状态转移：根据动作方向移动一格
  - 如果下一格是墙则不动
- 奖励：每一步为-1
- 策略：箭头方向
  - 每一个状态下的策略 $\pi(s)$

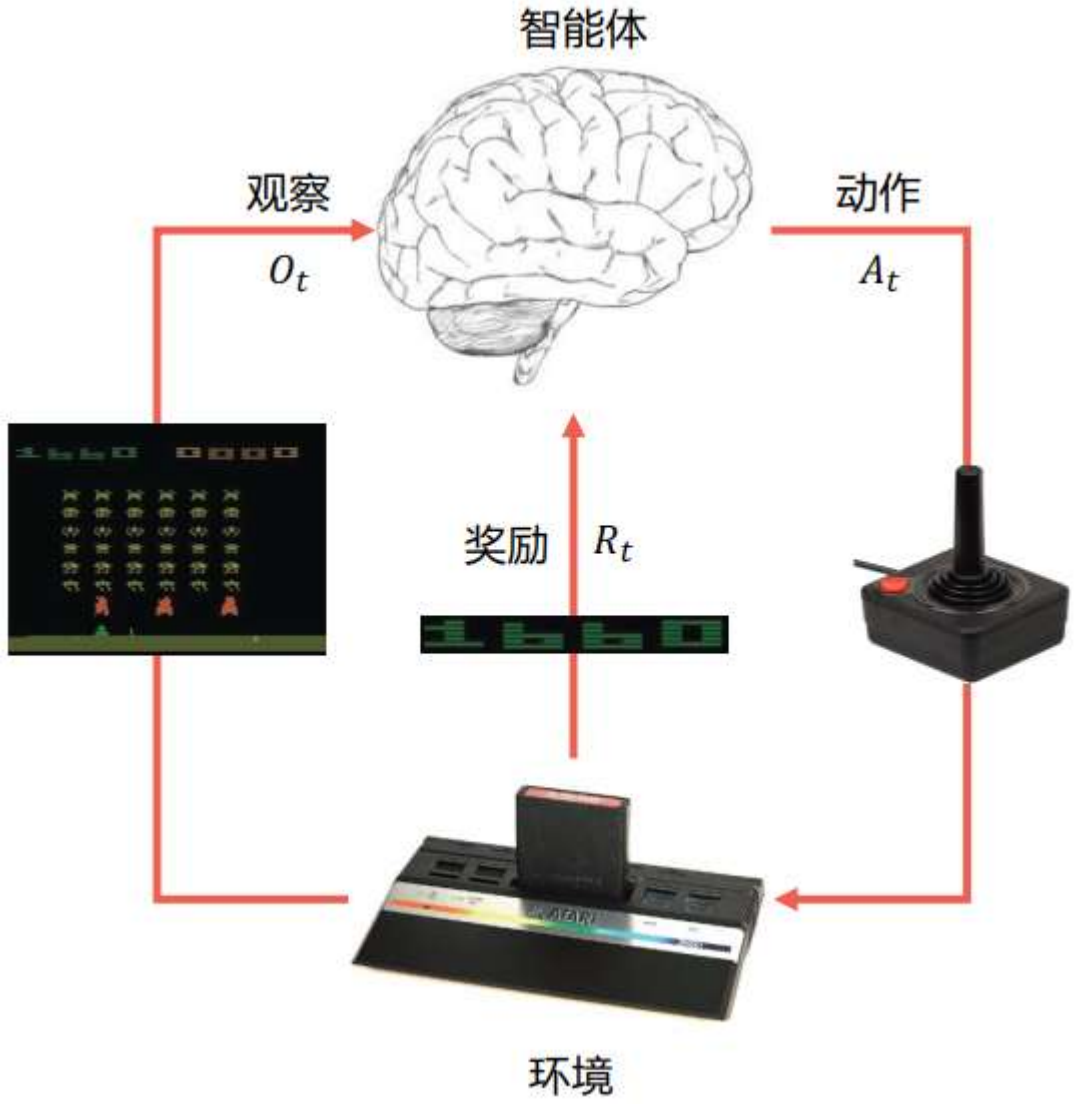
举例：迷宫



- 数字表示每个状态价值 $V_\pi(s)$

- 状态：智能体当前位置
- 动作：上、下、左、右
- 状态转移：根据动作方向移动一格
  - 如果下一格是墙则不动
- 奖励：每一步为-1

举例：Atari游戏



- 游戏规则未知
- 从交互游戏中进行学习
- 在操纵杆上选择动作并查看分数和状态图像

动态规划求解

- （策略）价值是一个标量，表示当前策略下，对于长期来说什么是“好”的
  - 数学定义：从某个状态开始，执行策略获得的累积（折扣）奖励期望

$$Q^{\pi}(s,a)=\mathbb{E}_{H^{\pi}}\left[r_0+\underbrace{\gamma r_1+\gamma^2 r_2+\cdots}_{\gamma Q^{\pi}(s_1,a_1)}|s_0=s,a_0=a,\pi\right]$$

$$=r(s)+\gamma\sum_{s'\in S}P_{s,a}(s')\sum_{a'\in A}\pi(a'|s')Q(s',a')$$

Bellman等式

  - 算子（向量）形式

$$Q^{\pi}=R+\gamma P^{\pi}Q^{\pi}\Rightarrow Q^{\pi}=(I-\gamma P^{\pi})^{-1}R$$

策略迭代

强化学习方法分类

- 基于价值（Value-based）：知道什么是好的，什么是坏的
  - 没有（显式）策略
  - 价值函数
- 基于策略（Policy-based）：知道怎么行动
  - 显式策略
  - 无价值函数
- 演员-评论家（Actor-Critic）：学生听老师的
  - 有显示策略
  - 有价值函数

价值和策略近似

$Q_\theta(s,a)$

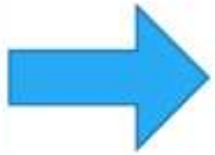
	a_1	a_2
s_1	3.2	4.1
s_2	0.8	1.2
s_3	2.3	-0.3

