第10章：深度强化学习展望

周炜星 谢文杰

华东理工大学金融学系

2023年秋

纲要

纲要

源于学科交叉

- 深度强化学习源于几十年前的系统论、控制论、信息论、人工智能等领域的思想和技术，是人工智能的重要组成部分。
- 深度强化学习融合了表示学习、深度学习和智能决策模块，在复杂环境决策中表现出了较好的应用前景。
- 融合博弈论和计算实验模型进行模拟仿真，可为政府和组织决策提供更加贴合实际的策略支持，达到政策模拟和政策评估的目的。
- 深度强化学习融合了诸多学科的思想精华，是非常复杂而实用的研究领域和研究方向。

用于序贯决策

人类个体的复杂性和个体之间关系的复杂性使得个体层面、团体层面和系统层面的决策行为都极其复杂，特别是序贯决策问题的复杂程度更高。

- 深度强化学习算法是解决复杂系统环境中序贯决策问题的重要方法。
- 在智能决策过程中，如何表示主体属性和环境因素是主要问题。
- 在复杂系统背景下，深度强化学习框架具有普适性。
- 从微观到宏观、从个体到系统、从关联关系到因果关系、从理论到方法，深度强化学习模型都能够建模现实复杂环境中的智能决策问题。
- 深度强化学习方法在金融经济系统性风险度量、识别、传染、防控和预警研究中具有较大的应用潜力。

强于深度学习

在深度强化学习飞速发展过程中，深度学习模型功不可没，包括深度神经网络、卷积神经网络、循环神经网络、图神经网络等。不同的深度神经网络模型适用于不同的决策变量和环境状态特征变量。

- 在机器学习和深度学习领域中，表示学习是各个子领域的基础。
- 对智能体特征和环境特征进行表示学习，学习的特征变量作为智能决策的决策变量，有效的环境特征表示向量在解决复杂问题时能够起到事半功倍的效果。
- 在现实应用场景中，决策环境复杂多变，很多影响因素会因隐私和采集难度而不可获得或者不可量化。

纲要

游戏控制崭露头角

深度强化学习是一个较新的研究领域，但强化学习并非一个新的智能学习方法。

- 早在二十世纪五十年代，强化学习方法和思想就已经出现。
- 2013年，DeepMind 团队发表的研究论文“Playing Atari with deep reinforcement learning”引起了极大反响，智能体从像素级数据中学会了智能游戏控制策略，开启了一轮深度强化学习的热潮。
- 2015年，DeepMind研究人员在国际顶级学术期刊Nature发表研究论文“Human-level control through deep reinforcement learning”，各个领域研究人员对深度强化学习表现了极大兴趣

