计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100107

一种数据高效的第三人称模仿学习方法

姜 冲[1] 章宗长[2] 陈子璇[1] 朱佳成[1] 蒋俊鹏[1]
1 苏州大学计算机科学与技术学院 江苏 苏州 215006
2 南京大学计算机软件新技术国家重点实验室 南京 210023
(20175227033@stu.suda.edu.cn)

摘 要 模仿学习提供了一种能够使智能体从专家示范中学习如何决策的框架。在学习过程中,智能体无需与专家进行交互,也不依赖于环境的奖励信号,而只需要大量的专家示范。经典的模仿学习方法需要使用第一人称的专家示范,该示范由一个状态序列以及对应的专家动作序列组成。但是,在现实生活中,专家示范通常以第三人称视频的形式存在。相比第一人称专家示范,第三人称示范的观察视角与智能体的存在差异,导致两者之间缺乏一一对应关系,因此第三人称示范无法被直接用于模仿学习中。针对此问题,文中提出了一种数据高效的第三人称模仿学习方法。首先,该方法在生成对抗模仿学习的基础上引入了图像差分方法,利用马尔可夫决策过程的马尔可夫性质以及其状态的时间连续性,去除环境背景、颜色等领域特征,以得到观察图像中与行为策略最相关的部分,并将其用于模仿学习;其次,该方法引入了一个变分判别器瓶颈,以对判别器进行限制,进一步削弱了领域特征对策略学习的影响。为了验证所提算法的性能,通过 MuJoCo 平台中的 3 个实验环境对其进行了测试,并与已有算法进行了比较。实验结果表明,与已有的模仿学习方法相比,该方法在第三人称模仿学习任务中具有更好的性能表现,并且不需要额外增加对样本的需求。

关键词:模仿学习;第三人称;领域特征;数据高效;图像差分;变分判别器瓶颈
中图法分类号 TP181

Data Efficient Third-person Imitation Learning Method

JIANG Chong[1], ZHANG Zong-zhang[2], CHEN Zi-xuan[1], ZHU Jia-cheng[1] and JIANG Jun-peng[1]
1 School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China
2 National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract Imitation learning provides a framework to make agent learn an efficient policy from expert demonstrations. During the learning process, the agent does not need to interact with the expert or get access to an explicit reward signal, but only needs a large number of expert demonstrations. Classical imitation learning methods usually need to imitate from first-person expert demonstrations, a sequence of states and actions that expert should have taken. However, most expert demonstrations exist in the form of third-person videos in reality. Different from the first-person expert demonstrations, there is a difference between the viewpoint of the third-person demonstrations and samples generated by the agent, resulting in a lack of one-to-one correspondence between them. Therefore, the third-person demonstrations cannot be directly used in imitation learning. To alleviate this problem, this paper presents a data efficient third-person imitation learning method. Firstly, this method introduces the image difference based on Generative Adversarial Imitation Learning(GAIL) to eliminate the domain features including the background of environment and colors by taking advantage of the Markov property of Markov decision process and the time continuity of states. And the most relevant part of policy can be achieved for imitation learning. Secondly, this paper introduces a variational discriminator bottleneck to limit the discriminator to alleviate the influence of domain features on the process of learning policy. In order to verify the performance of the proposed algorithm, this paper makes experiments on three MuJoCo tasks, and compares it with the existing algorithms. Experimental results indicate that the proposed method can achieve significant performance improvements

到稿日期:2019-11-14 返修日期:2020-04-16 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金面上项目(61876119);江苏省自然科学基金面上项目(BK20181432);中央高校基本科研业务费专项资金(14380005)
This work was supported by the National Natural Science Foundation of China (61876119), Natural Science Foundation of Jiangsu(BK20181432) and Fundamental Research Funds for the Central Universities(14380005).
通信作者:章宗长(zzzhang@nju.edu.cn)

over existing methods and does not require additional demonstrations, when dealing with imitation learning from third-person expert demonstrations.

Keywords　Imitation learning, Third-person, Domain feature, Data efficient, Image difference, Variational discriminator bottleneck

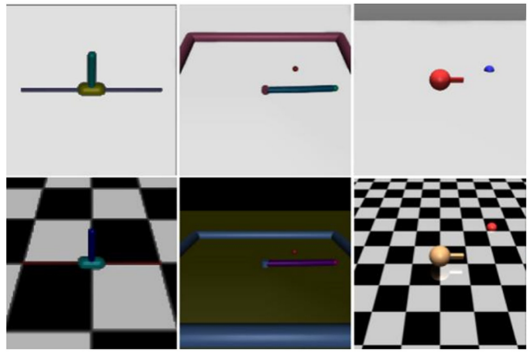
1　引言

强化学习(Reinforcement Learning，RL)作为机器学习领域的一个研究热点，目前已被广泛应用于机器人控制、仿真模拟、游戏博弈等领域[1-4]。强化学习的基本思想是通过不断地与环境进行交互，寻找能够最大化从环境中获得的累积奖赏的策略[5]。随着强化学习与深度学习的结合，强化学习方法已被应用到更加复杂的现实任务中。但是，强化学习需要根据奖赏函数来优化策略，当智能体处于奖赏函数未知的环境中时，强化学习方法就无法再进行策略优化了，这一限制使得强化学习算法难以适用于奖赏函数未知的现实任务。

