第53卷第5期 南京航空航天大学学报 Vol. 53 No. 5
2021年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2021

DOI:10.16356/j.1005-2615.2021.05.014

基于示范主动采样的行为克隆方法

黄文宇,黄圣君

(南京航空航天大学计算机科学与技术学院/人工智能学院,南京 211106)

摘要: 深度强化学习在学习过程中需要与环境进行大量的交互,训练效率低下。模仿学习通过从专家示范中学习,可以有效地应对这一挑战,但是需要收集大量的专家示范轨迹,在复杂任务中往往导致高昂的示范代价。本文提出一种基于主动学习的行为克隆算法,通过主动挑选示范起始状态来减小示范代价。该方法基于不确定性采样和不相似性采样两种策略,从状态候选集中挑选最有价值的状态作为起始状态,然后向专家查询固定长度的示范轨迹,希望从尽可能少的示范中学习出有效策略。在多个不同任务上的实验表明,本文方法可以用更少的示范轨迹进行行为克隆,降低了强化学习中的专家示范代价。

关键词: 强化学习;模仿学习;行为克隆;逆强化学习;主动学习

中图分类号: TP18 **文献标志码:** A **文章编号:** 1005-2615(2021)05-0766-06

Behavioral Cloning with Active Sampling of Demonstration

HUANG Wenyu, *HUANG Shengjun*

(College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China)

Abstract: Deep reinforcement learning has achieved great success in many applications. However, it usually needs large amount of interactions with the environment to learn the policy, which leads to inefficient training. Imitation learning is an important approach to tackle this challenge by learning from demonstrations, but it instead requires a large set of demonstrations provided by experts, which could be rather costly in many complex tasks. In this paper, we propose an active learning method to reduce the demonstration cost by actively selecting starting state for demonstration. The method is based on uncertainty sampling and dissimilarity sampling. It selects the best state from the candidate set and then queries expert for fixed length of trajectory, in order to train effective policy with fewer demonstrations. Experimental results in multiple environments demonstrate that the proposed method can achieve effective performance with significant lower demonstration cost.

Key words: reinforcement learning; imitation learning; behavioral cloning; inverse reinforcement learning; active learning

强化学习[1]旨在为智能决策任务学习出有效的策略,使智能体获得的长远奖赏最大。传统的强化学习更多关注离散状态和动作空间的任务,难以在状态和动作连续的任务上应用。深度强化学习通过将策略用深度神经网络来表示可以有效地解决这一问题。最近研究表明,深度强化学习在很多富有挑战性的任务上都取得了成功,例如围棋[2]、游戏[3]和模拟机器人任务[4-5]。但是深度强化学习在训练智能体的策略时需要与环境进行大量的交互,因此面临着训练效率低下的挑战。模仿学习通过从专家的示范中学习可以有效应对这一挑战,其主要思想是从专家的示范中去模仿专家的行为,因

基金项目: 航空动力基金(6141B09050342)资助项目。

收稿日期: 2020-11-10;**修订日期:** 2021-01-06

通信作者: 黄圣君,男,博士,教授,E-mail:huangsj@nuaa.edu.cn。

引用格式: 黄文宇,黄圣君.基于示范主动采样的行为克隆方法[J].南京航空航天大学学报,2021,53(5):766-771.
HUANG Wenyu, HUANG Shengjun. Behavioral cloning with active sampling of demonstration[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5):766-771.

而无需与环境进行交互。

模仿学习大体上可被分为两大类：行为克隆和逆强化学习。行为克隆[6-7]运用监督学习的方式直接从示范中学得一个策略，其将状态视为监督学习中的示例，将动作视为监督学习中的标签。与行为克隆直接学得一个策略不同，逆强化学习[8-9]首先学得一个奖赏函数，然后通过标准的强化学习算法学习策略。生成对抗模仿学习[10]是当前较为前沿的模仿学习方法，其主要思想是同时学习策略和判别器。判别器的目标是将专家生成的状态-动作对与智能体生成的状态-动作对有效区分，而智能体策略的目标是混淆判别器，使得判别器将智能体生成的状态-动作对判别为专家生成的状态-动作对。