游戏控制崭露头角

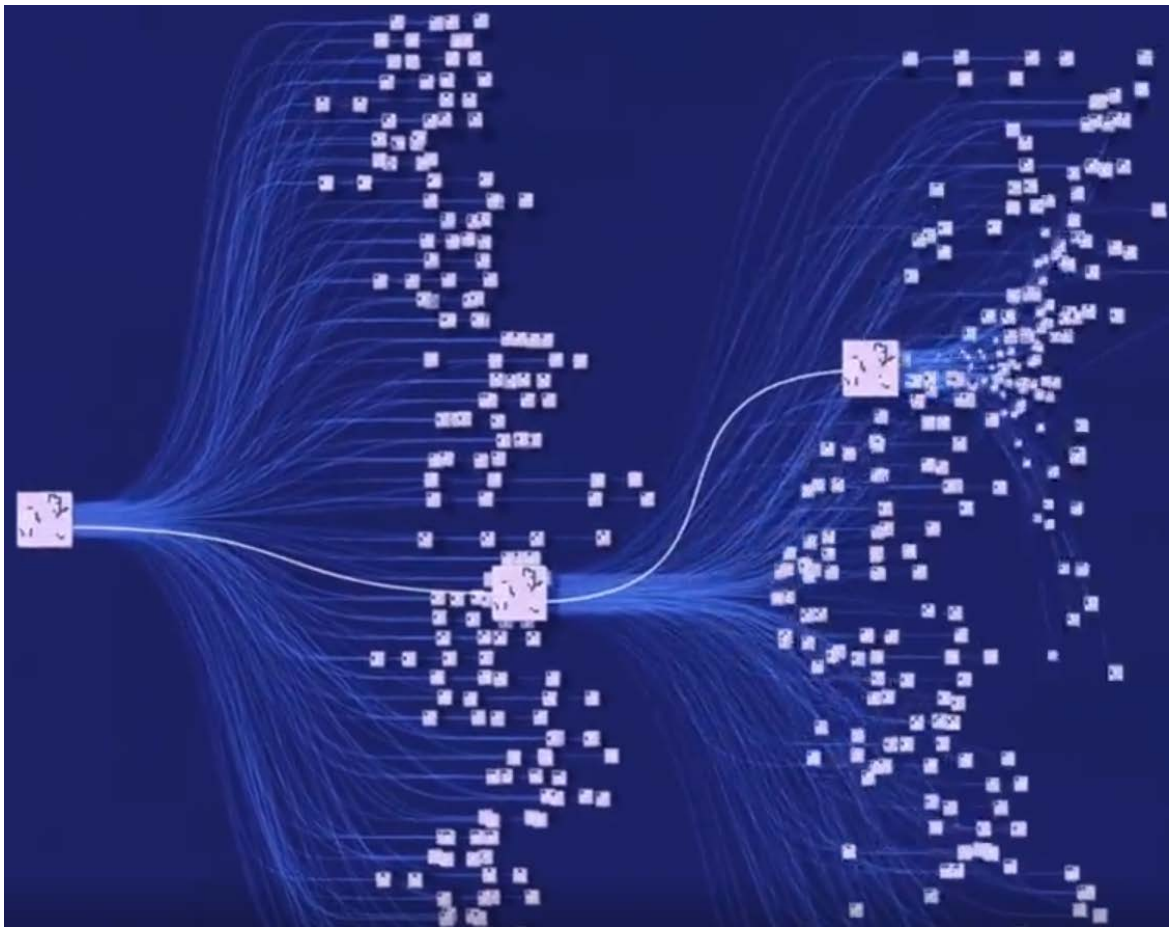


图 1: AIpha Go 基于蒙特卡罗树搜索示意图（图片来源视频截图）

AlphaGo风靡全球

- 2016年，DeepMind研究人员推出了AlphaGo，并在Nature杂志上发表了论文“Mastering the game of Go with deep neural networks and tree search”。
- 2017年， Deepmind在国际学术期刊Nature上发表的一篇研究论文“Mastering the game of Go without human knowledge”，推出了新版程序AlphaGo Zero。
- AlphaGo Zero抛弃了AlphaGo训练过程中的海量围棋棋谱经验数据，不需要监督学习来学习人类围棋棋谱，直接通过自我博弈（Self-play）、深度强化学习和蒙特卡罗树搜索等技术完成了围棋策略函数的训练。
- 围棋博弈属于完全信息博弈，而且是双人博弈，两个智能体之间交互进行决策。围棋博弈中的状态空间和动作空间都是有限的，只是空间大小超出了人类信息处理的极限。