模仿学习是一种能够从专家示范中模仿专家行为从而学习到专家策略的方法，它不依赖奖赏信号，也不需要与专家进行交互，因此模仿学习很好地缓解了强化学习依赖奖赏函数的问题[6-7]。模仿学习方法通常分为两类。1)行为克隆。这是一种监督学习方法，可以使智能体通过直接模仿专家轨迹的单步动作映射来学习专家策略，是一种较为简单的模仿学习方法。当专家样本能够覆盖整个状态空间时，行为克隆能够取得很好的性能表现。当专家样本较少时，智能体无法学习到每一个状态处的最优决策，并且由于不考虑长远影响，智能体在任一状态处产生的细微误差都将在序贯的决策过程中被逐步放大，从而导致级联误差。2)基于逆向强化学习的模仿学习方法。其主要思想是从专家轨迹中推导出一个奖赏函数，该奖赏函数可以使专家轨迹具有最优性，然后再基于该奖赏函数使用强化学习方法求解策略[8-9]。该方法求得的策略具有更好的泛化能力，并且对专家样本的需求量更小。

在一些复杂的现实问题中，奖赏函数获取难度较高，这一点严重地限制了强化学习方法的应用与发展。而模仿学习将强化学习对奖赏函数的依赖转换为对获取成本更低的专家示范的依赖，并且能够获得接近专家策略的性能表现。但是，传统的模仿学习方法需要第一人称的专家轨迹作为专家示范，这些轨迹由一系列的状态序列以及专家在各个状态上所采取的动作组成，获取难度较高。相比第一人称的专家示范，在现实生活中，第三人称专家示范更为常见且获取成本更低。

第三人称专家示范仅包含原始的第三人称观察图像，没有相应的动作信息。同时，不同视角所观察到的色彩、背景以及角度等均会不同(见图 1)，这就导致专家样本与智能体产生的样本之间存在明显的领域特征差异，受这种差异的干扰，智能体很难再从专家示范中学得专家策略。为了消除领域特征差异对策略学习的干扰，Bradly 等将领域混淆的思想与生成对抗模仿学习[10](Generative Adversarial Imitation Learning，GAIL)相结合，提出了一种基于第三人称示范的模仿学习方法[11](Third-Person Imitation Learning，TPIL)。TPIL 在 GAIL 的基础上构建了一个领域不可知(domain-agnostic)的特征提取器，能够模糊化领域特征，从观察中提取出领域不可知的特征表示，从而削弱由视角不同所带来的影响，使得 GAIL 可以使用第三人称专家示范进行学习。但是，为了获得领域不可知的特征提取器，TPIL 需要额外引入第二类专家示范，即在专家领域中使用非专家策略产生的示范，这大大增加了专家示范的收集难度。同时，额外引入的第二类专家示范的行为特征与专家和智能体的均不同，这也会影响判别器对智能体产生的样本的判别，从而影响策略的学习。



注：第一行是专家示范的第三人称视角，第二行是智能体自身的观察视角

图 1　实验环境 Inverted Pendulum，Reacher，Point 的示意图

Fig. 1　Diagrams of experiment environments of Inverted Pendulum, Reacher and Point

与 TPIL 直接模糊化领域特征的做法不同，Pratyusha 等[12]认为第三人称专家示范难以用于模仿学习的原因在于：智能体无法理解第三人称示范中专家的意图。因此，他们构造了一种分层模仿学习方法，可以将专家意图与控制策略分离开来。其中，高层模块为目标生成器，用于理解专家意图，将专家第三人称视角下的目标转换为智能体第一人称视角下的目标；低层模块为控制器，可以根据转换后的目标，预测智能体在当前状态下应该采取的动作。通过这种方式，智能体可以将第三人称视角中的专家意图转换到自身的第一人称视角中，避免了领域特征差异对策略学习的干扰。然而，目标生成器的训练不仅仅需要使用第三人称视角下的专家示范样本，还需要智能体在第一人称视角下执行相同的动作产生的样本，这无疑会大大增加样本的获取难度。同时，一个目标生成器只能用于一个特定的任务，对于任一新任务来说，都需要重新训练一个目标生成器。

针对上述问题，本文在文献[13]的基础上进行了进一步拓展，提出了一种数据高效的第三人称模仿学习方法 TPIL-ID-VDB(Third-Person Imitation Learning via Image Difference and Variational Discriminator Bottleneck)。该方法可以使用仅包含观察图像的专家样本进行模仿学习，同时可以有效地消除样本中的领域特征，保留样本中与策略最相关的行为特征，并且不需要额外引入第二类专家示范。具体地，本文基于马尔可夫假设，对相邻的连续两个观察图像做差分运算，

以此来直接去除观察图像中与行为特征无关的领域特征(如环境背景等),与此同时,该方法还可以得到观察图像中的变化区域,即样本的行为特征,提高数据的利用效率。其次,本文提出了一种变分判别器瓶颈(Variational Discriminator Bottleneck,VDB)的变种算法来进一步削弱领域差异信息对策略学习的影响。最终,我们在多个 MuJoCo[14] 任务上进行了实验。实验结果表明,在不需要额外的第二类专家示范的前提下,TPIL-ID-VDB 能够使智能体利用第三人称专家示范来学习专家策略,其性能表现也明显优于已有的方法,如 GAIL,TPIL。

