Reinforcement Learning Paradigm

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

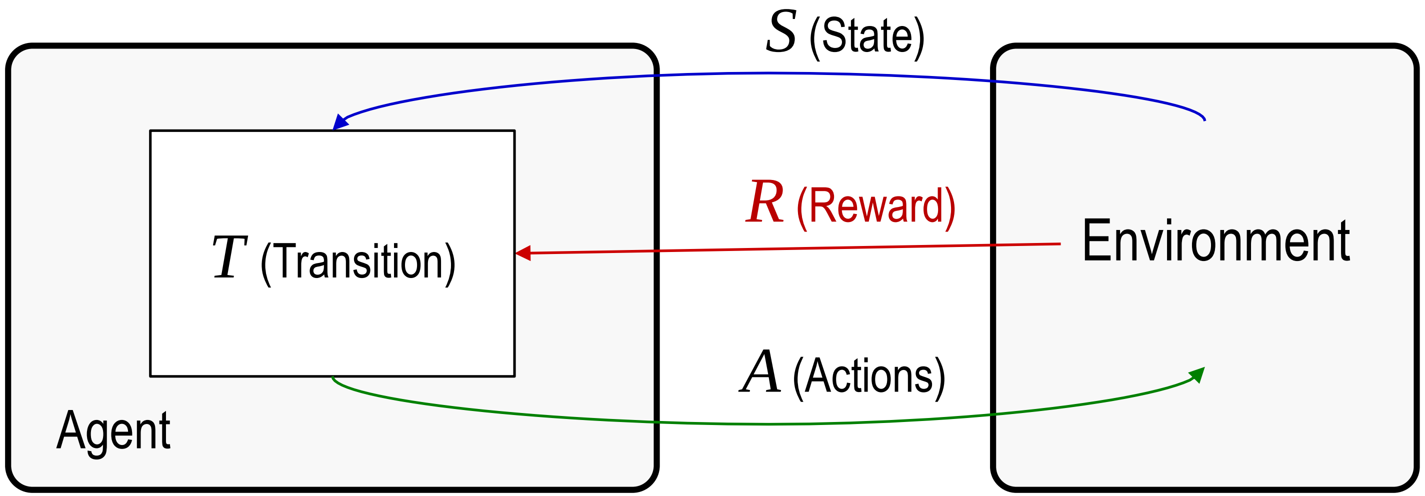
Contents:

11.3. Reinforcement Learning Paradigm

Contents:

What is Reinforcement Learning 什么是强化学习

- In reinforcement learning (RL), the learner is a decision-making agent, that takes actions in an environment and receives rewards for its actions.
  在强化学习中，其学习器是一个决策制定智能体，在环境下采取行动并获得这些动作的回报。
- After a set of trial-and-error runs, the agent should learn the best policy.
  经过一系列试错运行之后，该智能体能够学到最优策略。
- The policy is to maximize his reward over a course of actions and iterations with the environment.
  该策略是经过一个阶段的动作以及与环境的交互之后，使其回报最大化。



What is Reinforcement Learning 什么是强化学习

- Reinforcement Learning is inspired by behaviorist psychology.
  强化学习的灵感来自于行为心理学。
- Concerned with how agents take actions in an environment so as to maximize some notion of cumulative reward.
  关注于智能体如何在环境中采取行动，为了使累积回报最大化。
- Due to its generality, the problem is studied in many other disciplines, such as:
  由于其普遍性，许多其他学科都研究这一问题，例如：
  - game theory, control theory, operations research, information theory,
    博弈论、控制论、运筹学、信息论、
  - simulation-based optimization, multi-agent systems, swarm intelligence,
    仿真优化、多智能体系统、群体智能、
  - statistics and genetic algorithms.
    统计学和遗传算法。

