摘 要 尽管强化学习近年来在视频游戏等诸多任务中大杀四方，但是从高维的观察空间中获得对状态空间鲁棒的表征仍然是整个强化学习任务重尚未很好解决的关键问题。而由于表征能力的不足，一方面强化学习算法的训练难度大大增加，数据利用效率低；另一方面，在某特定任务环境下训练得到的智能体很容易过拟合以至于难以泛化。数据增强就是在这样的背景下应用于辅助强化学习算法获得更好的表征能力。本文将简要概括强化学习中数据增强的背景以及现有方法，并结合本人正在进行的研究内容探讨数据增强在强化学习任务中面临的诸多挑战。

关键词 强化学习，数据增强，表征学习

Challenges for Data Augmentation in Reinforcement Learning

Guozheng MA

Abstract In recent years, Reinforcement Learning (RL) has curved out a brilliant path in the field of video game, particularly in tasks with high-dimension inputs. However, multiple research papers have demonstrated a restriction that the policies learned by RL agents in such tasks might not have robust representation. A mainstream approach to learn robust representations is based on data augmentation. This paper will briefly summarize the background and existing methods of data augmentation in reinforcement learning, and discuss the challenges of data augmentation in reinforcement learning tasks based on my ongoing research.

Key words Reinforcement Learning, Data Augmentation, Representation Learning

1 引言

1.1 强化学习的表征

从图像信息中获得鲁棒的表征信息是强化学习在现实场景下广泛落地的前提。现有强化学习方法中，一般分为向量输入（state-based）和图像输入（visual-based）两种输入方式，一般认为图像输入的表征难度远远大于向量输入场景。在有很有限的先验经验的情况下，从高维观察空间（图像）中准确编码出那些决策所需的状态信息表征对于现有强化学习算法是有很大挑战的。[1]

表征能力的不足给强化学习带来两个严重的问题。一方面，在训练过程中，仅通过与环境交互获得的稀缺奖励信号拟合具有具有高表征能力的编码器是极度困难的，这就导致强化学习训练过程中数据利用效率（data efficient）低下且容易出现次优收敛的问题。[2] 另一方面，由于神经网络参数量大的特性，其能够拟合大量信息，经过在某一特定环境中大量的训练，尽管智能体可能在该任务环境中获得还不错的表现，但是其极易过拟合到观察空间中那些与任务本身特性不相关的信息中，这就导致智能体的鲁棒性很差，难以泛化到其他任务场景，哪怕只是观察图像中某些视觉信息发生变化。

近年来，大量工作基于图像的数据增强优化智能体的表征能力。可以从两个角度理解数据增强。首先，增加的数据被视为训练模型的额外数据，这将大大扩充现有的稀疏数据，因此能够得到更有效的训练。尤其在泛化目标域与源域差异一致的情况下，如仅有视觉的颜色变化，能狗有效增加源域数据和目标域数据之间的相似性实现隐式的对齐效果。另一个理解数据增强的角度认为数据增强是一种正则化方法，通过对模型进行规则化，使不同的增强数据点具有相同的输出 (或相同的内部表示)。[3]

1.2 强化学习的建模

强化学习是一种区别于有监督学习和无监督学习的机器学习任务类型，其差异主要在于获取信息的方式。强化学习通过智能体（Agent）与环境（Environment）进行交互获得信息。在交互过程中没有监督信息（比如类型标签等），只有环境给出的奖励信息，而且环境奖励往往不能实时给出，大概率存在较大的延迟。因此，时间序列的建模是强化学习中一个非常重要的部分。

1.2.1 MDPs

马尔可夫决策过程 (Markov Decision Processes, 缩写 MDPs) 是强化学习中最基础也是最常用的数学模型。相对于一般的决策过程，其具有马尔可夫性。一个 MDPs 可以用一个四元组 $< S, A, T, R >$ 表示。其中：

S 表示状态空间 (State Space)，是所有可能存在的状态构成的集合；

A 为动作空间 (Action Space)，是所有可能动作构成的集合；

T 为状态转移函数 (State-Transition Function)，是由当前状态 S_t 和智能体此刻动作 A_t 到下一时刻状态 $S_(t+1)$ 的映射关系。虽然状态转移