2 相关工作

2.1 预备知识

马尔可夫决策过程(Markov Decision Process,MDP)可以定义为这样一个四元组:(S,A,P,R)。其中,S 表示一个状态集合;A 表示动作集合;$P:S\times A\times S\to[0,1]$ 表示状态迁移概率;$P(s'|s,a)$ 为智能体在状态 s 处采取动作 a 后转移到状态 s' 的概率;$R(s,a)$ 为奖赏函数,表示智能体在状态 s 执行动作 a 后获得的奖赏值。在强化学习中,智能体的目标是通过不断地试错来寻找一个策略 $\pi:S\times A\to[0,1]$,以使它能够获得的长期累积奖赏值达到最大。通常我们可以使用策略的折扣累积奖赏(亦称回报,return)$J(\pi)=\mathbb{E}_{\pi}[\sum_{t=0}^{T}\gamma^t R(s_t,a_t)]$ 来评价该策略的好坏,其中 s_0 为初始状态,γ 为折扣率,$a_t\sim\pi(a_t|s_t)$,T 为有限的时间步长,最优策略为 $\pi^*=\arg\max_{\pi}J(\pi)$。

2.2 模仿学习

模仿学习的目的是从专家示范中求解一个能够复现专家示范的策略,从而摆脱对环境奖赏的依赖。我们将 N 条专家示范轨迹的集合定义为 $X_E=\{\tau_1,\tau_2,\cdots,\tau_N\}$,轨迹 $\tau_i=\{\phi_1,\phi_2,\cdots,\phi_T\}$ 是一个特征序列,$i=1,2,3,\cdots,N$。其中,ϕ_j 可以是任意一种度量形式,如状态动作对(s,a)或原始观察图像 o,$j=1,2,3,\cdots,T$。模仿学习的学习目标可以表示为:

$$\pi^*=\arg\min_{\pi}M(p(\phi),q(\phi)) \tag{1}$$

其中,$p(\phi)$表示专家轨迹特征,$q(\phi)$表示模仿者生成的轨迹特征,$M(p,q)$表示两者之间的相似性度量[7]。

GAIL[10-15] 是一种结合了生成对抗网络(Generative Adversarial Nets,GANs)[16] 与逆向强化学习[9] 的模仿学习方法,它将模仿学习中的学徒学习问题看作一个占有率度量匹配问题的对偶,并将专家轨迹看作一个关于状态动作对(s,a)的集合。其中,GANs 的主要思想是构建两个神经网络——生成器 G 与判别器 D。生成器 G 的目标是生成与真实数据相近的样本,使得判别器 D 无法准确地区分两者;判别器 D 的目标则是尽可能准确地将两者区分开来。G 与 D 在这种对抗式的训练过程中不断地改进,直至达到纳什均衡,此时 G 能够生成以假乱真的样本,而 D 则无法再对样本做出准确的判断。GAIL 将 GANs 的对抗式训练过程应用到了模仿学习中,用策略 π_θ 作为生成器,然后通过与判别器 D_ω 博弈来优化策略,进而达到从专家示范中学习专家策略的目的。其中,θ 为策略网络参数,ω 为判别器网络参数。在每轮迭代中,D_ω 对输入的专家样本或生成样本进行判别,然后根据判别误差来更新自身的网络参数 ω,而策略 π_θ 则将 D_ω 的判别结果作为奖赏来更新自身策略。该过程可以形式化地表示为:

$$\min_{\pi_\theta}\max_{D_\omega}\mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D_\omega(s,a))] \tag{2}$$

其中,π_E 表示专家策略。

相比基于逆向强化学习的模仿学习方法,GAIL 将策略直接作为学习的目标,避免了基于逆向强化学习的模仿学习中的 RL 子过程,大大缩减了计算成本,具有更高效的计算能力。

2.3 第三人称模仿学习

与 GAIL 不同,在 TPIL[11] 中,专家示范以第三人称观察的形式存在,即 $\tau_i=\{o_1,o_2,\cdots,o_T\}$,且专家示范与智能体生成的样本之间存在明显的领域特征差异。目前,也有一些模仿学习方法可以直接从观察中学习专家策略[17-18],但是这些方法使用的专家示范依然是由来自第一人称视角的观察组成的,专家样本与生成样本之间不存在领域特征差异。对于 GAIL 来说,样本的领域特征不会随策略的改变而改变,智能体无法通过调整策略来消除与专家样本之间的领域特征差异,而判别器则可以快速地发现这种差异,并以此来区分样本。此时,智能体无法再从判别器的判别结果中获取有用的信息来更新策略。为了消除领域特征对策略学习的影响,TPIL 在 GAIL 中引入了领域混淆的概念,即模糊化样本中的领域特征,使得判别器无法再根据领域特征来区分专家样本与生成样本。具体地,TPIL 增添了一个领域不可知的特征提取器 D_F,经过 D_F 得到的样本特征会被分别输入到一个判别器 D_D 和一个领域分类器 D_C 中。D_D 的目的与 GAIL 中的判别器相同,即根据样本的行为特征将其准确地划分为专家样本与生成样本,而 D_C 的目的则是根据样本的领域特征将第一人称样本与第三人称样本区分开。为了能够使 D_F 模糊化领域特征,达到领域混淆的目的,Bradly 等在 D_F 与 D_C 之间增加了一个梯度翻转层(Gradient Reversal Layer,GRL),GRL 会在反向传播更新网络参数时将 D_C 传回的梯度进行翻转,从而使 D_F 朝着领域不可知的方向进行更新[19-20]。然而,上述过程需要领域分类器 D_C 有足够多的信息来区分领域特征与行为特征,而单纯的生成样本与专家样本并不足以提供这些信息。为此,TPIL 额外引入了一类专家样本,即在专家领域中执行非专家策略生成的样本。引入额外的专家样本可以使 D_C 更加关注领域特征,并尽可能地根据领域特征来对样本进行分类,从而为特征提取器 D_F 提供准确的反向梯度信息,使 D_F 模糊化领域特征,达到领域不可知的效果。