尽管模仿学习可以缓解训练效率低下的问题，但是现有的模仿学习算法需要获得大量的专家示范作为训练数据，在实际任务中往往导致高昂的示范代价。例如Waymo公司为了训练自动驾驶的智能体收集了3 000万个专家驾驶的数据，专家提供每一个数据都需付出时间上的代价，在提供路况不好的驾驶数据时更是要面对安全风险。主动学习[11]是监督学习任务中降低标注成本的一类主流方法。它通过挑选最有价值的样本向专家查询，可以有效地降低训练所需样本。关于主动学习的大量研究都专注于如何设计好的选择标准以便更好地计算样本的价值。不确定性采样[12-13]是最常用的一种选择策略，它倾向于选择分类器的预测最不确定的样本。委员会采样[14-15]是另一种常用的选择策略，它从训练集的多个子集中学习多个模型，然后选取多个模型的预测分歧最大样本进行查询。最近，有一些工作试图将样本的信息量与代表性相结合来评估样本的价值[16-17]。

目前主动学习多应用于传统的分类任务，应用于模仿学习的工作较少。文献[18]将主动学习应用于逆强化学习，选取奖赏最不确定的状态并查询对应的动作。但该方法在现实任务中的应用存在较大局限性，因为对于专家而言，提供一条轨迹比起提供单个动作更方便。以驾驶为例，选取道路上的某个点，专家进行一段时间的控制显然比只做一个动作方便。文献[19]将主动学习应用于自动导航任务中，它从起点-终点对的候选集中挑选起点-终点对，查询起点至终点的路径，然而在其他任务中无法保证专家的示范轨迹一定会通过终点。同时，以上的两个工作都是应用在逆强化学习中，难以直接应用于行为克隆方法。本文提出了一种基于示范主动采样的行为克隆方法，目的是以更少的示范代价学得一个有效的策略。具体地，本文提出了不确定性采样和不相似性采样两种方法，挑选完状态后向专家查询固定长度的示范轨迹，并进一步用于策略更新。

1　背景知识

本文方法中的基础模型涉及到近端策略优化算法和行为克隆方法，因此本节先对其进行简要介绍。

1.1　近端策略优化算法

近端策略优化（Proximal policy optimization，PPO）算法是一种深度强化学习算法。设π_θ表示更新后的策略，θ为其参数，π_{old}表示旧的策略，$\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)$表示在状态$\boldsymbol{s}_t$采取动作$\boldsymbol{a}_t$的概率，$t$是时间步长，令$r_t(\theta)=\dfrac{\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\text{old}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}$，近端策略优化算法通过最小化如下目标训练智能体的策略，有

$$\hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t,\text{clip}(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)] \tag{1}$$

式中：$\hat{\mathbb{E}}_t$为期望的估计量；$\hat{A}_t$为优势函数的估计量；clip(•)为修剪函数，其作用是将$r_t(\theta)$限制在区间$[1-\epsilon,1+\epsilon]$内，使得更新后的策略参数与旧的策略参数相差不会过大，因此才能使得PPO利用重要性采样进行估计时足够精确，最终学得的策略性能足够好。

1.2　行为克隆

行为克隆是一种通过对示范集合运用监督学习，从而直接学习智能体策略的算法。设示范集合D由n个状态-动作对构成，即$D=\{(\boldsymbol{s}_1,\boldsymbol{a}_1),(\boldsymbol{s}_2,\boldsymbol{a}_2),\cdots,(\boldsymbol{s}_n,\boldsymbol{a}_n)\}$，其中$\boldsymbol{s}_i$为状态，$\boldsymbol{a}_i$为专家示范的动作。设智能体的策略为$\pi$，智能体的策略可通过最小化如下目标函数得到

$$\min\sum_i^n(\pi(\boldsymbol{s}_i)-\boldsymbol{a}_i)^2 \tag{2}$$

2　基于主动学习的行为克隆方法

本文提出的基于主动采样的行为克隆方法框架如图1所示。在每一轮迭代过程中，该方法首先从示范集合D中训练智能体的策略π，然后从候选集U_s中挑选最有价值的状态$\boldsymbol{s}_1$，并向专家查询示范轨迹，专家以该状态为起点，返回一条长度为n的示范轨迹$d=\{(\boldsymbol{s}_1,\boldsymbol{a}_1),\cdots,(\boldsymbol{s}_n,\boldsymbol{a}_n)\}$。之后该轨迹中的状态-动作对会被加入示范集合中，用作策略的重新训练。