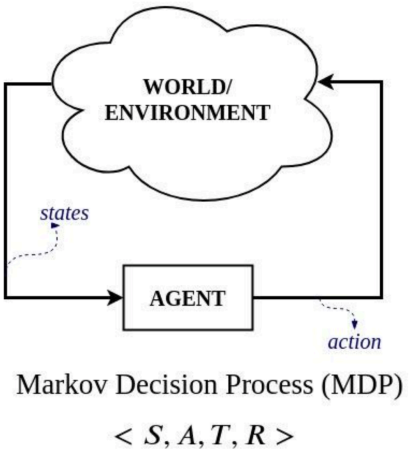

图 1 MDP

函数可以是确定的，但在绝大多数情况下，强化学习任务中的状态转移都是随机的，这时状态转移函数 P 为一个状态转移条件概率密度矩阵，如式 (1) 所示。且状态转移的随机性来自于环境特性，与智能体无关。

$$T^a_{ss'} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right] \tag{1}$$

R 为奖励函数 (Reward Function)，是指智能体在当前状态 S_t 的情况下执行一个动作 a_t 之后，环境反馈给智能体的一个数值。在 S_t 和 A_t 已知时，由于状态转移具有随机性，奖励 R_t 也不是唯一确定的，根据定义，奖励函数可以定义为式 (2)。但是，当下一个时刻的状态 $S_(t+1)$ 也已经观察到时，奖励 R_t 便是唯一确定的，因此可以设奖励 R_t 为 $S_t, A_t, S_(t+1)$ 三者的函数，

$$R^a_s = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right] \tag{2}$$

1.2.2 POMDPs

强化学习任务能够建模为 MDPs 的关键在于智能体的状态具有马尔可夫性，且智能体的状态能够被准确地获得。但是这样的要求在实际应用中是很难实现的。在现实场景中，我们无法直接获得全部状态信息，只能够根据设定任务关心的特征进行建模，从而获得某些特征量。即使在模拟环境中，智能体也仍然无法直接获得状态的全部特性，而只能通过观察到的输入信息学习状态的表征。

这种智能体只能观测到环境状态状态的部分信息的强化学习任务可以建模为部分可观的马尔可夫决策过程（Partial Observation Markov Decision Processes, 缩写 POMDPs)。相应的，MDPs 也可以称为完全可观的马尔可夫决策过程（Complete Observation Markov Decision Processes, 缩写 COMDPs)。

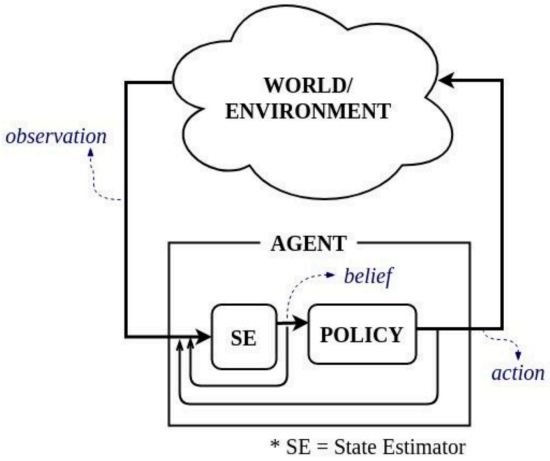

图 2 POMDP

POMDPs 可以定义为一个六元组 $< S, A, T, R, \Omega, O >$。相比于 MDPs，POMDPs 定义了观察空间 O 和观察函数 Ω。观察函数为智能体此刻执行动作 A_t 并转移到下一时刻状态 S_{t+1} 的条件下，观察为 O_{t+1} 的概率。

$$\Omega^a_{s'o} = \mathbb{P}\left[O_{t+1} = o \mid S_{t+1} = s', A_t = a\right] \tag{3}$$

目前最常用的虚拟环境 OpenAI gym[4] 就可以被定义为 POMDPs。因为这种虚拟的封装环境无法提供给智能体全部的实际状态空间信息，只能提供用来建模环境的参数构成一个向量作为观察空间（在输入为 state-based 时）。但是，在这种虚拟环境中以向量作为输入的强化学习过程，可以认为观察空间是能够完全反映状态信息（或者说能够完全反映不同状态下的差异的），因此可以认为是一种完全可观的马尔可夫决策过程。