2.4 变分判别器瓶颈

在 GAIL 中,策略 π_θ 与判别器 D_ω 的训练过程是一个博弈过程,二者可以在不断博弈的过程中改进自身,直至最终达到纳什均衡。但是,如果在博弈过程中判别器 D_ω 太强,策略 π_θ 就很难再从博弈中获取有用的信息来进行优化。变分判别器瓶颈方法(Variational Discriminator Bottleneck,VDB)可

以通过限制 GAIL 中判别器网络的信息流来调制判别器的判别准确率，从而缓解在生成器与判别器对抗训练的过程中判别器过强的问题[21]。其在判别器网络中额外引入了一个编码器 Enc，可以将判别器的输入样本 x 映射到一个随机编码 z，即 $z\sim Enc(z|x)$，然后通过施加一个上界 I_C 来限制 x 与 z 之间的互信息 $I(x,z)$，其中 I_C 为常量。通过这种方式，VDB 可以对判别器网络的输入信息流进行限制[22]，从而达到削弱判别器的目的，以更好地平衡判别器网络与策略网络。

3　数据高效的第三人称模仿学习

在现实生活中，从第三人称视角收集专家示范更加容易。但是，由于此类专家示范与生成样本之间存在领域特征差异，并且缺乏一一对应的关系，因此我们无法直接将第三人称示范应用到模仿学习中。本文提出了一种可以使用第三人称示范的模仿学习方法，结构框架如图 2 所示，其中 X_{π_θ} 表示使用策略 π_θ 采样得到的样本的集合。该方法具有如下特征：1）使用第三人称的原始图像作为示范；2）与 TPIL 相比，无需引入额外的第二类专家示范；3）样本的利用效率高。

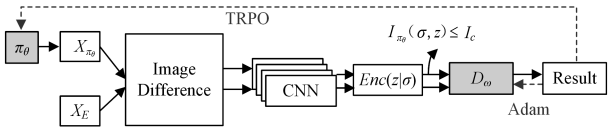


图 2　TPIL-ID-VDB 框架图

Fig. 2　Framework of TPIL-ID-VDB

首先，假设有专家轨迹 $\tau_i=\{o_1,o_2,o_3,\cdots\}$，其中 o_t 是原始第三人称观察图像。与 GAIL 相似，我们考虑将学徒学习问题看作一个占有率度量匹配问题的对偶，并将轨迹看作一个关于观察对(o,o')的集合，其中 o'为智能体或专家根据自身当前的状态观察 o 采取某一动作后转移到的下一个状态观察。本文算法的目标是从给定的观察对中学习得到 $\pi_\theta(a|o)$。

3.1　改进 1：结合图像差分

在 GAIL 中，判别器需要从输入的专家样本和生成样本中提取出各自的特征，并以此来判别样本的真假，最后根据判别结果更新网络参数。然而，如果生成样本与专家样本之间存在着明显的领域特征差异，且这种固有的领域特征差异无法通过改变 π_θ 来消除，则判别器就能够快速地学习到准确判别样本真假的方法，并使智能体无法再从判别器的反馈中获得有利于更新策略的信息。基于上述问题，本文首次将图像差分引入到 TPIL 中，提出了 TPIL-ID（Third-Person Imitation Learning via Image Difference），即使用图像差分方法处理输入到判别器中的样本，以削弱领域特征对判别器以及策略的影响，即图 2 中的图像差分（Image Difference）模块。同时，在 MDP 问题中，由于状态观察具有马尔可夫性质，任一时刻 $t+1$ 的观察 o_{t+1} 只取决于上一时刻 t 的观察 o_t 以及动作 a_t，并且状态观察是连续的，因此，在专家样本中仅包含观察信息而没有动作信息时，对相邻的两个观察进行图像差分可以直接地提取出样本中与行为特征相关的变化区域，使得判别器更加关注样本的行为特征而非领域特征（见图 3）。在 $t+1$ 时刻，样本的差分结果可表示为：

$$\sigma_{t+1}=F(o_{t+1}-o_t) \tag{3}$$

其中，F 是一个由卷积神经网络（Convolutional Neural Network，CNN）构成的特征提取器，用于将图像差分的结果编码成一系列低维、抽象的特征表示。图 2 中，CNN 模块表示特征提取器 F。相应地，我们可以将 GAIL 的目标函数变更为：

$$\min_{\pi_\theta}\max_{D_\omega}\mathbb{E}_{\pi_\theta}[\log D_\omega(\sigma)]+\mathbb{E}_{\pi_E}[\log(1-D_\omega(\sigma))] \tag{4}$$