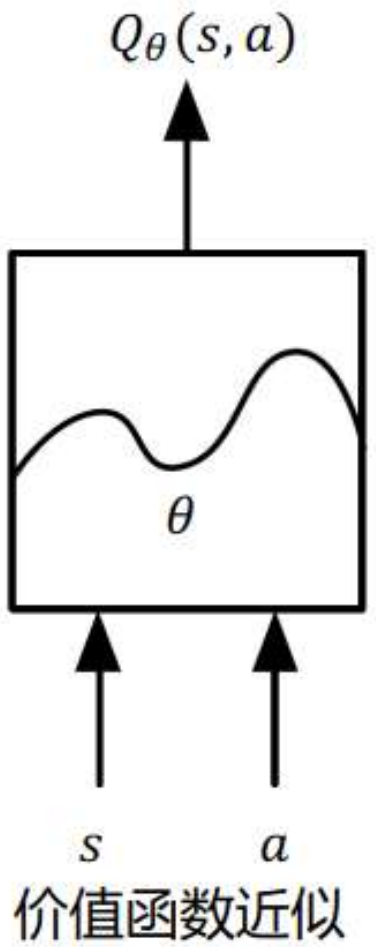
价值函数表格

$\pi_\theta(a|s)$

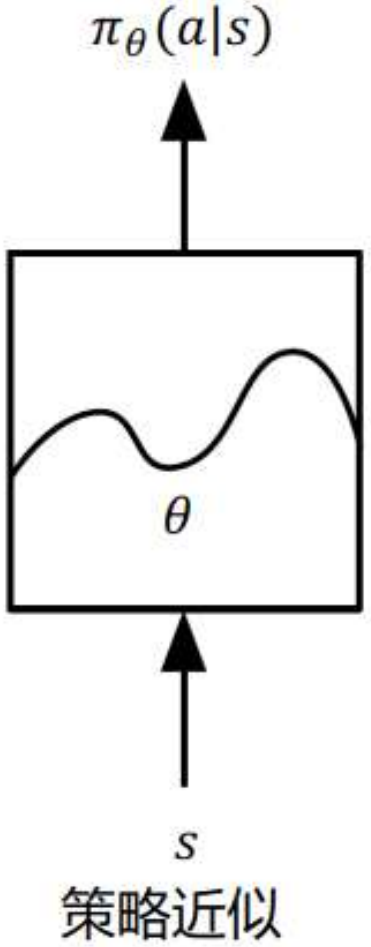
	a_1	a_2
s_1	0.2	0.8
s_2	0.4	0.6
s_3	0.9	0.1

策略表格





价值函数近似



策略近似

- 高维、黑盒环境中近似求解
- 深度神经网络近似求解→ 深度强化学习！（deep RL）

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

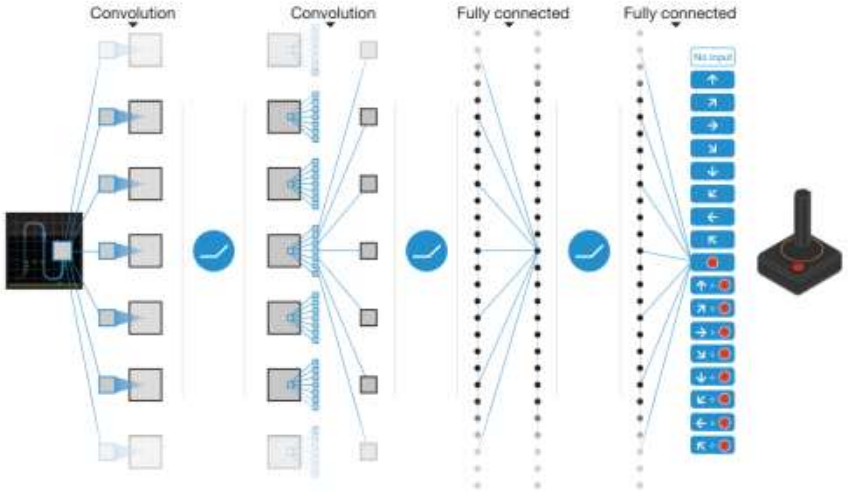
深度强化学习的崛起

- 突破性事件：Deepmind的Alpha系列智能体击败人类玩家

深度强化学习本质

深度强化学习

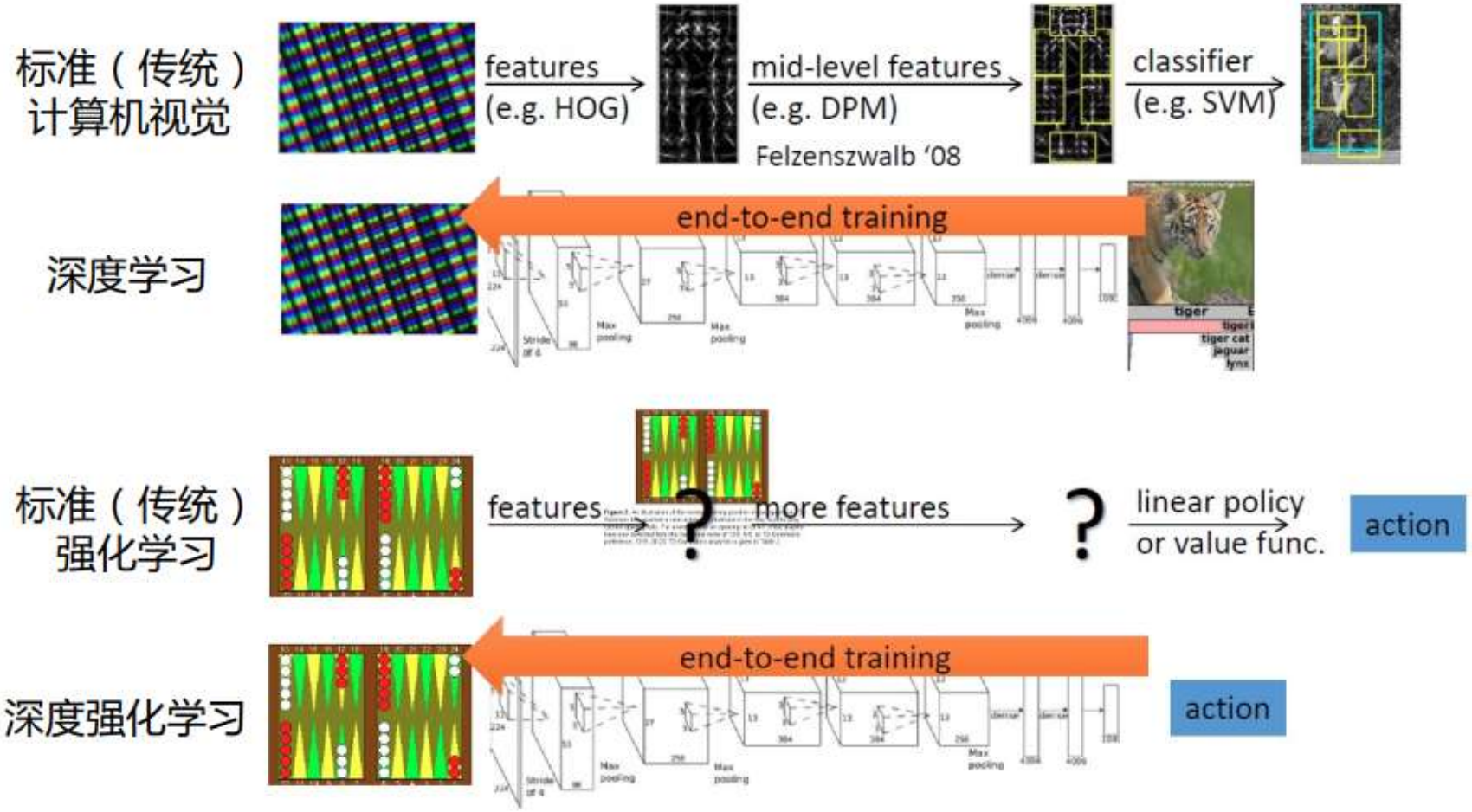
- 利用深度神经网络进行价值函数和策略近似
- 从而使强化学习算法能够以端到端的方式解决复杂问题