2.1　不确定性采样

第1种选择策略是从候选集中选取当前策略的决策动作最不确定的状态。其动机是，如果策略对于某状态的动作越不确定，那么以该状态为起点的轨迹对学习策略的帮助越大。以自动驾驶为例，

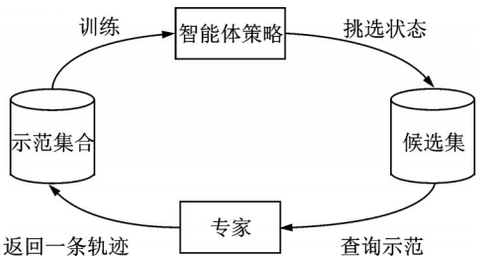

图1 基于主动采样的行为克隆方法框架

Fig.1 Framework of behavioval cloning with active sampling

假设任务的目的是训练一个能够有效驾驶的智能体，如果该智能体只在直道上训练过，那么它在遇到直道时将更确定如何控制。相反地，如果它遇到一个弯道，那么对于在弯道如何控制一定会有较大的不确定性。通过不确定性采样，一个弯道更有可能被挑选到，以其为起点的示范轨迹会对策略的性能提升作用更大。

在传统的主动学习中，不确定性利用分类器的预测去计算，如以各类别概率的熵作为不确定性，或者以最大的类别概率与第二大类别概率的差衡量不确定性[20]。然而该采样策略无法直接运用于强化学习问题，因为对于连续动作的任务不存在类似的分类器。考虑到智能体的随机策略由动作分布来表示，本文基于动作分布估计策略的不确定性，以动作分布的标准差作为不确定性的标准，有

$$\text{uncertainty}(s)=\text{std}(\pi(\boldsymbol{a}|\boldsymbol{s})) \tag{3}$$

式中：std(·)为标准差函数，用于计算分布的标准差；$\pi(\boldsymbol{a}|\boldsymbol{s})$为状态$\boldsymbol{s}$下动作$\boldsymbol{a}$的概率分布。

对于多维动作，动作向量的不确定性可视为所有动作元素的不确定性之和。在本文模型中，动作向量的每个元素a_i都遵循正态分布$N(\mu_i,\sigma_i)$，μ、σ分别为均值和标准差，因此a_i的不确定性可用σ衡量，状态$\boldsymbol{s}$的不确定性为

$$\text{uncertainty}(\boldsymbol{s})=\sum_i^m \sigma_i \tag{4}$$

式中m为动作向量的维度。计算完候选集中所有状态的不确定性之后，不确定性最大的状态会被挑选出来，以让专家提供最有价值的示范轨迹。

$$\boldsymbol{s}^*=\text{argmax}_{s\in U_s}\text{uncertainty}(\boldsymbol{s}) \tag{5}$$

2.2 不相似性采样

第2种策略挑选和示范集合中已有状态最不相似的状态。其动机是如果某状态和示范集合中的状态很相似，那么智能体可能已经学会如何在该状态进行决策，因此以该状态为起点的示范轨迹对智能体的帮助不大。再次以自动驾驶为例，假如智能体的策略已经在直道上表现得很好，如果遇到的状态仍然是一条直道，那么对应的示范轨迹对于学习策略的帮助很小。相反地，如果某状态是一个障碍物，由于策略从未在与该状态相似的状态下训练过，因此对应的示范轨迹对于策略的提升有更大的作用。

考虑到不相似的状态不太可能有相同的动作，因此用动作的差异来衡量状态的差异。对于随机策略来说，两个动作分布之间的距离是一个衡量分布差异的很好的标准。对于计算两个分布之间的距离，已经有不少的研究工作，比如KL散度[21]和最大均值差异(Maximum mean discrepancy，MMD)距离[22]。在这些方法中，本文选用Wasserstein距离[23]作为度量距离的标准。具体来说，为了计算状态$\boldsymbol{s}$与示范集合中状态的不相似度，需依次计算该状态与集合中每个状态的不相似度并求均值

$$\text{dissimilarity}(\boldsymbol{s})=\frac{1}{n}\sum_i^n W_2(\pi(\boldsymbol{a}|\boldsymbol{s}),\pi(\boldsymbol{a}_i|\boldsymbol{s}_i)) \tag{6}$$

