深度强化学习中的迁移学习综述

田鸿龙

LAMDA, Nanjing University

November 29, 2020

Table of Contents

论文

Transfer Learning in Deep Reinforcement Learning: A Survey

Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou

南京大学 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Motivation

强化学习面临相当多的问题，例如

- partial observability
- sparsity and delay in the environment feedback
- high-dimensional observations and action spaces
- the cost of acquiring interaction samples can be prohibitive
- safety concerns in many real-world domains
- ...

因此在深度强化学习中利用之前的知识势在必行

TL in DRL

强化学习中的迁移学习更复杂，我们往往是将知识从一个 MDP 迁移到另一个 MDP 上，而这两个 MDP 可能具有很大差别。
或者是将知识从专家迁移到智能体上，这也就是模仿学习。

Definition 1. (Transfer Learning) *Given a set of source domain* $\boldsymbol{\mathcal{M}}_s$ *and a target domain* $\mathcal{M}_t$, Transfer Learning *aims to learn an optimal policy* π^* *for the target domain, by leveraging exterior information* $\mathcal{D}_s$ *from* $\boldsymbol{\mathcal{M}}_s$ *as well as interior information* $\mathcal{D}_t$ *from* $\mathcal{M}_t$*:*

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{s\sim\mu_0^t, a\sim\pi}[Q^{\pi}_{\mathcal{M}}(s,a)],$$
$$\text{where } \pi = \phi_{\text{DL}}(\mathcal{D}_s \sim \mathcal{M}_s, \mathcal{D}_t \sim \mathcal{M}_t).$$

Related Topics

Imitation Learning

- 一些模仿学习算法只考虑了原域 (D_s) 的知识，忽略了目标域 (D_t) 的知识，例如行为克隆本质上是监督学习
- 另一些模仿学习算法既考虑了原域的知识，也考虑了目标域的知识，例如逆强化学习（和生成对抗模仿学习）
- 与 Learning from Demonstrations (LfD) 的区别：LfD 发生在与 RL 环境的真实交互过程中，其目的是通过专家演示来实现有效的策略改进

Related Topics(cont.)

Lifelong Learning

- 要求 agent 有能力学习在一个信息流中，多个时间上或者空间上相关的任务
- 要求学习一个新的任务之后不能忘掉过去的任务
- 获得终身学习的关键是在长期获得新信息和保留以前学到的知识以跨新任务转移之间进行权衡
- 终身学习往往比单纯的迁移学习更难

Related Topics(cont.)

Hierarchical Reinforcement Learning (HRL)

- HRL 往往对任务、动作和状态空间进行更高层次的抽象，从而形成更具有结构化的策略
- 因为策略结构可以通过抽象解耦，HRL 促进了跨相似领域的知识转移

Related Topics(cont.)

Multi-Agent Reinforcement Learning

- inter-agent transfer: reuse knowledge received from com-munication with another agent, which has different sensors and (possibly) internal representations
- intra-agent transfer: reuse of knowledge generated by the agent in new tasks or domains
- see: A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems

Table of Contents

Approach Categorization

What knowledge has been transferred?

- 即来自原域的知识的**形式**和**质量**
- 形式：例如一组专家经验、专家策略的行动概率分布，甚至可以是估计源/目标 MDP 中状态和动作对质量的潜在函数
- 知识形式和粒度上的差异影响了不同 TL 方法的内在逻辑
- 质量：oracle 策略还是次优的人类演示

Approach Categorization(cont.)

Where the transfer occurs?

- 有些 TL 方法适用于 M_s 和 M_t 相等的情况，而另一些方法则设计用于在不同 MDP 之间传递知识
- M_s 和 M_t 之间的差异因任务而异
- 可能动作空间相同而状态空间不同，例如 Atari
- 可能状态空间相同而动作空间不同，例如更改了机器人部件的型号
- 可能状态空间和动作空间都同，状态转移不同，例如 Sim2Real
- 可能状态空间和动作空间都同，reward 不同，例如不同的 skills

Approach Categorization(cont.)

How to transfer knowledge between source and target MPDs?

- 对 Ms 和 Mt 的相似性做了什么假设
- 从 Ms 到 Mt 的映射函数是预定义的还是自主生成的
- 学习过程的哪些组成部分，例如，学习策略 、价值函数 V，甚至转换动力学 T (对于基于模型的 RL)，可以从转移的知识中获益
- 这个映射的 offline learning 还是 online learning

Approach Categorization(cont.)

What goal to achieve for the transfer learning approach?

- **优化目标函数**：在 D_t 上使用什么优化目标，例如为了让 D_s 的 policy 在新的 MDP 上探索，使用最大熵强化学习
- **评估指标**：initial/convergence/episodic performance

Approach Categorization(cont.)