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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver et al. Playing Atari with Deep Reinforcement Learning. NIPS 2013 worksho

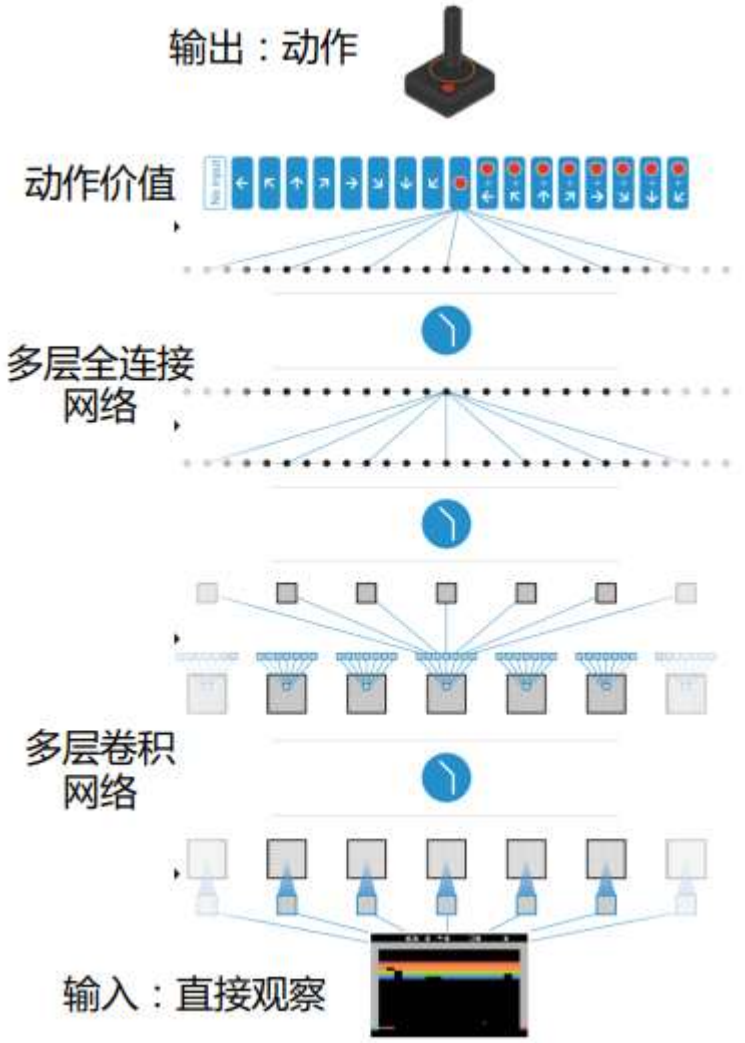
深度强化学习本质



- 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

深度强化学习关键变化

- DL+RL=?
  - 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
  - ……



深度强化学习前沿

基于模拟模型的强化学习
- 模拟器的无比重要性

目标策动的层次化强化学习
- 长程任务的中间目标是桥梁的基石