式中：n为集合中状态的数量；W_2为两个分布间的2-Wasserstein距离，定义为

$$W_2(p,q)=\inf\left(\mathbb{E}\left|\left|\boldsymbol{x}-\boldsymbol{y}\right|\right|_2^2\right)^{\frac{1}{2}} \tag{7}$$

式中：p和q为两个概率分布，且$\boldsymbol{x}\sim p$，$\boldsymbol{y}\sim q$。在本文模型中，以多元正态分布来表示随机策略，文献[24]证明了两个多元正态分布的2-Wasserstein距离的计算方式为

$$\left|\left|\boldsymbol{\mu}_x-\boldsymbol{\mu}_y\right|\right|^2+\text{tr}\left(\boldsymbol{\Sigma}_x+\boldsymbol{\Sigma}_y-2(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_y)^{\frac{1}{2}}\right) \tag{8}$$

式中：$\boldsymbol{x}\sim N(\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$，$\boldsymbol{y}\sim N(\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)$；$\boldsymbol{\mu}$为分布的均值向量；$\boldsymbol{\Sigma}$为分布的协方差矩阵；tr(·)为矩阵的迹。在本文的模型中，任意两个动作元素的协方差为0，因此$\boldsymbol{\Sigma}_x$和$\boldsymbol{\Sigma}_y$都为对角矩阵，式(8)可以改写为

$$W_2(p,q)=\left|\left|\boldsymbol{\mu}_x-\boldsymbol{\mu}_y\right|\right|^2+\left|\left|\boldsymbol{\Sigma}_x^{\frac{1}{2}}-\boldsymbol{\Sigma}_y^{\frac{1}{2}}\right|\right|_{\text{F}}^2 \tag{9}$$

式中加号右边的项很容易推出因为tr(·)函数里的项等于$\left(\boldsymbol{\Sigma}_x^{\frac{1}{2}}-\boldsymbol{\Sigma}_y^{\frac{1}{2}}\right)^2$，其中$\boldsymbol{\Sigma}_x^{\frac{1}{2}}$和$\boldsymbol{\Sigma}_y^{\frac{1}{2}}$都为对角矩阵，对角线上每个元素恰好是动作元素的标准差。

计算完每个状态的不相似度以后，不相似性最大的状态会被挑选出来并向专家查询，有

$$\boldsymbol{s}^*=\text{argmax}_{s\in U_s}\text{dissimilarity}(\boldsymbol{s}) \tag{10}$$

算法1总结了本文提出的方法。算法的输入是初始的示范集合D，包含了少量的示范，以及未标记状态集U_s，示范轨迹长度H，专家的策略π_{E}和迭代次数T。在每轮迭代中，算法首先根据不确定性采样或者不相似性采样选择状态，然后专家以提供的状态为起点做示范，示范结束后返回一条长度

为 H 的示范轨迹；接着示范轨迹中的状态-动作对会被加入示范集合中；同时，示范集合中的状态会从候选集中移除；最后更新智能体的策略。

算法 1　面向行为克隆的主动学习方法

输入：初始示范集合 D，未标记状态集 U_s，示范轨迹长度 H，专家的策略 π_{E}，迭代次数 T

输出：智能体的策略

```
从 D 中训练智能体的策略 π
for t = 1, …, T do
  for each s in U_s do
    计算 s 的不确定性或不相似性
  end for
  选择不确定性或者不相似性最大的状态 s_1
  初始化示范轨迹 τ = ∅
  for i = 1, …, H do
    a_i = π_E(s_i)
    将 (s_i, a_i) 加入 τ
    s_i 转移至 s_{i+1}
  end for
  将 τ 中所有状态-动作对加入 D
  将 τ 中所有状态从 U_s 中移除
  从 D 中学习策略 π
end for
```

3　实验过程和结果

3.1　任务介绍

实验中所有的任务都在 OpenAI Gym[25] 环境库中定义，并在 MuJoCo[26] 上模拟。下面对其进行简单介绍。

（1）HalfCheetah。此任务目标是让一个 2 维猎豹跑得尽可能地快（https://gym.openai.com/envs/HalfCheetah-v2/）。在这个任务中，状态由 17 维的向量表示，动作由 6 维的向量表示。