但是对于本文主要考虑的以图像作为输入的强化学习任务，智能体能够观察到的信息完全来源于输入图像，但是输入图像无法包含状态空间的完整信息。例如，自动驾驶中的图像输入，仅依靠当前图像是无法得知图像中车辆和行人的速度等信息。因此，图像输入（Visual-based）的强化学习任务是一种典型的 POMDPs 任务，智能体能够获得的图像信息中仅包含当前状态的部分信息。在任务信息的维度上，观察空间的信息是状态空间的子集。这种信息的不完整性在强化学习的算法设计与优化时需要予以考虑。

1.2.3 Block MDP

POMDPs 在建模时，关注的问题为观察空间的信息并不能完全包含完成任务所需的状态空间信息。但是在图像输入为代表的强化学习任务还存在另一个严重的问题，观察图像中存在大量与希望表征的任务状态无关的信息。这种任务无关（task-irrelevant）观察信息极易使得深度学习模型过拟合到那些与任务无关的图像信息中，这就导致训练得到的智能体泛化能力很差，即使仅仅发生了背景颜色的变化。例如，在自动驾驶任务中，智能体在晴朗天气进行训练，但其过拟合到天气信息中，这样在天气情况发生变化时，算法性能定将严重下降。为了更好地完成这类底层任务逻辑（动力学模型）完全相同，但是观察的图像信息可能存在多样的与任务无关的图像变化的任务，强化学习任务可以建模为分块马尔可夫决策过程（Block Observation Markov Decision Processes, 缩写 BMDP）。[5]

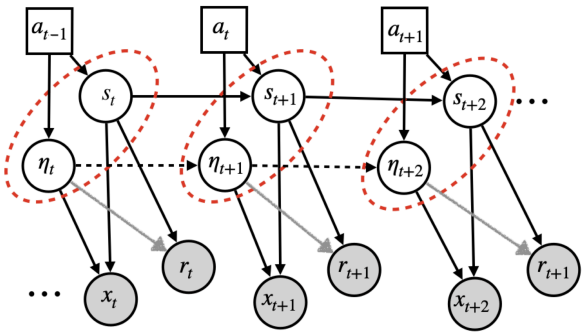

图 3 Block MDP

与 POMDPs 相同，Block MDP 同样定义了观察空间与从状态空间到观察空间的映射观察函数。但是与 POMDPs 不同的是，一个 Block MDP 定义为一类决策过程的集合，即 $\mathcal{M}_\varepsilon = \{(\mathcal{X}_e, \mathcal{A}, \mathcal{S}, \mathcal{P}, \mathcal{R}_e, \mathcal{O}_e, \gamma) | e \in \varepsilon\}$。一个 Block MDP 中的不同 MDP 共享相同的底层任务，即其状态空间 $\mathcal{S}$，动作空间 $\mathcal{A}$，状态转移函数 $\mathcal{P}$ 是相同的，与参数 e 无关。但是每一个 MDP 的观察空间 $\mathcal{X}$ 与从状态空间到观察空间的映射函数 $\mathcal{O}$ 是每个具体环境特异的。其中，状态空间 $\mathcal{S}$ 以及每个时刻的具体状态是智能体不可直接获得的，智能体智能根据直接输入的观察训练对于状态的表征。

Block MDP 的任务就是通过这样一系列满足底层任务相同，仅存在观察差异的 MDP 中的部分训练数据 $\varepsilon_{train} \subset \varepsilon$ 训练得到一个鲁棒的表征 $\phi : \mathcal{X} \to \mathbb{R}^d$，使得智能体在训练环境下学到的策略能够很好地泛化到属于同一个 Block MDP 的其他环境中。

2 强化学习中数据增强

本章将结合现有研究工作，尝试深入讨论强化学习中的数据增强三个关键的问题。

2.1 增强的最优不变性

状态变换的最优不变性（Optimality Invariant State Transformation）[6] 是数据增强需要满足的基本条件。在有监督问题中，不变性表现为增强前后该样本的真实标签应当不变。在强化学习中，DrQ[6] 定义数据增强对于状态的变换 τ 满足 $\tau : \mathcal{S} \times \mathcal{V} \mapsto \mathcal{S}$ 并且在映射后对应的 Q 值保持不变：

$$Q(s,a) = Q(\tau(s,\nu),a), \forall s \in \mathcal{S}, a \in \mathcal{A}, \nu \in \mathcal{V} \quad (4)$$