模仿学习
- 无奖励信号下跟随专家做策略学习

让强化学习算法更加高效

多智能体强化学习
- 分散式、去中心化的人工智能

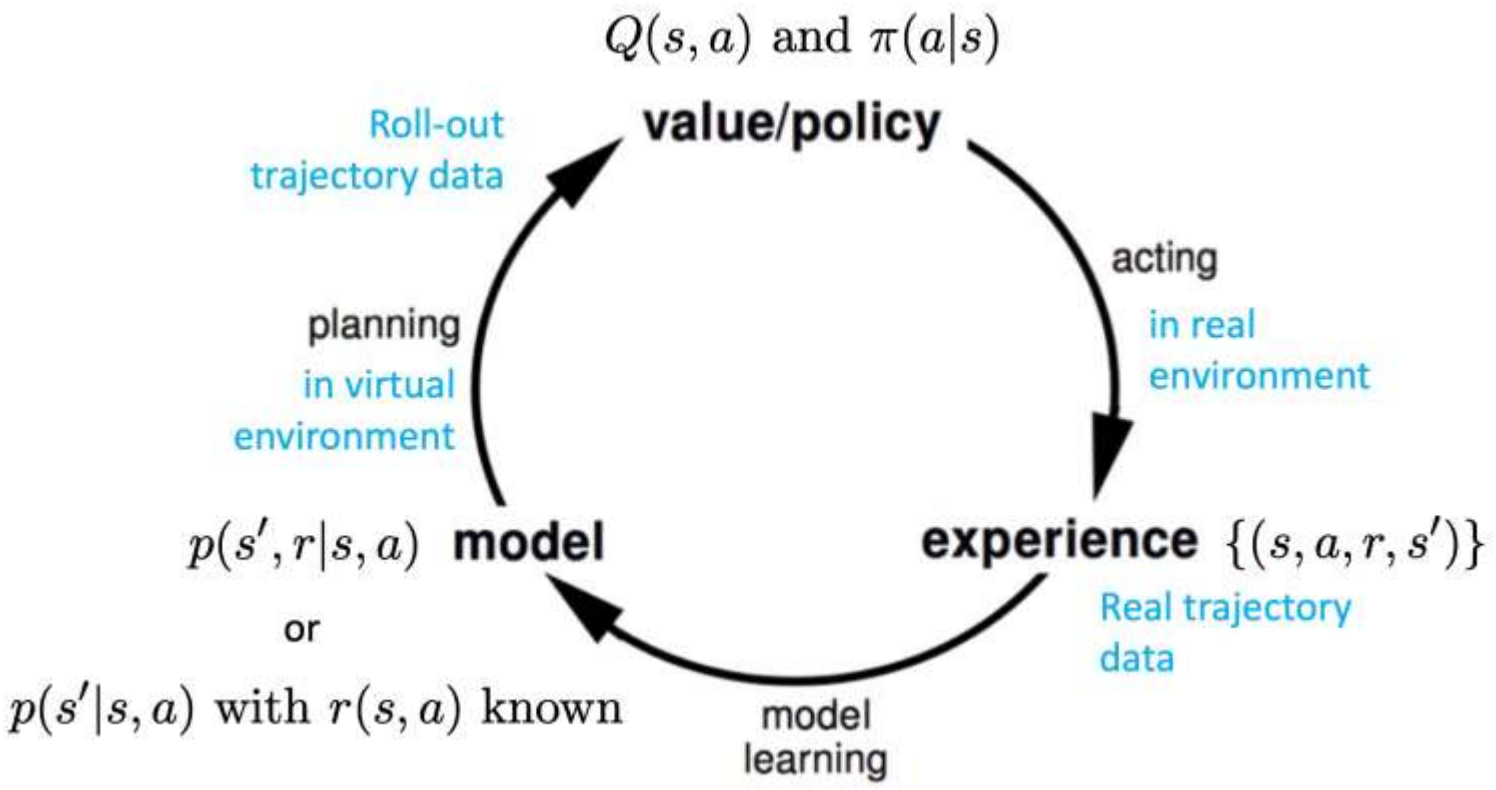
离线强化学习
- 训练过程中智能体不能和环境交互

让强化学习算法易于落地

强化学习决策大模型
- 探索以大的序列建模方式来完成序贯决策任务

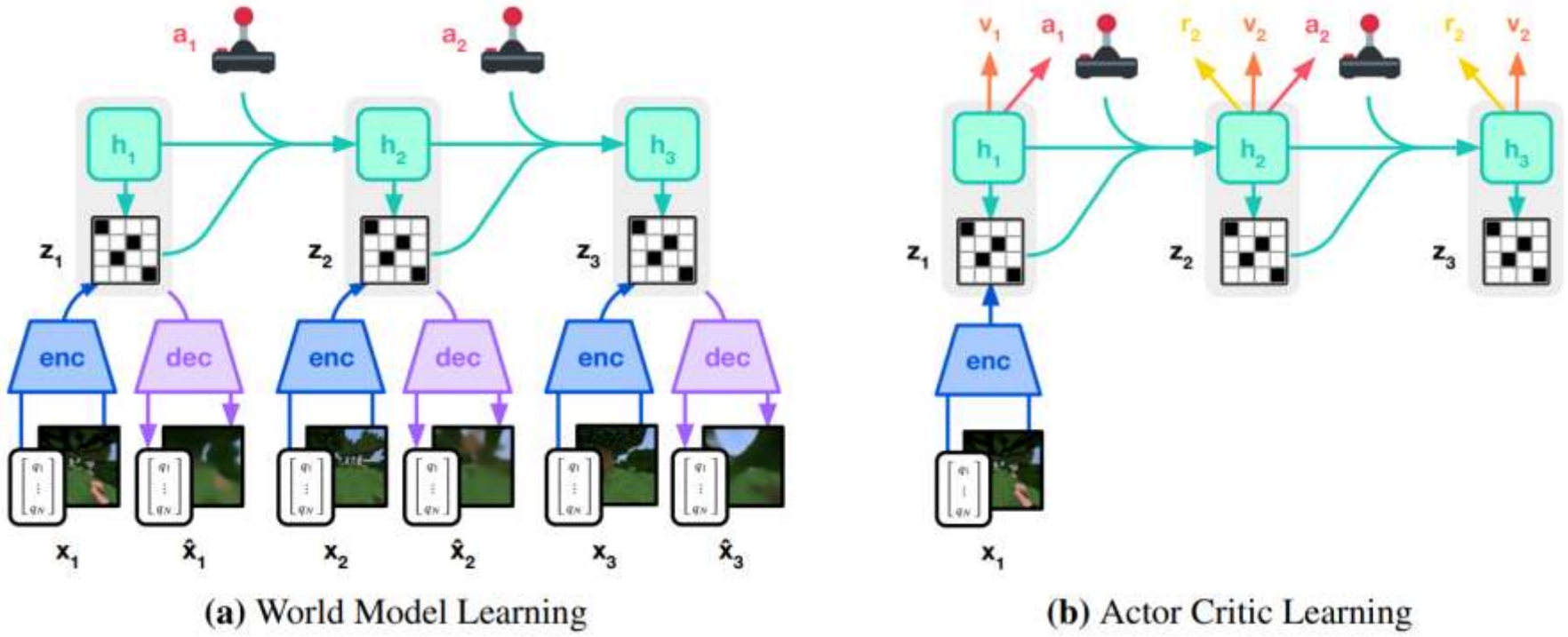
一项革新技术

基于模型的强化学习（Model-based RL）



- 建立环境模拟器，在模拟器中训练强化学习策略，减少对真实环境的影响，也可以生成更多特定场景数据

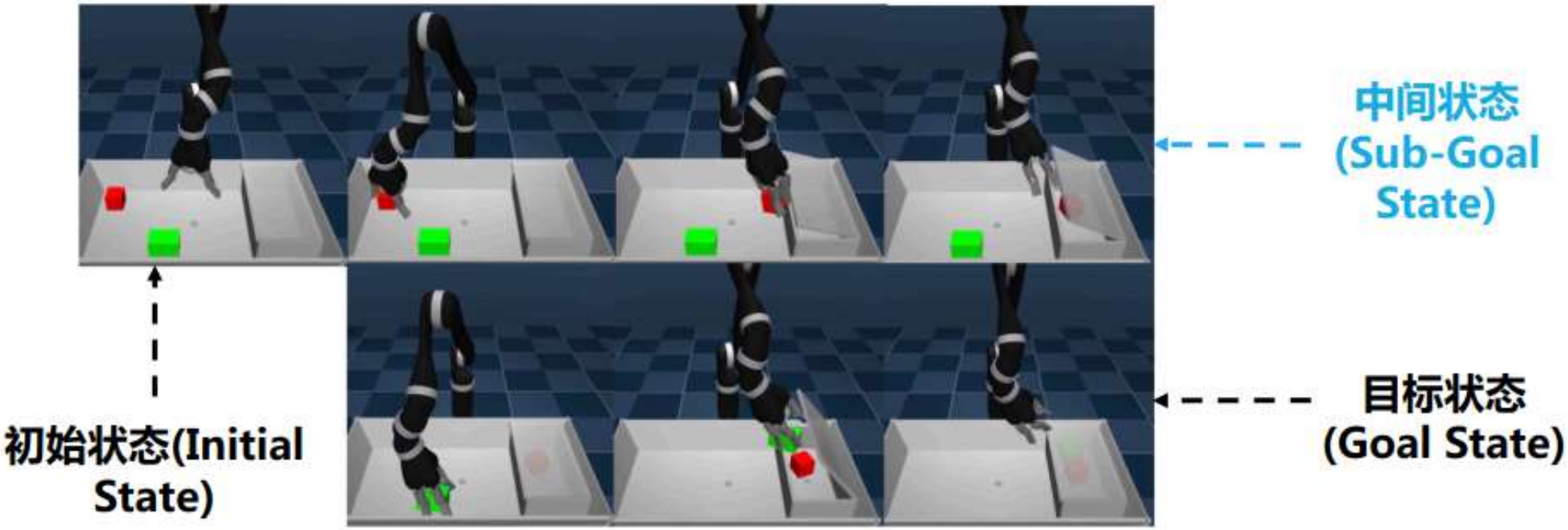
基于模型的强化学习（Model-based RL）



Hafner et al. Mastering Diverse Domains through World Models. 2023.