（2）Hopper。此任务目的是让一个只有一条腿的机器人尽可能快地向前跳（https://gym.openai.com/envs/Hopper-v2/）。状态由 11 维向量表示，动作由 3 维向量表示。

（3）Swimmer。此任务智能体是一个有 3 个关节的游泳机器人，它的目标是在粘性液体中尽可能快地游泳（https://gym.openai.com/envs/Swimmer-v2/）。状态由 8 维向量表示，动作由 2 维向量表示。

（4）Walker2d。此任务目的是让一个 2 维的双足动物机器人尽可能快地向前走（https://gym.openai.com/envs/Walker2d-v2/）。在这项任务中，状态由 17 维的向量表示，动作由 6 维的向量表示。

3.2　实验设置

专家和智能体策略均由 3 层神经网络构成，每层均为全连接层，激活函数为 tanh，其中隐藏层的神经元数量为 100，输入层的神经元数量等于状态的维度，输出层神经元的数量 2 倍于动作的维度，其中一半神经元输出每个动作元素的均值，另一半神经元输出每个动作元素的标准差。在实验中先用 PPO 算法训练策略作为专家策略，以用来模拟专家提供示范轨迹。在用行为克隆算法训练智能体的策略时，用 Adam 优化器进行优化，每次迭代选取的 Batch 大小为 128，算法迭代 10 000 次。由于任务的状态均为连续向量，状态空间无穷大，因此先用专家策略生成部分示范轨迹，然后将轨迹中的状态作为候选集。

由于本文提出的是一个全新问题，没有相关方法可以直接应用到该问题中，因此实验部分将提出的方法不确定性采样（Uncertainty）和不相似性采样（Dissimilarity）与随机采样（Random）对比。对于每个方法而言，其初始示范集合均一样，训练完智能体的策略后，将学得的智能体与环境进行交互，生成 50 条长为 1 000 的轨迹，计算 50 条轨迹的平均累积奖赏作为策略的性能，每个算法均进行 5 次实验，每次实验都随机初始化示范集合，然后取 5 次的平均值作为最终的性能。

为了验证提出方法的鲁棒性，在每个任务上，均设置了不同的示范轨迹的长度，在 HalfCheetah 任务上，将长度设置为 50，100 和 500；在 Hopper 上，轨迹的长度同样设置为 50，100 和 500；在 Swimmer 任务上，轨迹的长度设置为 20，100 和 500；在 Walker2d 任务上，长度设置为 300，500 和 1 000。

随着迭代次数的增加，将所查询的示范轨迹的总长度作为横坐标，将每轮迭代完成后，策略从环境获得的累计奖赏作为纵坐标，绘制了不同方法所对应的奖赏曲线，并对比在查询了同样长度的示范轨迹后，基于不同采样方法所训练得到的策略性能。

3.3　实验结果

图 2 是实验的结果，其中：每 1 行对应 1 个任务，每 1 行的每 1 列对应 1 种轨迹长度的设置；红线对应不确定性采样，蓝线对应不相似性采样，黑线对应随机采样。从图 2 中可以很容易看出，在所有任务中，在任意长度设置下，本文所提出的两种方法都显著地优于对比方法，其中不确定性采样的效果最好。在查询了同样长度轨迹后，基于不确定性采样和不相似性采样训练得到的策略，其获得的奖赏远大于基于随机采样的策略获得的奖赏。可以发现，本文所提出的方法以更少的示范学得了性能更优的策略，这表明提出的方法可以有效地降低示范代价。