通用智能备受期待

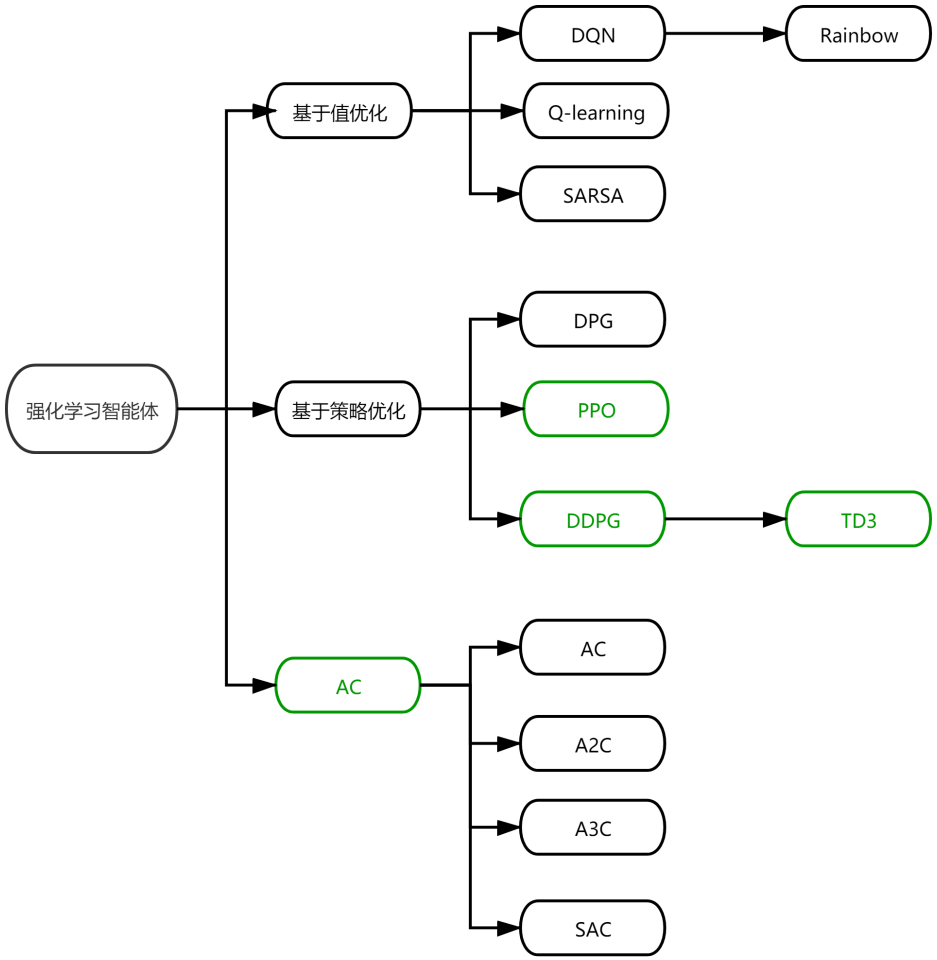
- 2018年，Deepmind 推出了 Alpha Zero 和 AlphaFold。
- 2020年，DeepMind的第二代AlphaFold 在国际蛋白质结构预测竞赛（CASP）获得了冠军，二代AlphaFold能够基于氨基酸序列精确地预测蛋白质的3D结构，其准确性可以与使用冷冻电子显微镜（CryoEM）、核磁共振或 X射线晶体学等实验技术解析的3D结构相媲美。
- 人工智能的目标是通用人工智能（Artificial General Intelligence，AGI）。通用人工智能是指具有一般人类的智慧，并且能够执行人类需要智力完成的任务，是一类通用的机器智能。通用人工智能又被称为强人工智能或者完全人工智能。

纲要

深度强化学习分类

- 基于值函数的强化学习（Value-based RL）和基于策略函数的强化学习（Policy-based RL）的算法
- 基于模型的强化学习（Model-based RL）和模型无关的强化学习（Model-free RL）算法
- 同策略（On-policy）学习算法和异策略（Off-policy）学习算法