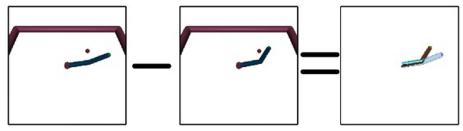


图 3　图像差分示意图（以 Reacher 为例）

Fig. 3　Diagram of image difference (take Reacher as an example)

TPIL-ID 使用图像差分方法改变样本形式，并将其作为判别器网络的输入状态，使得判别器难以再根据样本的领域特征来判别样本真假，智能体也可以从判别器的反馈中获取有用的信息来优化策略。与 TPIL 不同，TPIL-ID 不再需要额外的第二类专家样本，减小了样本的获取成本。另外，为了防止因相邻两个状态观察之间差异过小而难以得到理想的行为特征，我们使用间隔了 n 步的观察来进行图像差分，即 $o_{t+n}-o_t$。本文实验中，n 取 3。

3.2　改进 2：结合变分判别器瓶颈

通常情况下，上述方法能够消除由环境背景、颜色等不同所带来的领域差异信息。但是，由于观察视角不同，专家样本与生成样本中的部分物体的倾斜角度可能会存在一些差异，这种残留的领域特征差异仍然会在一定程度上导致判别器过强，从而影响策略的学习。针对这一问题，本文在 TPIL-ID 中进一步引入了 VDB，用于削弱判别器，以更好地平衡领域特征带来的影响。由此，TPIL-ID-VDB 的目标函数可以由式（4）变更为：

$$\min_{\pi_\theta}\max_{D_\omega}\mathbb{E}_{\pi_\theta}[\log D(z)]+\mathbb{E}_{\pi_E}[\log(1-D(z))] \tag{5}$$

$$\text{s.t. } I(\sigma,z)\leqslant I_C$$

其中，$z\sim Enc(z|\sigma)$，$I(\sigma,z)$表示 σ 与 z 之间的互信息，并且：

$$I(\sigma,z)=\int_{\sigma,z}p(\sigma)Enc(z|\sigma)\log\frac{Enc(z|\sigma)}{p(z)}\mathrm{d}\sigma\mathrm{d}z \tag{5}$$

其中，$p(\sigma)$与 $p(z)$表示关于 σ 与 z 的分布。然而，直接计算 $p(z)=\int_x Enc(z|\sigma)p(\sigma)$较为困难，因此 VDB 使用了一个标准高斯分布 $r(z)$作为 $p(z)$的近似值，以此来获得一个变分下界。因为 $\int p(z)\log p(z)\mathrm{d}z\geqslant\int p(z)\log r(z)\mathrm{d}z$，所以可以得到：

$$\begin{aligned}I(\sigma,z)&\leqslant\int_{\sigma,z}p(\sigma)Enc(z|\sigma)\log\frac{Enc(z|\sigma)}{r(z)}\mathrm{d}\sigma\mathrm{d}z\\&=\mathbb{E}_{\tilde{\pi}}[KL[Enc(z|\sigma)||r(z)]]\\&\leqslant I_C\end{aligned}$$

其中，KL 散度（KL divergence）可以用于度量两个概率分布之间的差异。通过这种方式，VDB 对编码器 Enc 的输出进行了约束，以达到限制输入到判别器的信息流的目的。在 VDB

中，$\tilde{\pi}=\frac{1}{2}\pi_\theta+\frac{1}{2}\pi_E$，表示专家策略与智能体策略的混合，即对输入的专家样本与生成样本均会进行限制。但是，本文中 $\tilde{\pi}$ 仅代表 π_θ，即 TPIL-ID-VDB 仅对生成样本的信息流进行限制，从而约束判别器对生成样本的判别准确率。因为策略的学习仅与判别器对生成样本的判别结果有关，而判别器对专家样本的判别准确度则不会对策略产生直接的影响。同时，该方法可以进一步削减 VDB 的计算成本，提高计算效率。具体如图 2 所示，我们将特征提取器 F 的输出 σ 输入到编码器 Enc 中，从而得到编码 z 作为判别器的输入，同时对生成样本的编码结果施加一个约束条件 $I_{\pi_\theta}(\sigma,z)\leqslant I_C$，以限制判别器对生成样本的判别。

为了优化此目标函数，我们引入一个拉格朗日乘子 β：

$$\min_{\pi_\theta}\max_{D_\omega}\mathbb{E}_{\pi_\theta}[\log D(z)]+\mathbb{E}_{\pi_E}[\log(1-D(z))]+\beta(\mathbb{E}_{\pi_\theta}[KL[Enc(z|\sigma)\parallel r(z))]-I_C) \tag{8}$$

其中，β 用于调节对信息流的限制程度，β 的更新方法如下：

$$\max(0,\beta+\alpha_\beta(\mathbb{E}_{\pi_\theta}[KL[Enc(z|\sigma)\parallel r(z))]-I_C)) \tag{9}$$

其中，α_β 为 β 的更新步长参数。

3.3 算法实现

由图 2 可知，TPIL-ID-VDB 主要需要更新两个网络，即判别器网络 D_ω 和策略网络 π_θ。对于 D_ω，我们使用 Adam 算法[23]来迭代地更新其网络参数。同时，我们将特征提取模块以及编码器模块都作为判别器的一部分加入到判别器网络中一起进行更新操作。对于策略网络 π_θ 来说，我们使用 TRPO 方法[24]来进行优化，以确保策略每次更新的回报函数增加或保持不变。具体算法实现如算法 1 所示。