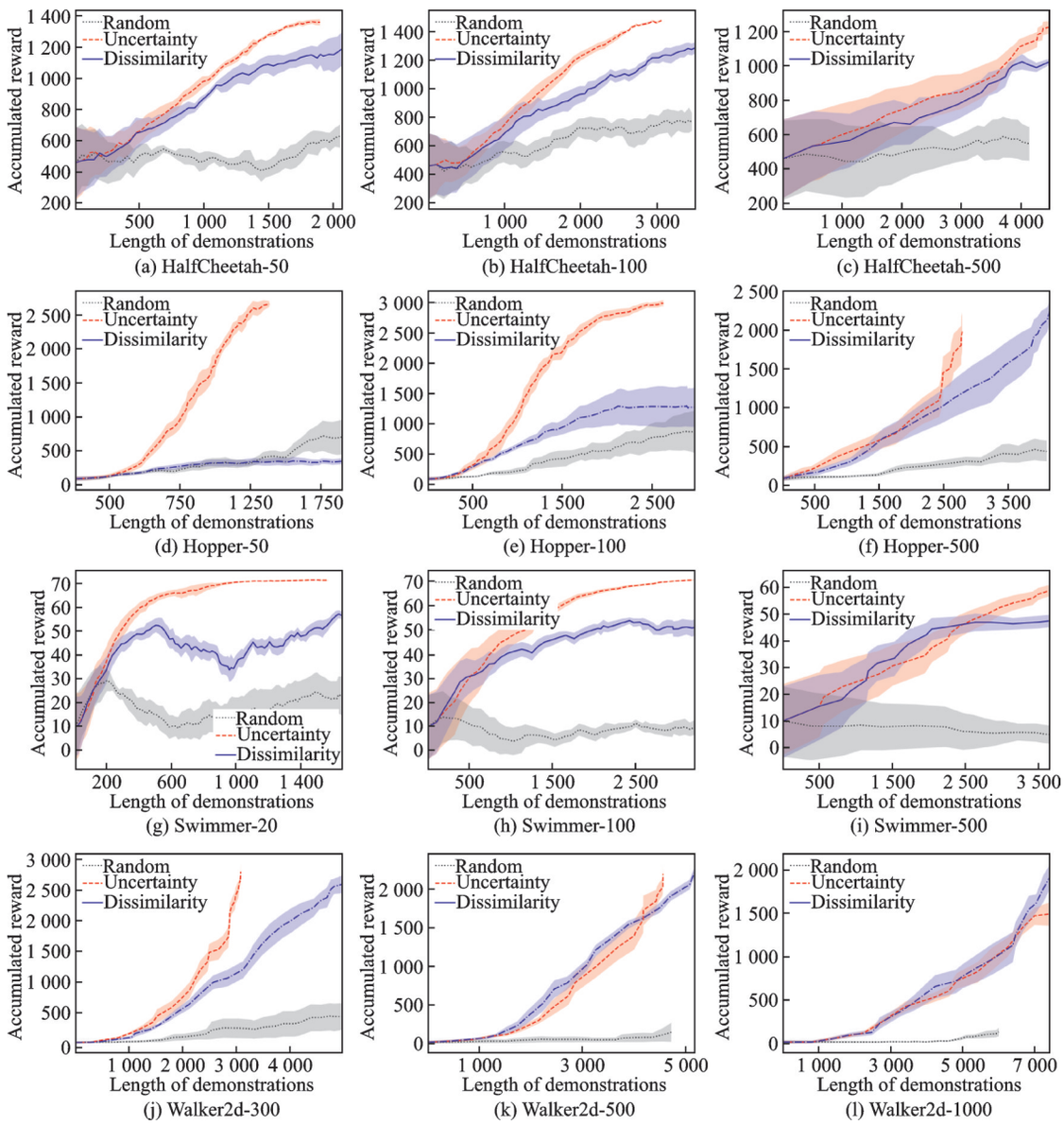

图 2 4个任务上的累计奖赏对比结果

Fig.2 Comparison results of accumulated reward on four tasks

3.4 讨 论

对于本文提出的方法，如果将其在轨迹长度更短时的表现与其在轨迹长度更长时的表现对比，可以发现本文方法在轨迹长度更短时的效果更优越，一个可能的原因是更长的轨迹会有更大的概率含有冗余信息。以自动驾驶为例，假如学得的策略已经可以在直道上进行有效控制，若此时挑选出的状态是一个障碍物，而障碍物后面又是一条直道，那么更长的轨迹就会有更大的概率包含后面的直道，从而有更大的概率含有冗余信息。

4 结 论

本文提出了基于示范主动采样的行为克隆方法，目的在于减少行为克隆算法的示范代价。具体的，本文提出了不确定性采样和不相似性采样两种方法，试图挑选出对于策略性能提升帮助最大的示范轨迹。实验结果表明，本文方法的效果显著优于对比方法，其中不确定性采样的效果最好。相比随机采样，本文方法显著地降低了示范代价，同时训练的策略性能更好。在以后的研究工作中，将计划设计一种自适应调整示范长度的方法，进一步提升基于主动采样行为克隆的实用性。

参考文献：

[1] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 2018: 1-2.