基于值函数和基于策略函数的强化学习



基于模型和无模型的深度强化学习

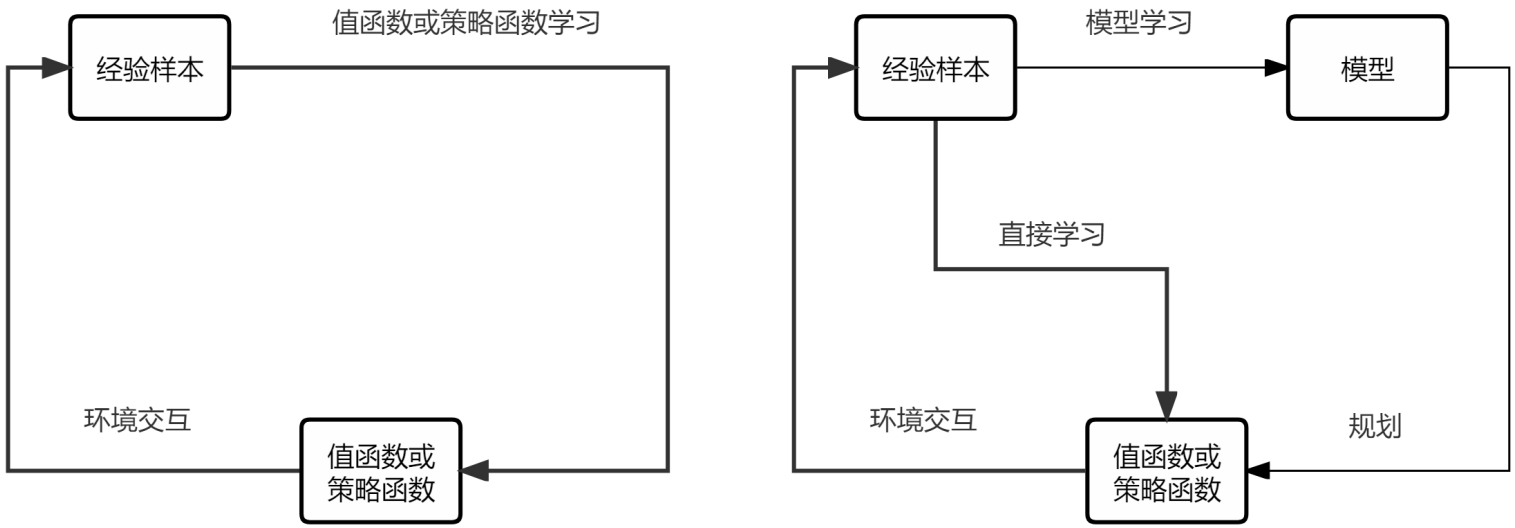


图 3: 无模型（左）和基于模型（右）的深度强化学习框架示意图

异策略和同策略学习

- 在强化学习智能体与环境交互过程中，我们可以将策略分成目标策略（Target policy）和行为策略（Behavior policy）。
- 目标策略是强化学习智能体需要学习和优化的策略。
- 行为策略是智能体与环境交互过程中获取经验轨迹样本数据的策略。
- 若智能体的目标策略与行为策略不同，则该类强化学习算法是异策略方法。
- 若智能体的目标策略与行为策略相同，则该类强化学习算法是同策略方法。

纲要

深度强化学习面临的挑战

深度强化学习算法在围棋、视频游戏、蛋白质折叠等实际场景中取得了耀眼的成绩，也遇到了不少挑战。

- 样本效率
- 灾难性遗忘
- 模拟与现实鸿沟
- 有效表示学习
- 可拓展性
- 奖励延迟
- 奖励稀疏
- 复杂动态环境
- 平衡探索与利用

平衡智能体的探索和利用

强化学习对探索和利用问题采用了不同的解决策略，但不可能彻底地解决探索和利用问题。在Q-learning算法中，智能体通过ϵ-贪心策略进行探索，具体表示如下：

$$P(a|s) = \begin{cases} 1-\epsilon+\frac{\epsilon}{|\mathcal{A}|}, & a = \arg\max_a Q(s,a;\theta) \\ \frac{\epsilon}{|\mathcal{A}|}, & a \neq \arg\max_a Q(s,a;\theta) \end{cases} . \quad (1)$$

平衡智能体的探索和利用

为了增加智能体探索性能，DQN改进算法采用了噪声网络。现实场景中的智能体面对相同的环境，理应做出相同的策略动作。噪声网络是在值函数网络参数θ中加入随机噪声：

$$\theta = \theta + \epsilon_\theta. \tag{2}$$

值函数网络参数θ加入噪声扰动后，增加了智能体行动的多样性。值函数网络参数θ在一次交互周期中保持不变，保证策略函数在一个交互周期中是相同的，即给定相同的状态时智能体能够输出相同的动作。