但是在实际应用中，我们是无法实现对状态 $\mathcal{S}$ 进行变换的，而只能对观察到的图像进行某种变换。借鉴 POMDPs 和 Block MDP，我们也定义观察空间 $\mathcal{O}$，则数据增强是在观察空间进行某种变换 τ 满足 $\tau : \mathcal{O} \times \mathcal{V} \mapsto \mathcal{O}_{aug}$，且这种变换能够保证其所表征的状态具有最优不变性。

2.2 增强的任务相关性

从这往后都是扯淡。

现有的图像数据增强可以认为是一种像素级的图像变换，即对图像中的所有区域不加区别地进行相同的变换。[7] 也就是说这种数据增强是与任务无关的，这就导致对于不同的任务，能够实现最好增强效果的增强方式是不同的。[8]

数据增强的应用过程可以概括为：对于状态 $\mathcal{S}$ 其观察图像为 $\mathcal{O}$，此刻模型的表征为 Rep。对观察图像为 $\mathcal{O}$ 进行增强得到 $\mathcal{O}_{aug}$，最优不变性要求增强前后的图像表示的状态 $\mathcal{S}$ 相同，模型得到对增强后图像的表征 Rep_{aug}。优化的目标函数可以不严谨地认为是 $\min ||Rep_{aug} - Rep||$ 并根据该目标函数对模型参数 θ 进行更新。假设模型具有鲁棒的表征能力，则在保证最优不变性时，增强前后模型的获得的表征相同，即 $Rep_{aug} = Rep$，但是在训练过程中可以认为模型的表征能力是不鲁棒的，即其不知道观察中那些信息是与任务相关的，哪些信息是与任务无关的。但是增强时如果改变了那些与任务相关的特征，模型的表征将发生变化，通过优化模型将学到这个特征是任务相关的。

3 总结与讨论

在强化学习任务中使用数据增强方法的目的是获得更加鲁棒的表征，从而提高样本利用效率（更好地训练）以及提高模型泛化能力（更好地迁移）。

但其实这两个目的应用的是数据增强不同方面的特性，提高样本利用效率希望通过数据增强满足增强后 Q 值不变的最优不变性，从而实现更小的估计方差。提升模型泛化能力则希望增强后的能够获得更加多样性的样本数据。简而言之，提升样本利用效率希望保证数据增强后的“稳定性”，而提升泛化能力则希望保证数据增强后的“多样性”，二者存在一定的矛盾。因此，在强化学习场景下用好数据增强，关键就在于如何平衡增强数据的“稳定性”和“多样性”。已有工作中，SVEA[9] 通过分别使用强增强和弱增强的数据估计 Q_{tgt} 和 Q_θ 的方法在保证 Q_{tgt} 稳定性的同时尽可能提升泛化能力；SECANT[10] 通过解耦策略优化和表征学习，在优化策略时，为了照顾敏感的强化学习训练而使用弱增强，在表征学习时使用模仿学习这种有监督学习的方式，因其抗噪性强，使用强增强；从而实现了根据提高样本利用率和提高泛化能力所需的不同特性分别使用其适合的增强方式。

参考文献

1 Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

2 Tao Yu, Cuiling Lan, Wenjun Zeng, Mingxiao Feng, Zhizheng Zhang, and Zhibo Chen. Playvirtual: Augmenting cycle-consistent virtual trajectories for reinforcement learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

3 Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 2021.

4 Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

5 Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

6 Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. *arXiv preprint arXiv:2004.13649*, 2020.

7 Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement Learning with Augmented Data. *arXiv preprint arXiv:2004.14990*, 2020.

8 Roberta Raileanu, Max Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862*, 2020.

9 Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation. *arXiv preprint arXiv:2107.00644*, 2021.

10 Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Anima Anandkumar. SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies. *arXiv e-prints*, pages arXiv–2106, 2021.