[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] PANG Z J, LIU R Z, MENG Z Y, et al. On reinforcement learning for full-length game of starcraft [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]:AAAI, 2019, 33: 4691-4698.

[4] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. (2017-8-28) [2020-8-5]. https://arxiv. org/abs/1707.06347.

[5] WANG Y, HE H, TAN X, et al. Trust region-guided proximal policy optimization [C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]:[s.n.], 2019: 626-636.

[6] ROSS S, GORDON G, BAGNELL D. A reduction of imitation learning and structured prediction to no-regret online learning [C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. [S.l.]:[s.n.], 2011: 627-635.

[7] TORABI F, WARNELL G, STONE P. Behavioral cloning from observation [EB/OL]. (2018-5-11) [2020-8-5]. https://arxiv.org/abs/1805.01954.

[8] NG A Y, RUSSELL S J. Algorithms for inverse reinforcement learning [C]//Proceedings of ICML. [S.l.]:[s.n.], 2000, 1: 2.

[9] WULFMEIER M, ONDRUSKA P, POSNER I. Maximum entropy deep inverse reinforcement learning [EB/OL]. (2016-3-11) [2020-8-5]. https://arxiv.org/abs/1507.04888.

[10] HO J, ERMON S. Generative adversarial imitation learning [C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]:[s.n.], 2016: 4565-4573.

[11] SETTLES B. BURR S. Active learning literature survey [R]. Madison: University of Wisconsin-Madison Department of Computer Sciences, 2009.

[12] TONG S, Koller D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2001, 2: 45-66.

[13] YAN Y F, HUANG S J. Cost-effective active learning for hierarchical multi-label classification [C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. [S.l.]: IJCAI, 2018:2962-2968.

[14] SEUNG H S, OPPER M, SOMPOLINSKY H. Query by committee [C]//Proceedings of the ACM Workshop on Computational Learning Theory. [S.l.]: ACM, 1992: 287-294.

[15] VANDONI J, ALDEA E, LE HÉGARAT-MASCLE S. Evidential query-by-committee active learning for pedestrian detection in high-density crowds [J]. International Journal of Approximate Reasoning, 2019, 104: 166-184.

[16] HUANG Shengjun, JIN Rong, ZHOU Zhihua. Active learning by querying informative and representative examples [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014 (10):1936-1949.

[17] WANG Zheng, YE Jieping. Querying discriminative and representative samples for batch mode active learning [J]. TKDD, 2015, 9(3):1-23.

[18] LOPES M, MELO F, MONTESANO L. Active learning for reward estimation in inverse reinforcement learning [C]//Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. [S.l.]:[s.n.], 2009:31-46.

[19] SILVER D, BAGNELL J A, STENTZ A. Active learning from demonstration for robust autonomous navigation [C]//Proceedings of 2012 IEEE International Conference on Robotics and Automation. [S.l.]: IEEE, 2012:200-207.

[20] SCHEFFER T, DECOMAIN C, WROBEL S. Active hidden Markov models for information extraction [C]//Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA). [S.l.]:[s.n.], 2001: 309-318.

[21] BISHOP C M. Pattern recognition and machine learning [M]. New York: Springer, 2006: 54-55.

[22] GRETTON A, BORGWARDT K M, RASCH M J, et al. A kernel two-sample test [J]. Journal of Machine Learning Research, 2012, 13:723-773.

[23] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN [EB/OL]. (2017-12-6) [2020-8-5]. https://arxiv.org/abs/1701.07875.

[24] DOWSON D C, LANDAU B V. The frechet distance between multivariate normal distributions [J]. Journal of Multivariate Analysis, 1982, 12(3):450-455.

[25] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. OpenAI Gym [EB/OL]. (2016-6-5) [2020-8-5]. https://arxiv.org/abs/1606.01540.

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

（编辑：刘彦东）