平衡智能体的探索和利用

为了增加智能体策略函数的探索性能，DDPG算法增加了智能体动作的随机性：

$$a = \pi_\theta(s) + \epsilon. \tag{3}$$

一般来说，动作空间中合法的动作区域可以表示为$a_{\text{Low}} \leq a \leq a_{\text{High}}$。TD3算法在给智能体动作值添加随机性过程中，限定添加随机性后的动作值在一个合法的区间内，因此对随机化后的动作行为进行了裁剪：

$$a = \text{clip}\left(\pi_\theta(s) + \epsilon, a_{\text{Low}}, a_{\text{High}}\right). \tag{4}$$

强化学习算法为了鼓励智能体更充分地探索环境，增加智能体策略的随机性是一个通用做法。

平衡智能体的探索和利用

为了鼓励智能体能够更充分探索环境，保证智能体策略不会过早收敛到单一化行为，SAC算法在最大化累积收益的同时最大化动作概率分布的熵$H(\pi(\cdot|s_t))$，使得行为概率分布更加分散：

$$\pi^* = \arg\max_{\pi} \mathrm{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t,s_{t+1})+\alpha H(\pi(\cdot|s_t))\right)\right], \quad (5)$$

其中α为算法超参数，调节动作概率分布熵在目标函数中所占比重，要求$\alpha > 0$，优化策略具有更大的熵，鼓励智能体行为多样性。

纲要

深度强化学习前沿

- 分层强化学习
- 分布式强化学习
- 多智能体强化学习
- 逆向强化学习
- 图强化学习

多智能体深度强化学习

多智能体部分可观马尔科夫决策过程可以表示为：$(\mathcal{S}, \mathcal{A}_1, \ldots, \mathcal{A}_N, P, R_1, \ldots, R_N, \Omega, O_1, \ldots, O_N, \gamma)$，其中$N$为智能体数量；

- $\mathcal{S}$ 为N个智能体的状态集合，每个智能体i都有各自的状态变量S_i。
- $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ 为N个智能体动作空间。
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ 为环境的状态转移模型，给定的动作条件下智能体从一个状态跳转至另一个状态。
- $\forall i, R_i : \mathcal{S} \times \mathcal{A}_i \times \mathcal{S} \to \mathcal{R}$ 是智能体i的即时奖励函数，其中 $\mathcal{R}$ 是一个连续函数且奖励值限定在范围$[0, R_{\max}]$中，且 $R_{\max} \in \mathbb{R}^+$。
- Ω 是观测变量集合。
- $\forall i, O_i : \mathcal{S} \times \Omega \to [0, 1]$ 是一组条件观测概率。
- $\gamma \in [0, 1)$ 为折扣系数。