- 世界模型将感知输入编码为离散表示z，该表示由具有循环状态h的序列模型在给定动作a的情况下进行预测。
- Actor和Critic从由世界模型预测的抽象表示的轨迹中进行学习

目标策动的强化学习（Goal-oriented RL）

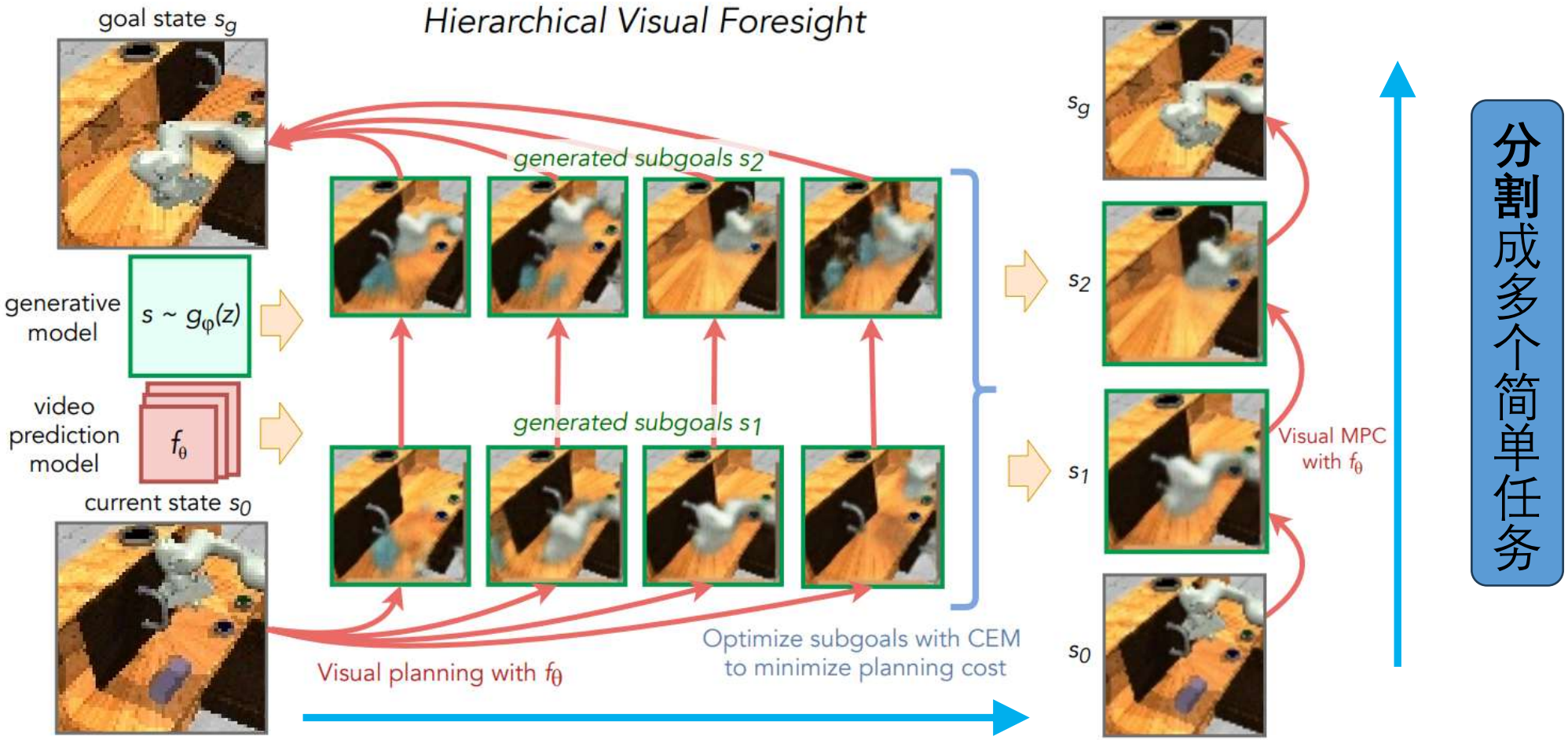


- 累计建模误差（compounding model error）
- 稀疏反馈（Sparse Reward）

核心思想：**生成**中间状态，将长期限任务（long-horizon task）**分割**成多个简单的短期限任务

目标策动的强化学习（Goal-oriented RL）

模拟器场景



分割成多个简单任务

生成中间状态

Suraj Nair, Chelsea Finn. Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation. ICLR, 2020

