中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于环境反馈机制的四足机器人运动技能学习

张思远, 朱晓庆, 阮晓钢, 李春阳, 刘鑫源

引用本文:
张思远,朱晓庆,阮晓钢,李春阳,刘鑫源. 基于环境反馈机制的四足机器人运动技能学习[J]. 控制与决策, 2024, 39(5): 1461–1468.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1714

基于环境反馈机制的四足机器人运动技能学习

张思远[1,2], 朱晓庆[1,2†], 阮晓钢[1,2], 李春阳[1,2], 刘鑫源[1,2]

(1. 北京工业大学 信息学部，北京 100124；2. 北京计算智能与智能系统重点实验室，北京 100124)

摘 要: 哺乳动物的运动学习机制已得到广泛研究,犬科动物可以根据环境反馈的引导性信息自主地学习运动技能,对其提供更为特定的训练引导可以加快其对相关任务的学习速度. 受上述启发,在软演员-评论家算法(SAC)的基础上提出一种基于期望状态奖励引导的强化学习算法(DSG-SAC),利用环境中的状态反馈机制来引导四足机器人进行有效探索,可以提高四足机器人仿生步态学习效果,并提高训练效率. 在该算法中,策略网络与评价网络先近似拟合期望状态观测与当前状态的误差,再经过当前状态的正反馈后输出评价函数与动作,使四足机器人朝着期望的方向动作. 将所提出算法在四足机器人上进行验证,通过实验结果可知,所提出的算法能够完成四足机器人的仿生步态学习. 进一步,设计消融实验来探讨超参数温度系数和折扣因子对算法的影响,实验结果表明,改进后的算法具有比单纯的SAC算法更加优越的性能.

关键词: 强化学习；四足机器人；仿生步态学习；环境探索；状态反馈引导

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1714

引用格式: 张思远, 朱晓庆, 阮晓钢, 等. 基于环境反馈机制的四足机器人运动技能学习 [J]. 控制与决策, 2024, 39(5): 1461-1468.

Motor skill learning of quadruped robot based on environmental feedback mechanism

ZHANG Si-yuan[1,2], *ZHU Xiao-qing*[1,2†], *RUAN Xiao-gang*[1,2], *LI Chun-yang*[1,2], *LIU Xin-yuan*[1,2]

(1. Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China；2. Beijing Key Laboratory of Computational Intelligence and Intelligent System，Beijing 100124，China)

Abstract: The motor learning mechanism of mammals has been extensively studied, and the learning speed of canines for relevant tasks can be accelerated by conducting guided training on them. According to the above inspiration, this paper proposes a reinforcement learning algorithm based on desired state reward guidance (DSG-SAC) on the basis of soft actor-critic algorithm (SAC). This algorithm uses the state feedback mechanism in the environment to guide the quadruped robot to explore effectively, which can improve the bionic gait learning effect of the quadruped robot and improve the training efficiency. In this algorithm, the strategy network and the evaluation network first approximate the error between the desired state observation and the current state, and after the positive feedback from the current state, the evaluation function and the action are output, so that the quadruped robot moves in the desired direction. In this thesis, the algorithm is verified on a quadruped robot, and the experimental results can be concluded that the proposed algorithm can complete the bionic gait learning of the quadruped robot. Ablation experiments are designed to investigate the effects of hyperparametric temperature coefficients and discount factors on the algorithm, and finally experiments are designed to verify that the improved algorithm has superior performance than the simple SAC algorithm.

Keywords: reinforcement learning；quadrupedal robot；bionic gait learning；environmental exploration；state feedback guidance

0 引 言

为了方便处理日益复杂且危险的任务,机器人技术发展得愈发成熟. 受自然界四足哺乳动物启发的四足机器人具有独特的仿生结构,4条支撑腿使其具有比两足机器人更高的承载能力和稳定性[1]. 四足机器人的结构比六足机器人更简单[2],在便于建模仿真的同时也能够产生速度更快的运动,与轮式机器人相比[3],四足机器人可以快速通过复杂地形,具有

收稿日期: 2022-09-28；**录用日期:** 2023-06-14.

基金项目: 国家自然科学基金项目(62103009)；北京市自然科学基金项目(4202005).

责任编委: 高会军.