Formalization of Reinforcement Learning 强化学习的形式化

- Reinforcement learning consists of: 强化学习包含
  - a set of agent states, 一组智能体的状态，$s_t \in S$;
  - a set actions of the agent, 一组智能体的动作，$a_t \in A$;
  - a transition from states to actions, 一个从状态到动作的转换函数，$T(s_t, a_t, s_{t+1})$;
  - a reward function, 一个回报函数，$R(s_t, a_t, s_{t+1})$.
- To look for a policy, 寻找一个策略，$\pi(s_t)$.
- Don't know T or R 尚未知道T或R
  - I.e. don't know which states are good or what the actions do.
    即，不知道哪个状态好或者要做什么动作。
  - Must actually try actions and states out to learn.
    必须实际去尝试要学习的行动和状态。

Supervised vs. Unsupervised vs. Reinforcement Learning 三种范式之比较

Supervised learning 有监督学习	➢ Input/output pairs are presented by labeled data (training examples). 通过标注数据（训练样本）提供输入和输出对儿。 ➢ *Learn-by-examples* 从样本中学习
Unsupervised learning 无监督学习	➢ To find the structure hidden in collections of unlabeled data. 发现无标注数据集中隐藏的结构。 ➢ *Learning-by-itself* 自我学习
Reinforcement learning 强化学习	➢ Input/output pairs are never presented, focus on online performance. 不提供输入和输出对儿，专注于在线的性能优化。 ➢ *Online-learning* 在线学习

Contents:

Types of Reinforcement Learning 强化学习的类型

- 1) Model-based 基于模型

  building a model of the environment. 构建环境的模型。

  - First acting in Markov decision process (MDP) and learning *T*, *R*;

    首先以马可夫决策过程方式动作，并学习*T*和*R*；
  - Then doing value iteration or policy iteration with learned *T*, *R*.

    然后用学习的*T*和*R*进行数值迭代或策略迭代。

- 2) Model-free 无模型

  learning a policy without any model. 学习策略而没有任何模型。

  - Bypassing the need to learn *T*, *R*, using direct evaluation policy.

    避开学习*T*和*R*的过程，采用直接评估策略。
  - Prediction-based temporal difference (TD) methods.

    基于预测的时间差分 (TD) 法。

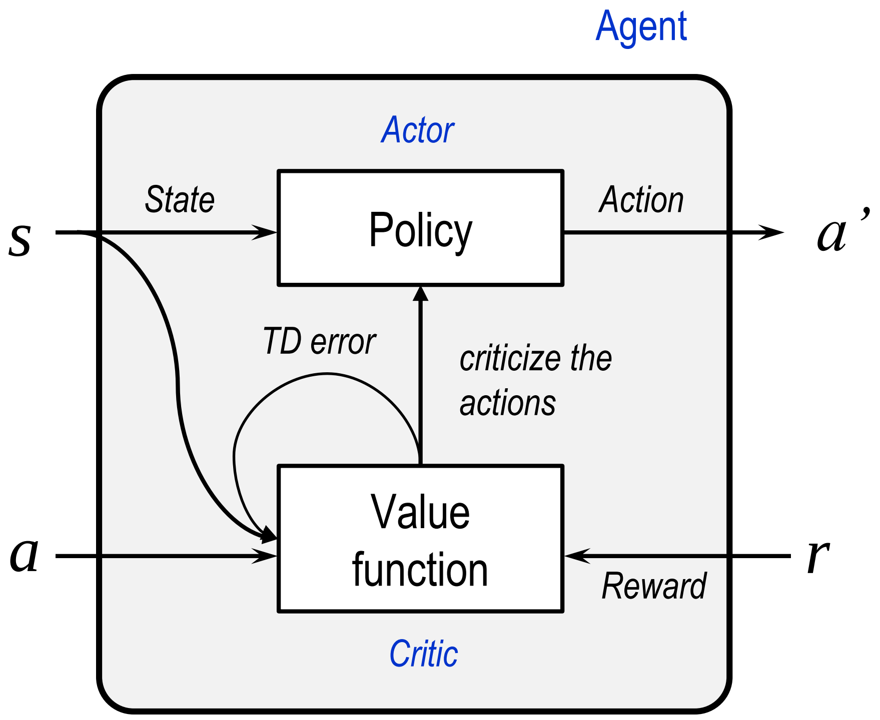
1) Model-based Reinforcement Learning 基于模型的强化学习