逆向强化学习

逆向强化学习是一种模仿学习。在序贯决策问题中，智能体的模仿学习基于专家策略的示范行为数据学习一个最优化策略。专家示范行为数据并不包含即时奖励数据。

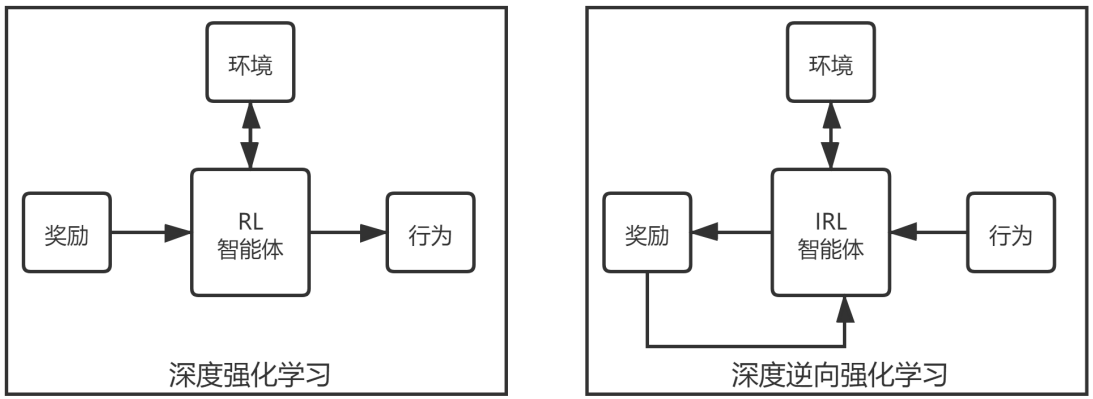


图 4: 深度强化学习和深度逆向强化学习比较示意图

模仿学习

模仿学习（Imitation Learning）包含大量算法，逆向强化学习作为一种常用方法，得到了广泛关注。

- 模仿学习又被称为学徒学习（Apprenticeship Learning）或基于演示的学习（Learning By Demonstration）。
- 行为克隆（Behavior Cloning）
- 逆向强化学习（Inverse Reinforcement Learning）
- 生成对抗模仿学习（Generative Adversarial Imitation Learning）等

行为克隆

如果存在一个专家数据集合，数据中包含了专家在不同环境状态下的决策行为数据，那么基于监督学习思想，我们可以直接将专家数据中的状态和动作对作为监督学习的数据集$\mathcal{D}$:

$$\mathcal{D} = \{(s_i, a_i)|i = 1, 2, 3, ..., N\}. \tag{6}$$

其中N为样本数量。
专家的示范数据中没有强化学习模型中常用的(s_i, a_i, r_i)经验轨迹数据。在获取专家数据的过程中，智能体以观察者身份来记录数据(s_i, a_i)，在状态s下专家实施了动作a，记录下来并存入数据集$\mathcal{D}$。经验数据状态-动作对(s_i, a_i)包含的信息可以作为智能体策略函数π的输入数据和输出信息。

行为克隆

智能体策略函数为状态s到动作a的映射，满足

$$a = \pi_\theta(s). \tag{7}$$

在监督学习框架下，我们可以构建最优化目标函数：

$$\hat{\theta} = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \left(a_i - \pi_\theta(s_i)\right)^2. \tag{8}$$