†通讯作者. E-mail: alex.zhuxq@bjut.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

更好的地形适应能力和机动性,在军事、工业、救援等相关领域中得到广泛应用,也吸引了许多学者对其进行研究并提出了许多算法[4],例如波士顿动力公司的Cheetah[5-6]、苏黎世联邦理工学院的ANYmal[7]等. 但是,经典的机器人控制算法,例如中央模式发生器(CPG)[8]、模型预测控制(MPC)[9]等,通过模仿动物行走时产生的节律步态,使四足机器人腿部产生周期性动作,这些信号是由人为调制产生的、针对特定地形的节律信号,并不是通过环境交互机制使其自主学习出的步态,无法实现仿生意义上的参考轨迹,因此不是理想的参考轨迹[10],同时这种调制步态一旦遇到动态未知的环境,四足机器人就无法在复杂地形上稳定行走,存在很大的局限性[11].

强化学习[12]通过启发式方法让智能体与环境进行交互,通过反馈的状态对智能体进行奖励或者惩罚,使智能体具有自我学习的能力[13]. 简单的强化学习(例如Q-learning、SARSA[14-15]等)只能处理离散动作序列,无法处理连续动作环境下智能体的学习问题. 因此,将深度神经网络引入强化学习,这种深度强化学习(例如DQN、DDPG、PPO等[16-18])通过神经网络来拟合Q值与价值函数,可以在连续动作场景下产生Q值,使智能体做出相应动作进而完成学习过程. 深度强化学习的出现为机器人控制提供了全新思路[19-20],目前,已成功地实现在Minitaurs、Spot中四足行走,并逐渐成为机器人运动技能学习的研究热点之一[21-22].

随着对强化学习的进一步研究,强化学习的一些局限也随之出现,例如:智能体与环境交互反馈耗时较长,导致训练时间较长;智能体在与环境交互过程中存在随机性,如果使用单纯强化学习而不加外部引导或者约束,很容易使训练结果不朝着期望的方向进行,同时不加约束也会一定程度上导致训练时间较长.

为了解决上述问题,本文提出一种基于期望状态奖励引导的软演员-评论家算法(desired state reward guidance-soft actor critic, DSG-SAC),控制四足机器人进行步态学习,并通过仿真与目前较为热门的强化学习算法进行对比,验证本文所提出的算法具有更好的性能.

1 相关工作

1.1 强化学习

强化学习(reinforcement learning, RL)[23]的目的是让智能体在不确定的环境中自主地学习一个可行策略,智能体通过环境反馈的状态值做出动作,然后根据动作进入下一个状态并获得对应的奖励或者惩罚,再更新参数,重复这一过程,实现智能体自主学习. 强化学习过程本质为马尔可夫决策过程(Markov decision process, MDP). MDP可以用一个四元数组表示$E=\langle S,A,P,R\rangle$. 其中:S为MDP过程中智能体的状态集合;A为智能体的动作集合;P为状态转移概率;R为智能体与环境交互所得到的奖励函数,即

$$R_t=\sum_{i=0}^{T}\gamma^i r_i. \tag{1}$$

奖励函数定义为未来一段时间内的累计折扣奖励之和. 其中:r_i为第i时刻的奖励值;γ为折扣因子,代表未来奖励值对当前奖励的重要程度,时间越远的奖励对当前状态的评估影响越小. 通过奖励函数可以得出用于定义状态价值的状态价值函数(state value function),即

$$V(s)=r(s)+\gamma\sum_{s'\in S}P(s'|s)V(s'). \tag{2}$$

其中:s表示当前状态,s'表示下一时刻的状态. 通过迭代可以得到MDP的状态-动作价值函数为

$$q(s,a)=R(s,a)+\gamma\sum_{s'\in S}P(s'|s,a)V(s'). \tag{3}$$

设智能体的初始状态为S_1,在策略Π中服从某一分布ρ^π,则强化学习的任务为学习一个策略,使智能体在与环境交互过程中的目标函数J最大,有

$$J=E_{s\sim\rho,a\sim\pi}[R_i]. \tag{4}$$

1.2 SAC算法

SAC算法[24-25]是以actor-critic框架为基础,融合soft Q-learning中最大熵原理[26]的策略梯度算法. 该算法与标准RL的不同之处在于SAC将策略的熵整合到奖励目标中,以期望同时最大化预期回报和熵来鼓励智能体尽可能学习一个随机行为策略,避免智能体陷入局部最优策略,有

$$J(\theta)=\sum_{t=1}^{T}E_{(s_t,a_t)\sim\rho_{\pi\theta}}[r(s_t,a_t)+\alpha\mathcal{H}(\pi_\theta(\cdot|s_t))]. \tag{5}$$

其中:$\alpha\mathcal{H}(\pi_\theta(\cdot|s_t))$是关于策略的熵,$\alpha$为温度系数,主要定义熵在目标函数中的重要性. 熵最大化能够引导智能体在策略中更多地进行探索,避免智能体因错过更好的动作选择而陷入局部最优.