算法 1 TPIL-ID-VDB

输入：第三人称专家示范 X_E

输出：学习得到的策略 π_θ

1. 初始化：判别器网络参数 ω，策略网络参数 θ，特征提取器 F，编码器 Enc，最大时间步 T_{max}
2. REPEAT：
3. 使用策略 π_θ 采样得到轨迹集合 X_{π_θ}
4. 分别从 X_E 和 X_{π_θ} 中抽取相同数量的专家样本 x_E 与生成样本 x_{π_θ}
5. 图像差分：

$$\sigma_\theta=F_{x_{\pi_\theta}}(o_{t+3}-o_t)$$

$$\sigma_E=F_{x_E}(o_{t+3}-o_t)$$

6. 编码：$z\sim Enc(z|\sigma)$
7. 使用如下梯度更新判别器参数 ω：

$$\Delta_\omega=\mathbb{E}_{\pi_\theta}[\nabla_\omega\log D_\omega(z)]+\mathbb{E}_{\pi_E}[\nabla_\omega\log(1-D_\omega(z))]+\beta(\mathbb{E}_{\pi_\theta}[KL[Enc(z|\sigma)\parallel r(z)]]-I_C)$$

8. 将 $\log D_{\omega,x_{\pi_\theta}}(z)$ 作为奖赏函数，使用 TRPO 方法更新策略参数 θ。
9. 时间步 $T=T+1$
10. UNTIL $T>T_{max}$

其中，下标 x_E 表示样本来自于专家示范，下标 x_{π_θ} 表示样本来自于策略 π_θ，以示区分。与 TPIL 不同，本文方法 TPIL-ID-VDB 不需要引入领域分类器，因此也就不需要额外的第二类专家示范来提供信息以区分领域特征与行为特征。同时，TPIL-ID-VDB 仅使用一张包含了样本行为特征的差分图像作为判别器网络的输入，大大减少了迭代训练过程中的计算量。

4 实验

本节首先介绍了实验所使用的训练环境和训练过程中所使用的参数设置，然后在 3 种 MuJoCo 环境中评估了 GAIL，TPIL，TPIL-ID-VDB 的训练效果。其中，GAIL 分别使用了第一人称专家示范与第三人称专家示范进行实验；TPIL 在有额外的第二类专家样本和没有额外的第二类专家样本的情况下分别进行了实验。

4.1 实验环境及参数设置

与 TPIL 相同，本文主要在 MuJoCo 物理引擎上的 3 个环境中进行了实验，分别是 Reacher，Inverted Pendulum 和 Point。我们首先使用强化学习方法在每个环境中训练得到专家策略，然后修改环境的观察角度以及环境背景等，再执行专家策略，提取出每一个时间步的观察图像，从而构成第三人称专家示范轨迹。同时，我们还使用一个随机策略在相同环境下采样得到了 TPIL 所需要的额外的第二类专家样本，图 1 给出了专家示范与智能体生成的样本在领域特征上的差异。在 Inverted Pendulum 中，我们使用 25 条专家轨迹作为专家示范；在 Reacher 中，我们使用 200 条轨迹作为专家示范；在 Point 中，我们使用 200 条轨迹作为专家示范。在训练过程中，这些专家轨迹会被打乱成单个的状态迁移对 (o,o')，其中 o 为 RGB 图像。

Reacher 是有两个自由度的手臂，其目的是控制手臂将其末端移动到目标点上，手臂末端距离目标点越近，奖赏就越高。为了更加明显地区分两种不同领域，我们对专家示范的环境背景、物体颜色、相机角度都进行了明显的修改。

Inverted Pendulum 是一个倒立摆，其主要目的是尽可能地维持平衡，使得竖着的杆子不会倒下，维持的时间越久累积回报就越高。我们在专家示范中对其环境背景以及物体的颜色都进行了较大的改动，与 TPIL 不同的是，我们还改变了相机的角度。

Point 是平面上的一个质点，其目标是移动到平面上的一个目标点上，质点离目标点越近，奖赏就越高。实验中，我们对专家示范的背景、目标点颜色、质点颜色等都进行了明显的修改。

在实验中，TPIL-ID-VDB 主要更新两个网络，即策略网络 π_θ 和判别器网络 D_ω。其中，π_θ 是由两个有 100 个神经元的隐藏层构成，激活函数为 tanh。判别器网络 D_ω 中包含有图像差分模块、特征提取模块、编码器模块以及判别器模块。图像差分模块包含了一个差分运算，构成了判别器网络的输入层；特征提取模块由两个卷积层和两个池化层构成，激活函数为 ReLU；编码器由一个全连接层构成，激活函数为 ReLU；判别器模块由两个全连接层和一个输出层构成，激活函数为 ReLU。

4.2 实验评估

在模仿学习中，评估实验性能的指标通常是策略的训练

过程以及策略最终的评估结果。其中，训练过程指为每一个训练阶段计算出平均情节回报，并以此做出从训练开始到结束的曲线图。图 4 给出了 GAIL，TPIL 以及 TPIL-ID-VDB 在 Reacher，Inverted Pendulum，Point 3 个实验环境上的训练结果，每个训练图的横坐标为迭代次数，纵坐标表示在每个训练阶段情节的平均回报。图 4 中，每一个实验使用了 5 个随机种子，并取最终多次实验结果的平均值。