机器学习中的监督学习方法是各种学习算法的基础，有助于理解行为克隆方法。

行为克隆方法与逆向强化学习的区别

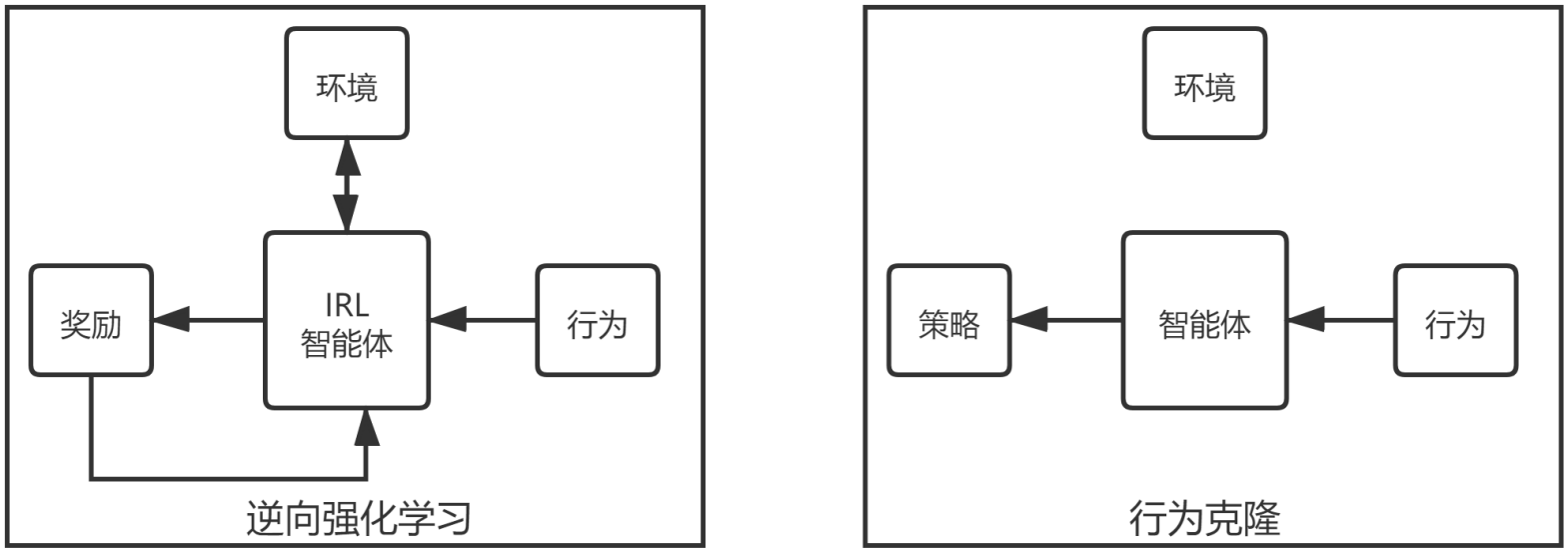


图 5: 行为克隆与逆向强化学习比较示意图

图强化学习

- 图强化学习是研究图相关或网络相关决策问题的重要方法和工具。
- 两大主要技术为深度强化学习算法和图神经网络模型。
- 图神经网络模型是专门针对图数据或网络数据的深度学习方法，能够有效地挖掘和学习网络节点、网络连边和全局网络的特征表示。
- 深度强化学习方法是图神经网络模型参数优化的重要方法。
- 图强化学习方法融合了图神经网络模型和深度强化学习模型。