SAC不直接对软Q函数、状态值函数和策略的收敛性进行评估和改进,而是将软Q函数和状态值函数建模为参数化的神经网络,将策略建模为具有由神经网络给出均值和协方差的高斯分布,并交替使用随机梯度下降优化3个网络. 因此,SAC主要学习以下3

个目标函数.

1) 参数θ的soft Q函数Q_θ. soft Q函数的参数可以通过soft贝尔曼残差来训练,其目标函数为

$$J_Q(\theta)=E_{(s_t,a_t)\sim\mathcal{D}}\Big[\frac{1}{2}(Q_\theta(s_t,a_t)-\hat{Q}(s_t,a_t))^2\Big]. \tag{6}$$

其中:$\mathcal{D}$为缓存器中采样的状态动作分布;参数θ可以通过随机梯度进行优化,即

$$\hat{\nabla}_\theta J_Q(\theta)=\nabla_\theta Q_\theta(a_t,s_t)(Q_\theta(s_t,a_t)-r(s_t,a_t)-\gamma V_{\bar{\psi}}(s_{t+1})). \tag{7}$$

2) 参数为ψ的soft状态值函数V_ψ. 软状态值函数可以用最小化平方残差来训练,即

$$J_V(\psi)=E_{s_t\sim\mathcal{D}}\Big[\frac{1}{2}(V_\psi(s_t)-E_{a_t\sim\pi_\theta}(Q_\theta(s_t,a_t)-\log\pi_\phi(a_t|s_t)))^2\Big]. \tag{8}$$

方程(8)的梯度可以用无偏估计器来估计,有

$$\hat{\nabla}_\psi J_V(\psi)=\nabla_\psi V_\psi(s_t)(V_\psi(s_t)-Q_\theta(s_t,a_t)+\log\pi_\phi(a_t|s_t)). \tag{9}$$

3) 带有参数ϕ的策略π_ϕ. 在soft策略迭代过程中,为了保证新策略能够在软值上得到改善,使用新旧策略的KL散度来定义新策略

$$\pi_{\text{new}}=\arg\min_{\pi'\in\Pi} D_{\text{KL}}\Big(\pi'(\cdot|s_t)\|\frac{\exp(Q^{\pi_{\text{old}}}(s_t,\cdot))}{Z^{\pi_{\text{old}}}(s_t)}\Big). \tag{10}$$

但因策略是由神经网络估计的,在进行参数更新时可以进行微分,故可以利用重参数化技巧重新定义策略的部分参数,降低神经网络估计出的方差

$$a_t=f_\phi(\epsilon_t,s_t). \tag{11}$$

因此目标函数可改写为

$$J_\pi(\phi)=E_{s_t\sim\mathcal{D},\epsilon_t\sim\mathcal{N}}[\log\pi_\phi(f_\phi(\epsilon_t,s_t)\|s_t)-Q_\theta(s_t,f_\phi(\epsilon_t,s_t))]. \tag{12}$$

通过随机梯度下降对参数进行更新

$$\hat{\nabla}_\phi J_\pi(\phi)=\nabla_\phi\log\pi_\phi(a_t|s_t)+(\nabla_{a_t}\log\pi_\phi(a_t|s_t)-\nabla_{a_t}Q(s_t,a_t))\nabla_\phi f_\phi(\epsilon_t;s_t). \tag{13}$$

2　系统建模

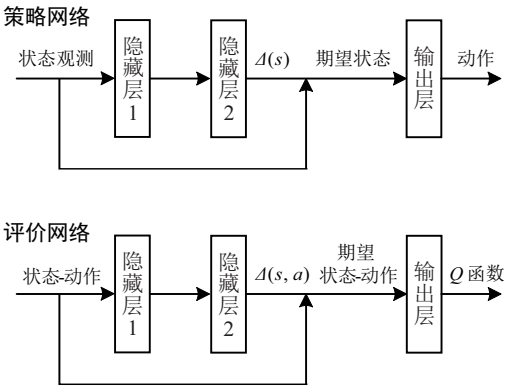
2.1　基于期望状态奖励引导的SAC算法

在深度强化学习算法中,智能体在环境中的状态是获得奖励的重要条件之一,也是智能体进行下一步动作的先验条件之一. 在连续动作空间任务下智能体采取动作后进入下一个状态,但这个状态并不都是理想的状态,无法通过神经网络拟合出期望的下个动作. 在连续动作过程中,本文将智能体自然转换的状态与理想状态的数值偏差定义为状态残差$\Delta(s)$,即