- Idea 思想
  - Learning the model empirically through experience. And solving for values as if the learned model were correct.
    通过实践经验学习模型。若学到的模型正确，则用于数值求解。
- Simple empirical model learning 简单的经验模型学习
  - Counting outcomes for each s, a.
    对每个s和a，对结果进行计数。
  - Normalizing to give estimate of $T(s_t, a_t, s_{t+1})$.
    对给定的估计$T(s_t, a_t, s_{t+1})$ 做正则化处理。
  - Discovering $R(s_t, a_t, s_{t+1})$ when we experience (s_t, a_t, s_{t+1}).
    当实践 (s_t, a_t, s_{t+1}) 时，去发现$R(s_t, a_t, s_{t+1})$。
- Solving Markov decision process with the learned model.
  用学到的模型求解马可夫决策过程。

2) Model-free Reinforcement Learning 无模型强化学习

Actor-Critic methods 动作者·评判者方法

- The TD version of Policy Iteration (On-policy).
  策略迭代 (On-policy) 的时间差分版。
- A structure to explicitly represent policy independent of value function.
  一种明确表示独立于价值函数的策略的结构。
- Policy (actor), is used to select actions.
  策略 (动作者) 用于选择动作。
- Value function (critic), used to evaluate actions made by actor.
  价值函数 (评判者) 用于评估动作者所完成的动作。



Actor-critic methods
动作者·评判者方法

TD error: $\delta_t = r_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

Preference: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$

2) Model-free Reinforcement Learning 无模型强化学习

□ Q-learning

- The TD version of Value Iteration (Off-policy).
  价值迭代 (Off-policy) 的时间差分版。
- Incrementally estimate Q-values for actions, based on rewards and Q-value function.
  基于回报值和Q-value函数，递增估计动作的Q值。
- Update rule is a variation of TD learning, using Q-values and a built-in max-operator over the Q-values of the next state:
  更新规则是一种时间差分学习的变体，采用Q值与内置的下个状态Q值的最大运算符：

$$Q(s_t, a_t) = \sum_a T(s_t, a_t, s_{t+1}) \left[R(s_t, a_t, s_{t+1}) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})\right]$$

- Sample-based, action-value function Q will be learned.
  学习到基于样本的、动作-值函数Q。

Contents:

New Algorithms of Reinforcement Learning 强化学习的新算法

- Deep Q-Network (DQN)
  深度Q-Network
  - CNN + Q-Learning (NIPS’13, Nature’15).
    将CNN与Q-Learning相结合。
- Deterministic Policy Gradients (DPG)
  确定性策略梯度
  - Estimate much more efficiently than usual stochastic policy gradient (ICML’14).
    与常用的随机策略梯度相比，可以更有效地进行估计。
- Asynchronous Advantage Actor-Critic (A3C)
  异步优势动作者·评判者
  - A variant of actor-critic method, using asynchronous gradient descent for optimization of DNN controllers. (arXiv:1602.01783)
    一种作者·评判者的变体，采用异步梯度下降来优化DNN控制器。