纲要

深度强化学习建模框架

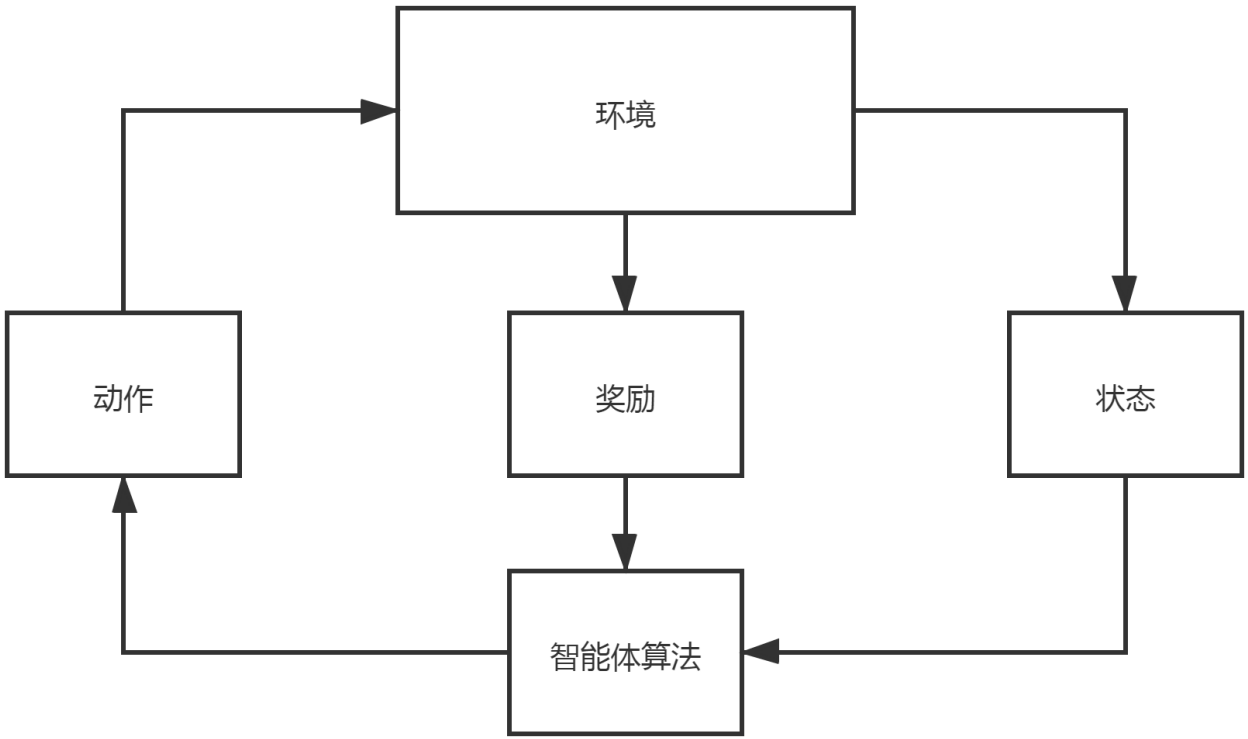


图 6: 深度强化学习模型框架

环境状态空间

- 量化环境状态是深度强化学习模型最先要考虑的问题。
- 在智能体决策过程中，智能体接收到环境状态，并基于策略函数做出最优决策行为动作。
- 深度强化学习过程中的值函数、策略函数、状态转移函数和奖励函数都包含了环境状态变量。
- 表示学习和深度学习中存在大量的针对不同数据类型的深度学习算法，如深度神经网络、卷积神经网络、循环神经网络、图神经网络等，分别适用于离散型、序列型、矩阵型、网络型数据等。
- 视频游戏中的游戏画面是智能体感知的环境状态，围棋游戏中棋盘落子情况是智能体感知的环境状态，金融市场中智能交易机器人可以将众多金融市场变量作为环境状态变量。

智能体动作空间

- 深度强化学习的目的是学习一个智能策略，智能体基于策略对环境状态作出智能反应，输出智能动作。
- 实践者根据环境和智能体特征，选定动作类型，如离散型、连续型或者混合型。
- 视频游戏中的动作就是人类玩家的键盘按键，可以转化成离散型变量。
- 股票交易智能体的动作可以是持仓比例，可用连续型变量表示。
- 现实场景中的问题非常明确，一般策略输出动作即为问题的解或解的组成部分。

即时奖励模型

奖励函数R是环境模型的一部分，环境模型未知时我们可以对其进行参数化：

$$r = R_w(s, a), \tag{9}$$

即环境模型基于奖励函数在给定的状态s和动作a下反馈一个奖励信号r。智能体的目标函数可定义为累积奖励，深度强化学习算法通过优化策略函数来最大化智能体的累积奖励。

状态转移模型

状态转移函数P也是环境模型的一部分，当环境模型未知时，我们可对其进行参数化：

$$s_{t+1} = P_w(s_t, a_t). \tag{10}$$

即环境模型基于给定的状态s_t和动作a_t返回下一个环境状态s_{t+1}。

环境模型

- 状态空间、动作空间、状态转移模型和奖励函数模型设定以后，我们融合四个模块构建环境模型。
- 深度强化学习的基本模型框架可以表示成马尔科夫决策过程五元组(S, A, R, P, γ)，其中S为模型状态空间，A为智能体动作空间，R为模型奖励函数，P为模型状态转移函数，γ为奖励折扣系数。
- 在与环境进行交互过程中，智能体基于策略函数在状态$s \in S$下做出动作a，环境接收智能体行动a后返回智能体即时奖励值r以及下一个环境状态s'。

策略函数模型（智能体模型）

深度强化学习智能体的训练目标是学习一个策略函数，因此我们需要表示智能策略函数，策略函数表示有很多种选择，如深度学习中的深度神经网络、卷积神经网络、循环神经网络、图神经网络等。智能体的策略函数定义为π，满足

$$a = \pi_\theta(s), \tag{11}$$

其中θ为策略函数参数，即深度神经网络模型参数。

深度强化学习算法

- 状态空间、动作空间、状态转移模型、奖励函数、环境模型和策略函数模型都是深度强化学习系统重要组成部分。
- 在完成深度强化学习智能系统的主体架构后，我们就需要选择关键模块——深度强化学习算法。
- 深度强化学习算法众多，是解决问题的关键。
- 在现实复杂决策问题中，可以从七个方面（状态空间、动作空间、状态转移模型、奖励函数、环境模型、策略函数模型和深度强化学习算法）入手构建一个较好的深度强化学习模型。

掌握的问题

1. 举例深度强化学习在智能投顾场景的应用。
2. 举例说明基于值函数和基于策略函数的深度强化学习方法的区别和联系？
3. 强化学习有哪些分类方式？试举出4种。
4. 简要叙述深度强化学习面临的虚实映射鸿沟。
5. 深度强化学习模型包括哪些核心模块？