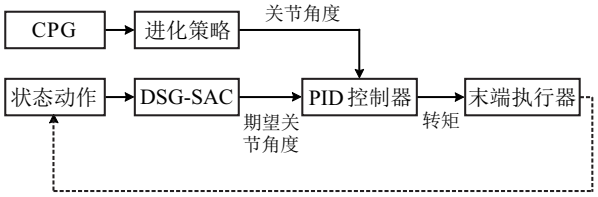
$$\Delta(s)=s_d-s_{\text{obs}}. \tag{14}$$

其中:s_d为期望状态,s_{obs}为观测状态. 因此,若能在训练过程中对智能体进行状态上的引导,减小期望状态与理想状态的残差,则可以鼓励智能体朝着期望的方向做出动作,加快学习速度,优化学习效果.

受上述智能体状态引导的启发[27],本文基于现阶段最新的深度强化学习算法SAC并结合状态引导,提出基于期望状态奖励引导的SAC算法(DSG-SAC). 算法假设存在一个从当前智能体的状态到状态残差的潜在映射,找出这个映射关系并与当前状态叠加之后就能得到期望状态. 由于神经网络具有强大的拟合性能,本文选择神经网络来拟合状态残差. DSG-SAC算法由策略网络和评价网络组成,其策略网络和评价网络结构如图1(a)所示.



(a) DSG-SAC网络结构



(b) 系统框图

图1　算法的系统框架

与传统SAC算法不同,DSG-SAC的策略网络(actor)以智能体当前的观测状态为输入,不直接拟合智能体下一步的动作,而是先经过网络第1部分拟合出状态残差,经过状态反馈之后再作为策略网络的第2部分的输入,得到动作分布的期望与方差,经过采样生成下一个动作. 评价网络(critic)则是以智能体的观测状态和动作共同作为网络的输入,经过评价网络的第1部分拟合观测状态-动作对的残差,再输入网络第2部分估计一个Q函数值,用于评价当前策略的

好坏. 价值函数网络结构与Q函数网络结构一致,参数更新过程也一致. 策略网络和评价网络对状态与动作残差的拟合与估计动作分布和Q函数值过程一致,两个过程的参数更新在策略迭代过程中利用策略梯度同时更新. 经过状态奖励引导之后,算法的目标函数更新如下列各式所示:

$$J_Q(\theta)=E_{(s_d,a_d)\sim\mathcal{D}}\Big[\frac{1}{2}(Q_\theta(s_d,a_d)-\hat{Q}(s_d,a_d))^2\Big], \tag{15}$$

$$J_V(\psi)=E_{s_d\sim\mathcal{D}}\Big[\frac{1}{2}(V_\psi(s_d)-E_{a_d\sim\pi_\theta}(Q_\theta(s_d,a_d)-\log\pi_\phi(a_d|s_d)))^2\Big], \tag{16}$$

$$J_\pi(\phi)=E_{s_d\sim\mathcal{D},\epsilon_d\sim\mathcal{N}}[\log\pi_\phi(f_\phi(\epsilon_d,s_d)|s_d)-Q_\theta(s_d,f_\phi(\epsilon_d,s_d))]. \tag{17}$$

对3个修改后的目标函数进行梯度求导可以得出状态引导之后的策略梯度,最后在迭代到最大步数之后对参数进行软更新. DSG-SAC算法的伪代码如表1所示.

表 1 DSG-SAC算法伪代码

算法1 算法伪代码

input: 策略参数ϕ, Q函数参数θ_1和θ_2, 价值函数参数ψ, 经验池$\mathcal{D}$
initialize: target Q函数参数$\overline{\theta}_1\leftarrow\theta_1,\overline{\theta}_2\leftarrow\theta_2$
for 每个回合 do
  for 每个时间步 do
    状态奖励引导 $s_d\leftarrow s_t+\Delta(s)$
    根据当前状态引导选择动作 $a_{t+1}\sim\pi_\phi(a_t|s_d)$
    进入下一个状态 $s_{t+1}\sim p(s_{t+1}|s_d,a_{t+1})$
    将策略信息存入经验池中 $\mathcal{D}\sim\mathcal{D}(s_t,a_t,r(s_t,a_t),s_{t+1})$
  end for
  for 每个策略梯度 do
  更新参数
    $\theta_i\leftarrow\theta_i-\lambda_Q\hat{\nabla}_{\theta_i}J_Q(\theta_i)$, for $i\in1,2$
    $\phi\leftarrow\phi-\lambda_\pi\hat{\nabla}_\phi J_\pi(\phi)$
    $\overline{\theta}_i\leftarrow\tau\theta_i+(1-\tau)\overline{\theta}_i$, for $i\in1,2$
  end for
end for
output $\phi,\theta_1,\theta_2,\psi$