模仿学习（Imitation Learning）

Computer Games

reward

Mnih et al. '15

Real World Scenarios

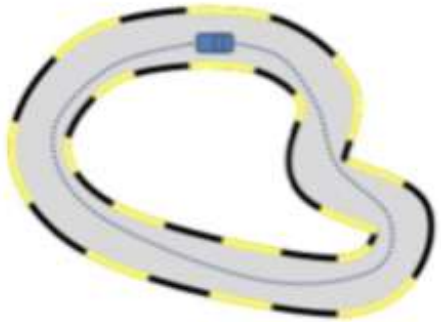
robotics

dialog

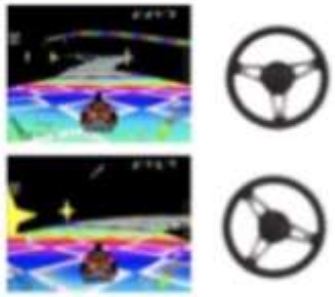
autonomous driving

what is the reward?
often use a proxy

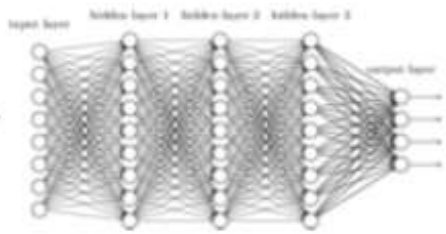
Expert Demonstrations



(*s*, *a*) pairs

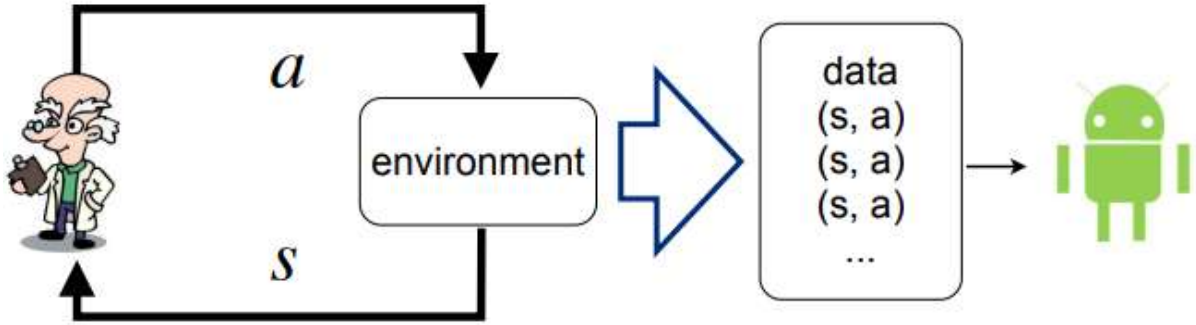


Imitation Learning

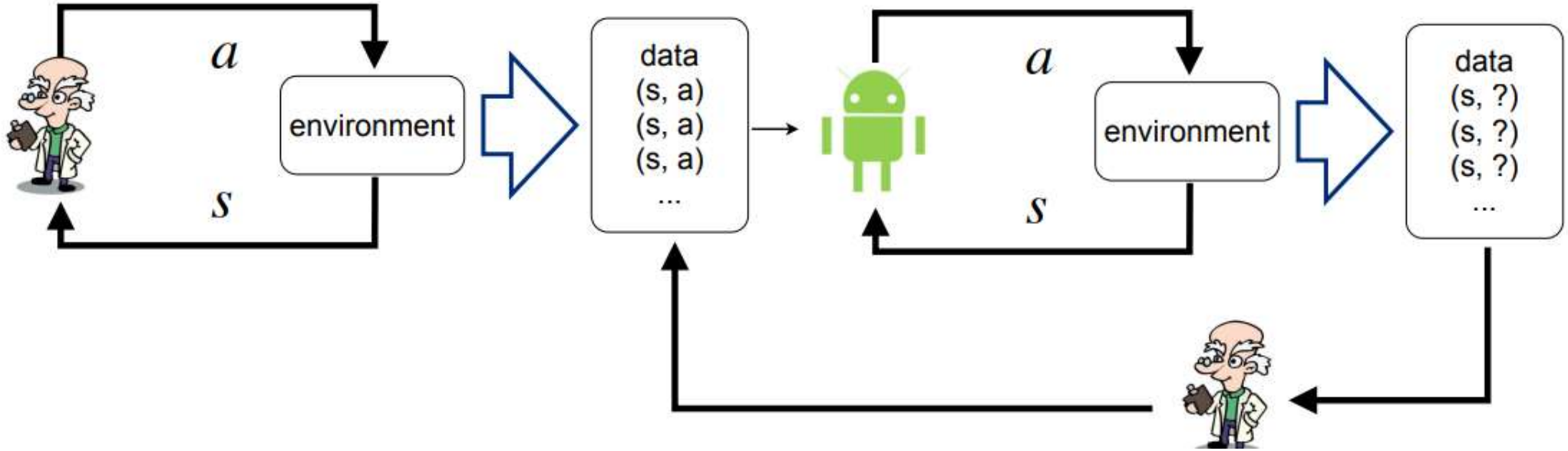


模仿学习（Imitation Learning）

简单模仿

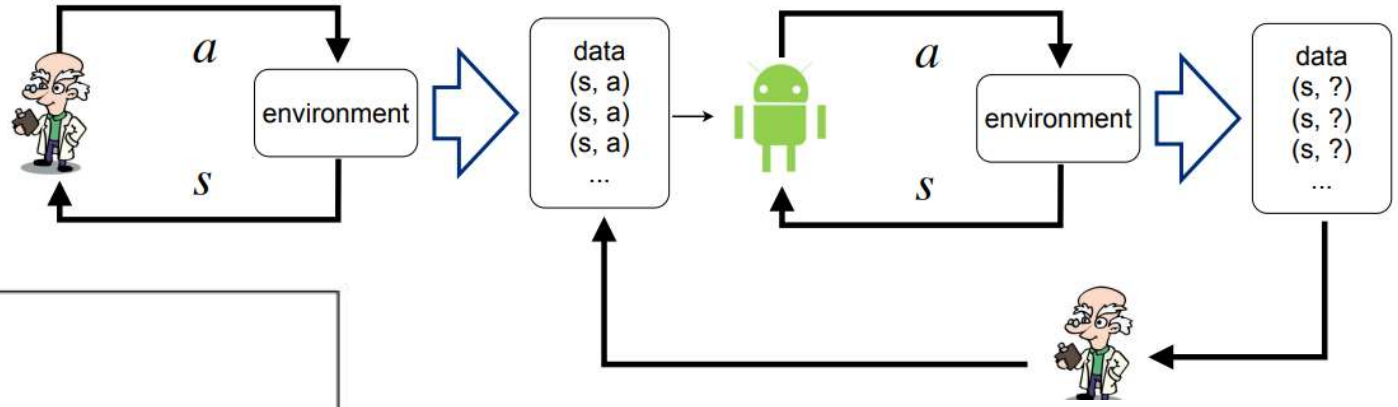


迭代收集更多数据



模仿学习（Imitation Learning）

DAGGER（Dataset Aggregation）



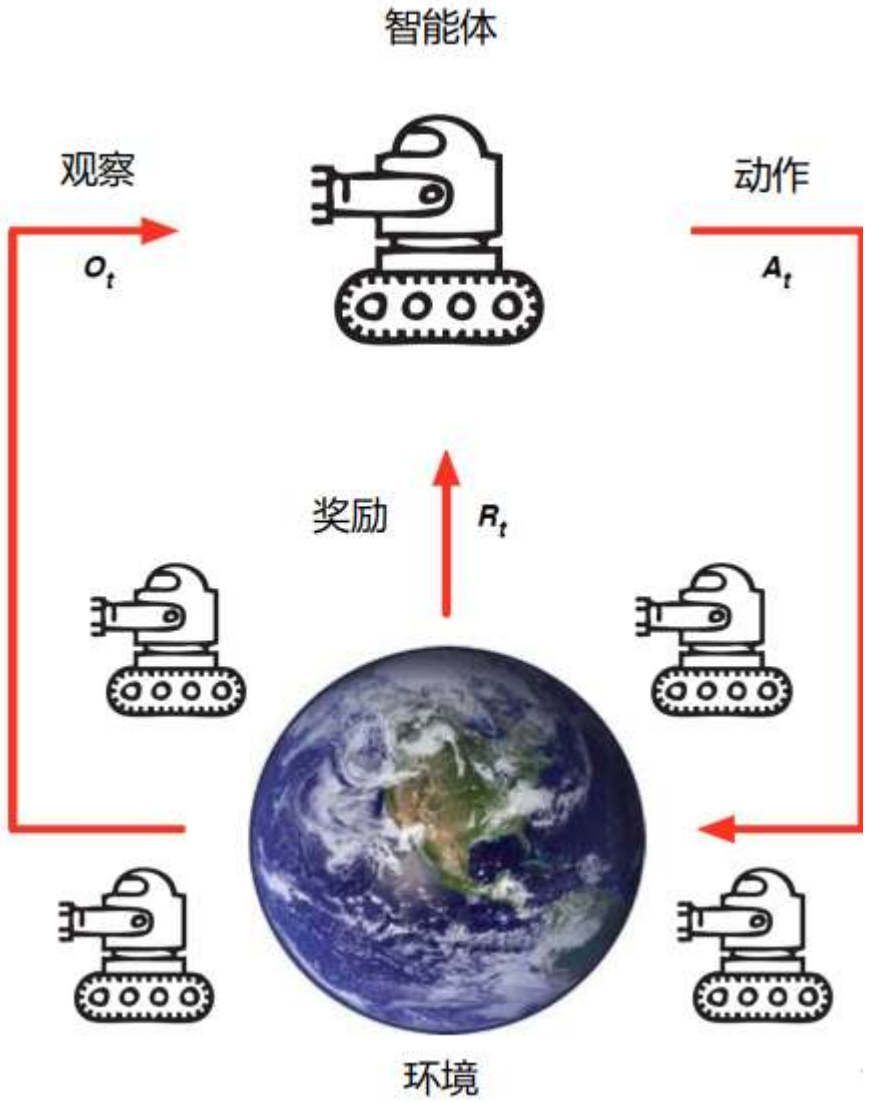
Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.
  Sample T-step trajectories using π_i.
  Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
end for
Return best $\hat{\pi}_i$ on validation.