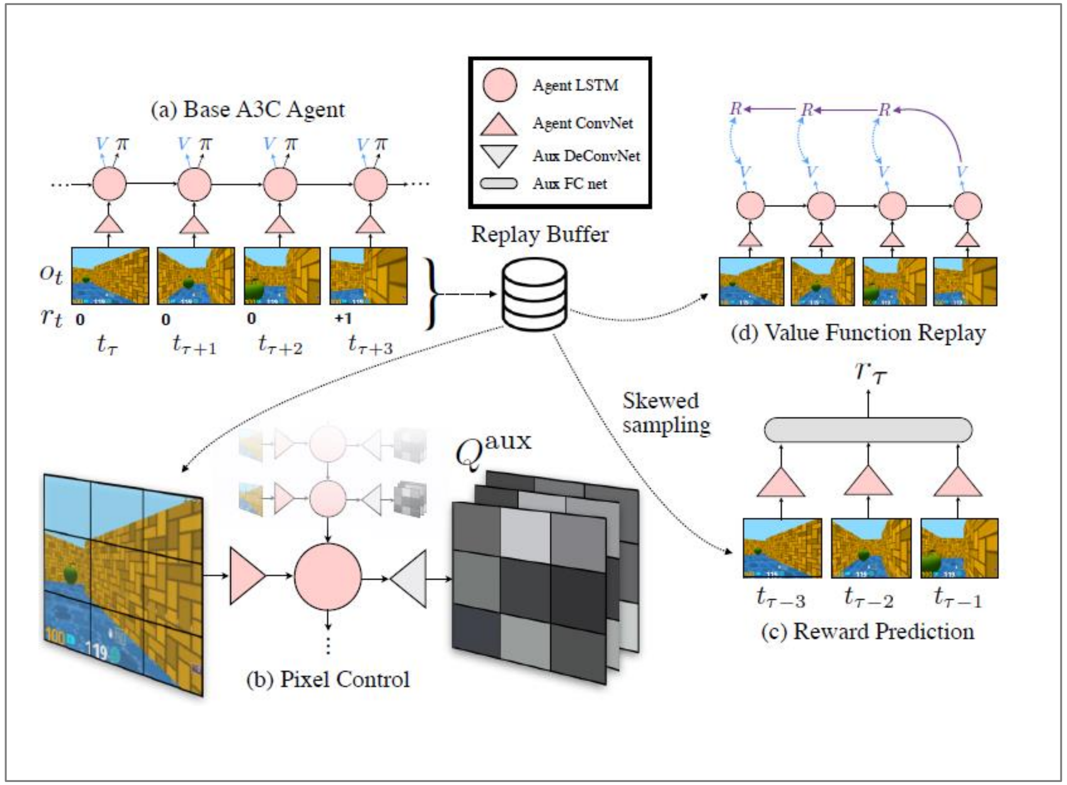
New Algorithms of Reinforcement Learning 强化学习的新算法

- UNsupervised REinforcement and Auxiliary Learning (UNREAL)
  无监督强化及辅助学习
  - For the environments containing a much wider variety of possible training signals (arXiv:1611.05397).
    针对包含更广泛的各种可能的训练信号环境。
  - It also maximize many other pseudo-reward functions simultaneously.
    还可以同时将许多其它的伪回报函数进行最大化。
- Neural Episodic Control (NEC)
  神经情景控制
  - Can rapidly assimilate new experiences and act upon them (arXiv:1703.01988).
    可以迅速地吸收新的经验，并且对其采取行动。

Case Study: UNREAL (Unsupervised Reinforcement and Auxiliary Learning)