2.2 基于期望状态奖励引导的四足机器人步态学习

该四足机器人的每条腿有3个关节,分别为髋关节的偏航关节、俯仰关节和膝关节,每个关节通过电机输出力矩实现位置控制. 每条腿的足端设置有接触传感器,通过接触传感器的触发来判断四足机器人的状态是运动还是站立. 机器人的位姿和加速度由IMU来测量,同时设置距离传感器以测量机器人在运动过程中x轴、y轴、z轴上的位移. 表2是四足机器人的关键参数.

表 2 四足机器人的技术参数

参数	值
站立尺寸/mm	370×270×295
质量/kg	11
关节自由度/个	12
髋关节动作空间/(°)	$-46\sim46$
摆动关节动作空间/(°)	$-60\sim240$
膝关节动作空间/(°)	$-154.5\sim-52.5$

将改进后的SAC算法应用于四足机器人,其状态观测空间维度为37. 其中包含: 各个关节电机的角度和方向分别为12项,足端接触检测4项,距离传感器数据3项,IMU测得的位姿与加速度各3项. 动作空间包含12项,分别为各关节电机需要转过的角度. 将状态观测空间的37维数据作为策略网络的输入,预测四足机器人下一步的动作,其中两个隐含层分别有256个神经元,网络结构为37×256×256×37×1. 将状态观测空间和动作空间共同作为评价网络的输入,得出一个用于评价策略网络的Q值,两个隐含层神经元为256个神经元,网络结构为49×256×256×49×1,当机器人动作之后进入下一个状态,得到新的状态观测值作为输入来预测下一步动作与Q值,同时更新网络参数. 两个网络的激活函数均为Relu,使用Adam作为优化器并采用随机梯度下降更新参数. 实验的最终目标是让四足机器人学会以最高速度稳定行走,参考这个目标设置如下奖励函数[28-29]进行模型训练.

1) 躯干速度奖励r_t^v. 即

$$r_t^v=v_{\max}[1-\exp(-0.5v)], \tag{18}$$

其中$v_{\max}$和v分别为最大速度和当前速度.

2) 足端接触奖励r_t^c. 用于判断四足机器人行走过程中落地是否稳定,当四足机器人有两条以上的腿没有接触地面时,被认为行走不稳,有

$$r_t^c=-\max(n_{\text{loss}}-2,0), \tag{19}$$

其中n_{loss}为四足机器人足端未接触地面的数量.

3) 姿态奖励r_t^o. 控制四足机器人在行走过程中保持身体姿态稳定,即

$$r_t^o=1-\tan[3.75(r^2+p^2)], \tag{20}$$

其中r和p分别为四足机器人的偏航角和俯仰角. 当四足机器人位姿出现偏差、奖励变小、偏差过大时,奖励变为惩罚.

最后,将所有奖励根据重要性进行加权求和,以期望训练效果达到最佳. 整个奖励函数为R,即

$$R=k_vr_t^v+k_cr_t^c+k_or_t^o, \tag{21}$$

其中k_v、k_c和k_o分别为对应奖励的系数.

实验的整体流程框图如图1(b)所示.

3　实验结果与讨论

3.1　实验仿真与参数设计

本实验的仿真环境为pybullet框架, 使用pybullet模拟器加载四足机器人模型的urdf文件, 在pytorch框架下对所提出算法优化之后的四足机器人进行训练. 整个实验所用的计算机为搭载i7- 9750H的个人笔记本电脑, 无GPU加速.

为了验证所提出算法的性能, 本文从3个方面进行实验评估: 第1个实验(基于DSG-SAC算法的四足机器人步态学习过程仿真视频: https://www.bilibili.com/video/BV1q84y1C7Mj/)是使用该算法对四足机器人步态学习进行训练, 验证算法的可行性; 第2个实验分别改变关键参数的值, 探究该算法的最佳参数; 第3个实验(DSG-SAC算法与现有算法学习效果对比仿真视频: https://www.bilibili.com/video/BV1Gd4y1t7bK/)是将该算法与传统SAC算法进行对比, 验证算法的优越性.