Stéphane Ross, Geoffrey J. Gordon, Drew Bagnell: A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In: AISTATS 2011, 627-635

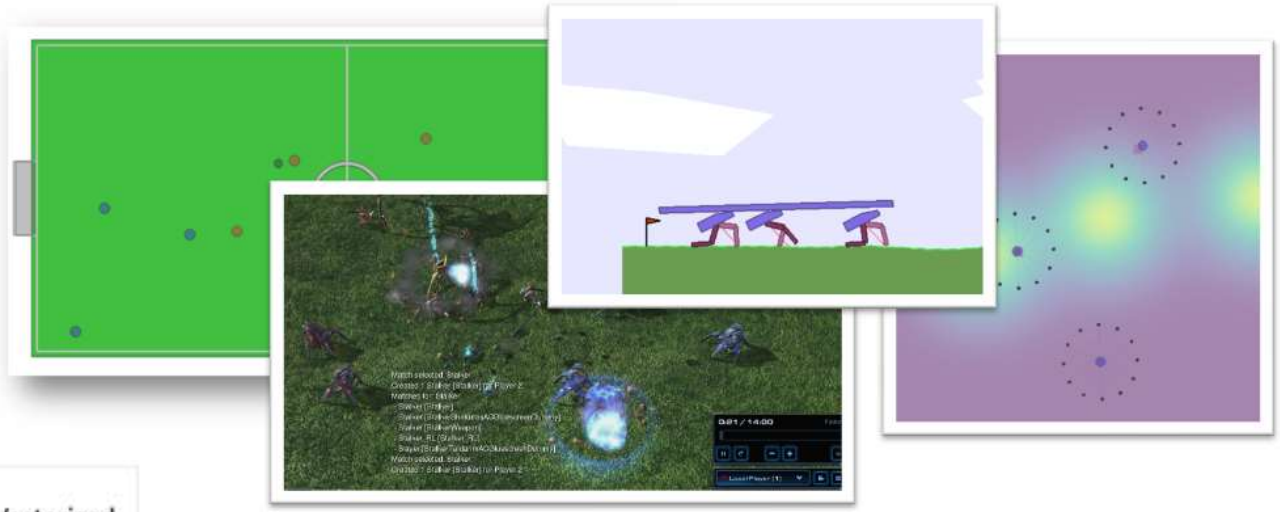
多智能体强化学习（Multi-Agent RL）

- 环境包含有不断进行学习和更新的其他智能体
- 在任何一个智能体的视角下，环境是非稳态的（nonstationary）
  - 环境迁移的分布会发生改变

合作？竞争？



多智能体强化学习（Multi-Agent RL）



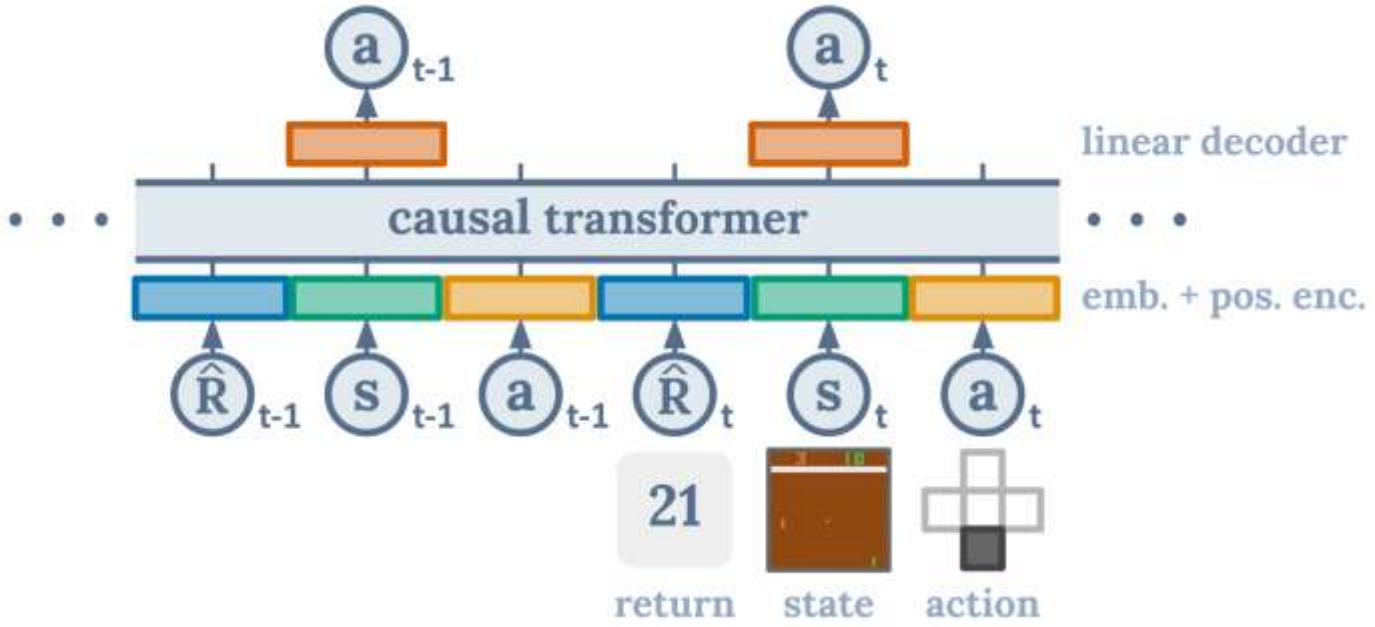
Environment	Tasks	Cooperation	Global state	Reward function	Action space	Vectorized
VMAS	27	Cooperative + Competitive	No	Shared + Independent + Global	Continuous + Discrete	Yes
SMACv2	15	Cooperative	Yes	Global	Discrete	No
MPE	8	Cooperative + Competitive	Yes	Shared + Independent	Continuous + Discrete	No
SISL	2	Cooperative	No	Shared	Continuous	No
MeltingPot	49	Cooperative + Competitive	Yes	Independent	Discrete	No
MAgent2	1	Cooperative + Competitive	Yes	Global in groups	Discrete	No