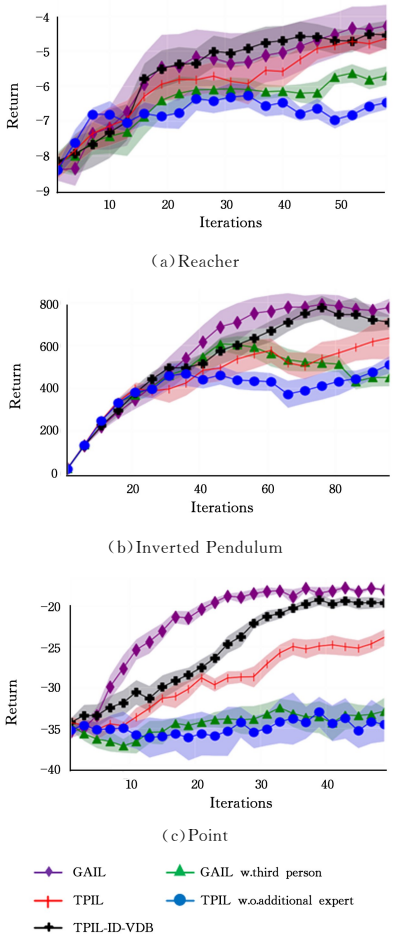


图 4　在 Reacher，Inverted Pendulum，Point 3 种实验环境上的实验结果对比

Fig. 4　Comparison of experimental results on Reacher，Inverted Pendulum and Point

图 4 中，菱形曲线表示 GAIL 使用第一人称专家示范的训练结果；三角形曲线为 GAIL 使用第三人称专家示范的训练结果；竖形曲线为 TPIL 的训练结果；圆形曲线为不使用额外专家示范的 TPIL 的训练结果；十字形曲线为文中提出的 TPIL-ID-VDB 的实验结果。从图中可以看出，TPIL-ID-VDB 可以使用第三人称专家示范进行模仿学习，并获得接近使用第一人称专家示范的 GAIL 的性能表现，同时也优于 TPIL。另外，从 TPIL 与不使用额外专家示范的 TPIL 的对比中可以看出，TPIL 的性能表现确实依赖于额外的第二类专家示范。与使用第三人称专家示范的 GAIL、不使用额外的专家示范的 TPIL 相比，TPIL-ID-VDB 的性能表现更加突出，这说明 TPIL-ID-VDB 可以更好地适用于第三人称模仿学习任务，并且对专家样本的需求少于 TPIL。

在 TPIL-ID-VDB 中，I_C 与 β 是两个较为重要的参数。其中，常量 I_C 为判别器瓶颈，用于限制判别器的信息流，以降低其对生成样本的判别准确度。在使用第三人称示范进行模仿学习时，由于存在领域特征差异，判别器可以轻易地获取大量信息对样本进行准确的判别，从而导致策略难以从判别结果中获取足够的梯度信息进行更新。而 I_C 取值越小，判别器的判别准确度越低，能够提供给策略的梯度信息就越多，策略可以更好地进行更新。但是，I_C 的取值也并非越小越好，因为 I_C 太小会导致策略更新过快，判别器与策略仍然无法保持平衡，二者都无法再从博弈中获得有效的信息来进行自身更新，最终只会学到一个较差的策略。系数 β 用于调节判别器瓶颈的限制程度，当 β 值接近于 0 时，算法的性能就会恢复到与 TPIL-ID 相同的水平，即判别器瓶颈对判别器的影响接近于 0；当 β 取较大的值时，判别器受判别器瓶颈的影响较大，会提供给策略更多的梯度信息，此时策略在前期的学习速度会加快，但是这也会严重破坏策略与判别器之间的平衡，导致最终收敛得到一个较差的策略。在本文的实验中，I_C 取 0.5，β 则不取固定的值，而是采用自适应的更新方法进行不断更新。

本文还对通过 GAIL，TPIL，TPIL-ID-VDB 训练得到的策略模型进行了评估，评估结果为 50 条采样轨迹的平均值，如表 1 所列。表 1 中的数据表明，在使用第三人称专家示范进行模仿学习时，TPIL-ID-VDB 的性能表现接近于使用第一人称专家示范的 GAIL，也略好于使用了额外的专家示范的 TPIL，这表明 TPIL-ID-VDB 能够使用第三人称专家示范进行模仿学习，且不会增加对专家样本的需求。而不使用额外的专家示范的 TPIL 在 Reacher 与 Point 中获得了最低的评估值，在 Inverted Pendulum 中的表现也低于使用了额外专家示范的 TPIL，这进一步说明了 TPIL 依赖于额外的第二类专家示范。

表 1　策略评估结果

Table 1　Results of policy evaluation

	Reacher	Inverted pendulum	Point
GAIL w. first-person	−4.2	798.2	−18.5
GAIL w. third-person	−5.8	472.6	−34.2
TPIL	−4.5	683.1	−23.1
TPIL w. o. additional expert	−6.3	532.7	−34.9
TPIL-ID-VDB	**−4.4**	**775.5**	**−19.7**