How applicable a TL approach is?

- TL 算法是 policy-agnostic, 还是依赖于某个 set of algorithms
- 算法可以迁移哪些知识（见 What knowledge has been transferred?）
- 算法在哪些“不同”上迁移（见 Where the transfer occurs?）

Approach Categorization(cont.)

What is the accessibility of the target MDP?

- 往往认为从源域访问知识的成本更低
- 由于目标 MDP 中的高采样成本，agent 可能无法直接访问目标 MDP，或者只能有非常有限的 MDP 交互
- 例如 Sim2Real，需要在目标域考虑安全性，损耗等问题（例如那只伤痕累累的狗……）

Approach Categorization(cont.)

How sample efficient the TL approach is ?

- Zero-shot transfer：不需要在目标 MDP 交互
- Few-shot transfer：目标 MDP 交互很少
- Sample-efficient transfer：（多数算法处于这个级别）比在目标 MDP 直接训练效率更高
- 与目标 MDP 中从头开始的训练相比，TL 方法使目标 agent 具有更好的初始性能（且在转移知识的指导下更快地收敛）

Potential Differences Among Tasks

- S (State-space)
- A (Action-space)
- R (Reward function)
- T (Transition dynamics)
- μ_0 (Initial states)
- τ (Trajectories)

Evaluation metrics

- Jumpstart Performance (jp): 初始表现
- Asymptotic Performance (ap): 最终表现
- Accumulated Rewards (ar): 积累 reward，就是 reward 曲线下面积
- Time to Threshold (tt): 达到某个 Threshold 需要的时间
- Performance with Fixed Training Epochs(pe): 训练了固定轮数达到的表现
- Performance Sensitivity(ps): 算法受超参数的影响

Evaluation metrics(cont.)

- Transfer Ratio: with TL 和 Without TL 下 Asymptotic Performance 的比值
- Required Knowledge Quantity (rkqt): 为了达到一定的性能阈值，迁移学习所需的知识数量，例如源任务的数量、专家策略的数量或用于实现知识转移的演示交互的数量
- Required Knowledge Quality (rkql): 为了达到一定的性能阈值，迁移学习所需的知识质量，例如 TL 方法是否依赖于源域中的近似 oracle 知识，或者给定次优知识，TL 技术也能工作吗

Table of Contents

Table of Contents

Reward Shaping

Reward Shaping 利用先验知识重构目标 MDP 的奖赏分布，从而对 agent 的行为选择产生偏差

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{R})) \rightarrow \mathcal{M}' = \left(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{R}'\right)$$

Reward Shaping(cont.)

Potential based Reward Shaping (PBRS)

$$F\left(s, a, s'\right) = \gamma\Phi\left(s'\right) - \Phi(s)$$

Potential Based state-action Advice (PBA)

$$F\left(s, a, s', a'\right) = \gamma\Phi\left(s', a'\right) - \Phi(s, a)$$

Dynamic Potential Based (DPB)

$$F\left(s, t, s', t'\right) = \gamma\Phi\left(s', t'\right) - \Phi(s, t)$$

Table of Contents

LfD

传递的知识以外部演示的形式出现

- oracal 策略
- 一个接近最优的专家策略
- 次优专家策略

Table of Contents

Policy Transfer

传递的知识是来自源任务的专家（教师）策略

Problem Setting. *(Policy Transfer) A set of source tasks* $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$ *are provided along with their expert (teacher) policies:* $\pi_{E_1}, \pi_{E_2}, \ldots, \pi_{E_K}$*. A student policy* π *for a target domain is learned by transferring knowledge from each* π_{E_i}*, with* $1 \leq i \leq K$*.*

Policy Transfer(cont.)

- Transfer via Policy Distillation: 类似机器学习领域的知识蒸馏的概念，用若干个模型“教”一个新的模型
- Transfer via Policy Reuse: 直接重用源域的策略来构建目标域的策略

Table of Contents

Inter-Task Mapping

在目标域和源域之间学习一个映射函数，主要考虑两个问题

- 映射函数适用于哪个领域
- 映射函数是如何被利用的

基本假设

> **Assumption.** *(Existence of Mapping) A one-to-one mapping exist between the source and the target MDP.*

Table of Contents

Reusing Representations

- 不需要学习任务之间显式映射
- 表示要么可以直接重用，要么存在任务不变的特征空间
- 知识就可以在特征空间上的任务之间传递
- 例如 progressive network

Learning Disentangled Representations

state space, action space, or even reward distribution space can be disentangled into independent, orthogonal sub-domains

- successor representation (SR)
- universal value function approximating (UVFA)