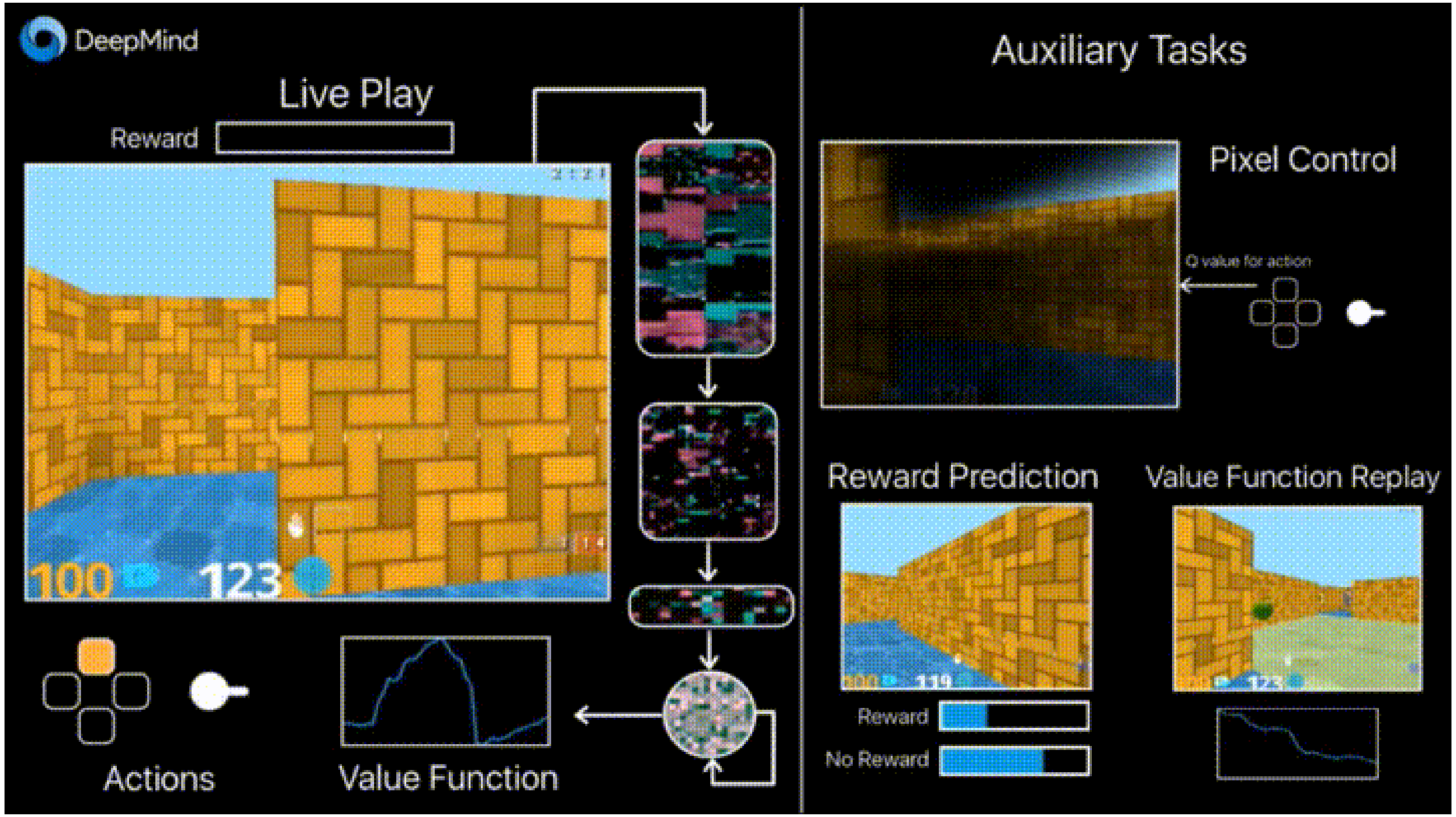
- (a) Base A3C Agent 基础A3C智能体
  a CNN-LSTM agent trained *on-policy* with A3C loss.
  一个CNN-LSTM智能体，经过A3C损失on-policy训练。
- (b) Pixel Control 像素控制
  training auxiliary policies Q^{aux} to maximise change in pixel intensity of different regions.
  训练辅助策略Q^{aux}，使不同区域像素强度变化达到最大化。
- (c) Reward Prediction 回报预测
  given three recent frames, predict the reward that will be obtained in next unobserved timestep.
  给定三个最近的帧，预测将在下一个未观测时阶获得的回报。
- (d) Value Function Replay 价值函数回放
  further training of value function using agent network to promote faster value iteration.
  进一步训练价值函数，采用智能体网络来推进迅速价值迭代。

Source: "Reinforcement learning with unsupervised auxiliary tasks", arXiv:1611.05397, DeepMind



Overview of the UNREAL agent.
UNREAL智能体概览

Case Study: UNREAL (Unsupervised Reinforcement and Auxiliary Learning)



3D Labyrinth on Atari, averaging 880% expert human performance.

Atari上的3D迷宫游戏，平均性能达到人类玩家的880%

AI

11.3. Reinforcement Learning Paradigm

Contents:

Typical Applications of Reinforcement Learning 强化学习的典型应用

- Robots 机器人
  - Robotic arms 机器人手臂
    be controlled to find the most efficient motor combination.
    控制得到最有效的电机组合。
  - Robot navigation 机器人导航
    collision avoidance behavior can be learned by negative feedback.
    可通过负反馈来学会碰撞躲避行为。
- Computer games 计算机游戏
  - Backgammom，西洋双陆棋
  - Chess，国际象棋
  - Go. 围棋

Thank you for your attention!

AI