结束语　模仿学习方法可以从专家示范中学习得到专家策略，但是传统的模仿学习方法通常需要第一人称的专家示范，而相比于第一人称的专家示范，第三人称专家示范更易获得。为了能够使用第三人称专家示范进行模范学习，本文提出了一种数据高效的第三人称模仿学习方法 TPIL-ID-VDB。其主要思想是：1）在判别器网络中引入图像差分模块，使得判别器更多地关注样本中与行为特征相关的变化区域，并削弱领域特征所带来的影响；2）引入变分判别器瓶颈，以限制判别器对生成样本的判别，进一步缓解因存在领域特征差异而导致判别器过强的问题。本文选取了 3 种 MuJoCo 环境，验证了 TPIL-ID-VDB 在第三人称模仿学习问题上的性能表现，同时与不同实验条件下的 GAIL 和 TPIL 进行了对比，实验结

果显示了 TPIL-ID-VDB 的优异性。

然而，我们注意到 TPIL-ID-VDB 仍然存在一些有待改进的地方，如在图像差分模块中，观察图像中相对静止的目标点等有用信息也会被一同去除。因此，下一步的研究重点是如何更好地保留观察图像中的有用信息，以更好地指导智能体进行模仿学习。

参 考 文 献

[1] LIU Q, ZHAI J W, ZHANG Z Z, et al. A Survey of Deep Reinforcement Learning [J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[2] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489.

[3] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676): 354-359.

[4] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning [C] // Proceedings of the Workshops at the 27th Neural Information Processing Systems (NIPS). 2013: 201-220.

[5] SUTTON R S, BARTO A G. Reinforcement learning: An introduction (2nd edition) [M]. MIT Press, 2018.

[6] SCHAAL S. Is imitation learning the route to humanoid robots? [J]. Trendsin Cognitive Sciences, 1999, 3(6): 233-242.

[7] OSA T, PAJARINEN J, NEUMANN G, et al. An algorithmic perspective on imitation learning [J]. Foundationsand Trends in Robotics, 2018, 7(1/2): 1-179.

[8] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning [C] // Proceedings of the 21st International Conference on Machine Learning (ICML). 2004: 1-8.

[9] NG A Y, RUSSELL S J. Algorithms for inverse reinforcement learning [C] // Proceedings of the 17th International Conference on Machine Learning (ICML). 2000: 663-670.

[10] HO J, ERMON S. Generative adversarial imitation learning [C] // Proceedings of the 30th Neural Information Processing Systems (NIPS). 2016: 4565-4573.

[11] STADIE B C, ABBEEL P, SUTSKEVER I. Third-person imitation learning [C] // Proceedings of the 5th International Conference on Learning Representations (ICLR). 2017.

[12] SHARMA P, PATHAK D, GUPTA A. Third-person visual imitation learning via decoupled hierarchical controller [C] // Proceedings of the 33rd Neural Information Processing Systems (NIPS). 2019: 2593-2603.

[13] JIANG C, ZHANG Z Z, CHEN Z X, et al. Third-person imitation learning via image difference and variational discriminator bottleneck (student abstract version) [C] // Proceedings of the 44th AAAI Conference on Artificial Intelligence (AAAI). 2020.

[14] TODOROV E, EREZ T, TASSA Y. Mujoco: A physics engine for model-based control [C] // 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012: 5026-5033.

[15] LIN J H, ZHANG Z Z, JIANG C, et al. A Survey of imitation learning based on Generative Adversarial Nets [J]. Chinese Journal of Computers, 2020, 43(2): 326-351.

[16] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] // Proceedings of the 28th Neural Information Processing Systems (NIPS). 2014: 2672-2680.

[17] MEREL J, TASSA Y, TB D, et al. Learning human behaviors from motion capture by adversarial imitation [J]. arXiv: 1707.02201, 2017.

[18] TORABI F, WARNELL G, STONE P. Generative adversarial imitation from observation [J]. arXiv: 1807.06158, 2018.

[19] TZENG E, HOFFMAN J, ZHANG N, et al. Deep domain confusion: maximizing for domain invariance [J]. arXiv: 1412.3474, 2014.

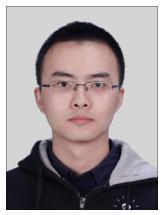
[20] GANIN Y, LEMPITSKY V. Unsupervised domain adaptation by backpropagation [J]. arXiv: 1409.7495, 2014.

[21] PENG X B, KANAZAWA A, TOYER S, et al. Variational discriminator bottleneck: improving imitation learning, inverse RL, and GANs by constraining information flow [J]. arXiv: 1810.00821, 2018.

[22] ALEMI A A, FISCHER I, DILLON J V, et al. Deep variationalinformation bottleneck [J]. arXiv: 1612.00410, 2016.

[23] KINGMA D P, BA J L. Adam: a method for stochastic optimization [C] // Proceedings of the 4th International Conference on Learning Representations (ICLR). 2015.

[24] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization [C] // Proceedings of the 32nd International Conference on Machine Learning (ICML). 2015: 1889-1897.



JIANG Chong, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include imitation learning and reinforcement learning.

ZHANG Zong-zhang, born in 1985, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include reinforcement learning, intelligent planning and multi-agent systems.