BenchMARL

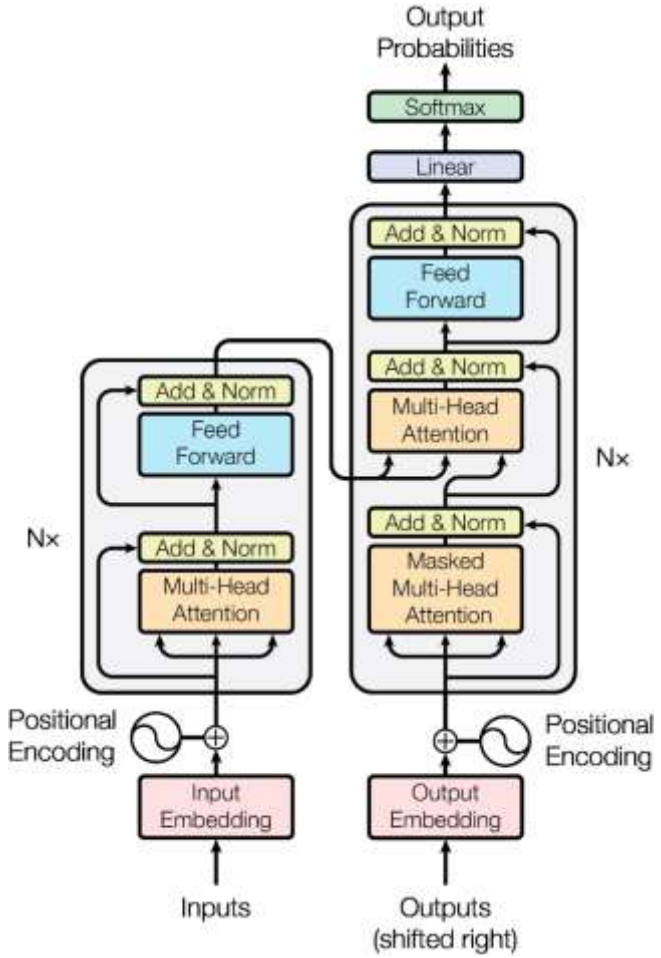
https://github.com/facebookresearch/BenchMARL

强化学习大模型（Large Model for RL）

- 把强化学习建模成一个序列预测问题；
- 使用Transformer类的大模型来做动作解码



Decision Transformer



Transformer编解码器

Chen, Lili, et al. “Decision transformer: Reinforcement learning via sequence modeling.” NeurIPS 2021

目录

强化学习落地场景

- 无人驾驶
- 游戏AI
- 交通灯调度
- 网约车派单
- 组合优化
- 推荐搜索系统
- 数据中心节能优化
- 对话系统
- 机器人控制
- 路由选路
- 工业互联网场景
- ...

强化学习应用案例

机器狗控制

Ziwen Zhuang, Zipeng Fu et al. Robot Parkour Learning. CoRL 2023.

强化学习应用案例

- 机器狗控制

宇树科技
https://www.bilibili.com/video/BV1CM4m1m74h/?spm_id_from=333.1387.homepage.video_card.click

强化学习应用案例

灵巧机械手控制

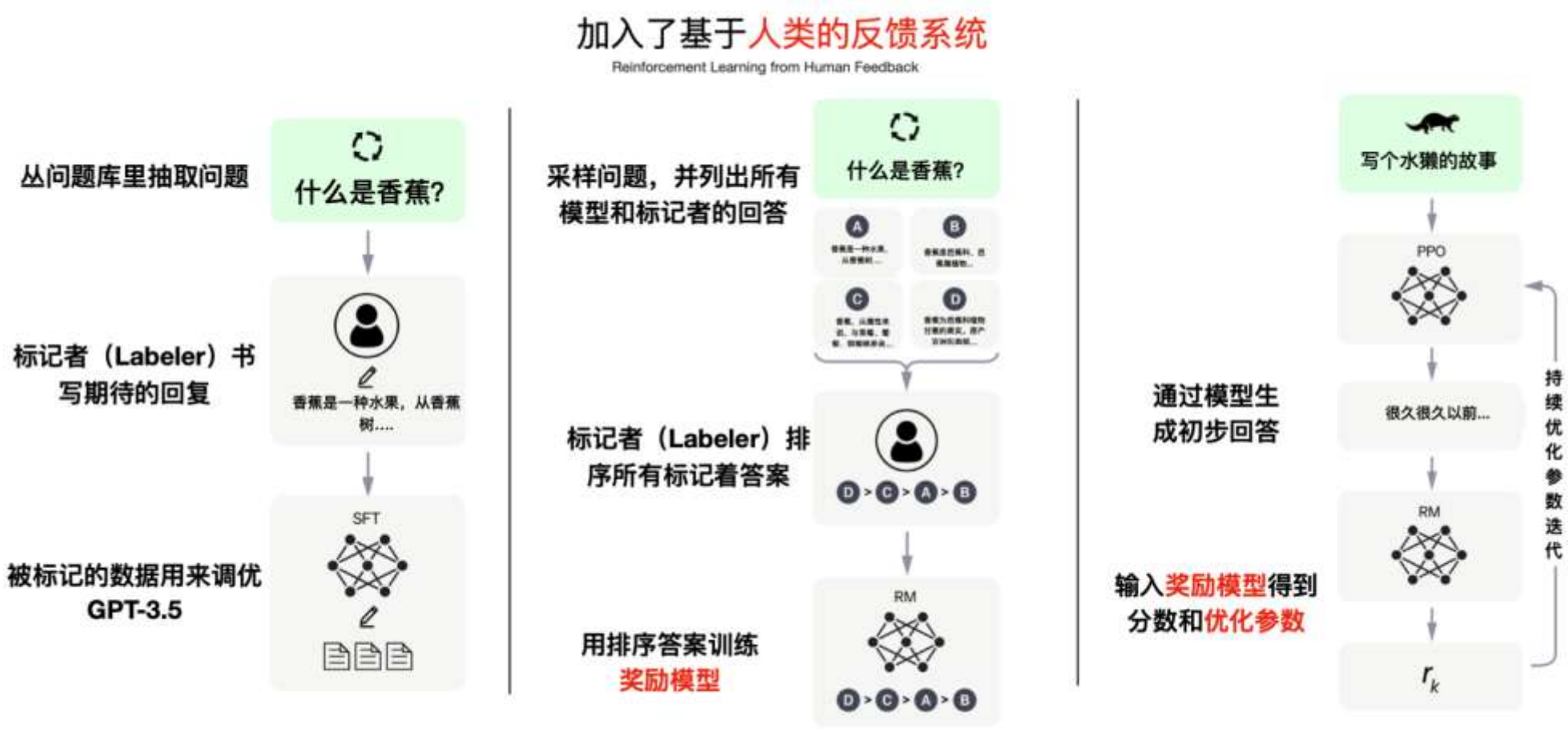
Solving Rubik‘s Cube with a Robot Hand, OpenAI, 2019

强化学习应用案例

- 基于模仿学习的移动操作

Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation
https://mobile-aloha.github.io/

ChatGPT中的强化学习



Ouyang et al. Training language models to follow instructions with human feedback. 2022. https://openai.com/research/instruction-following

总结

RL做什么

- 序贯决策任务
- 让AI做完一切事情，而不仅仅是一个辅助的角色

RL技术发展

- 2013年12月的NIPS workshop论文开启了深度强化学习时代
- 目前深度强化学习方法已经可以解决部分序列决策任务，但距离真正普及还有很长的路要走；在大模型时代强化学习大有可为

挑战是什么

- 决策权力交给AI，人对AI有更高的要求
- 强化学习技术人才短缺，决策场景千变万化，并不统一
- 当前强化学习算法对数据和算力的需求极大

Q & A