3.2　机器人训练结果分析

与状态输入力矩输出这种端到端的控制方式不同, 本实验在训练过程中加入叠加机制, 在强化学习算法中, 以机器人当前状态空间和动作空间作为输入, 机器人各个关节的期望关节角度为输出, 再通过机器人逆运动学将关节角度转为每个关节电机的力矩, 最后, 由底层控制器控制末端执行器动作, 以提高训练效果. 在本实验中算法DSG-SAC的相关参数如表3所示.

表 3　DSG-SAC算法参数

参数	值
策略网络学习率lr_a	0.000 3
评价网络学习率lr_c	0.000 3
折扣因子γ	0.95
软更新系数τ	0.005
温度系数α	0.2
批量大小batchsize/个	256
回合最大步数max step/步	600
最大样本存储memorycap/个	1 000 000

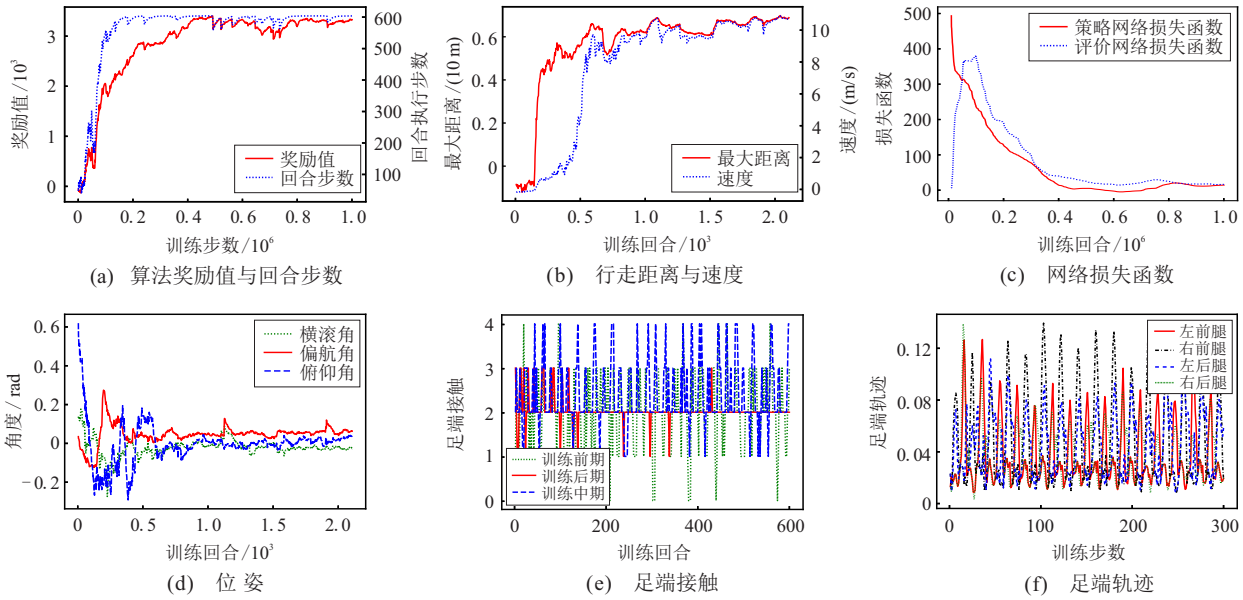


图 2　DSG-SAC算法在四足机器人环境中的训练结果

图2为该算法部署在四足机器人仿真环境中的训练结果.

图2(a)中实线为算法训练的奖励变化情况, 其横坐标为机器人训练的步数, 纵坐标为机器人训练过程中的奖励值. 如图2(a)所示, 四足机器人在训练初期得到的奖励为负值, 在训练步数为10 k～200 k期间奖励值快速上升, 训练步数达到400 k左右时趋于收敛, 奖励值逐渐稳定在3 300左右. 图2(a)中虚线展示了四足机器人在训练过程中每个回合的最大步数, 结合DSG-SAC算法的每个训练回合的终止条件可以得出结论: 四足机器人在刚开始训练时只能短时间内保持稳定, 之后便因摔倒导致每回合训练很快中止. 训练后期, 四足机器人每回合都能达到最大的训练步数, 产生稳定的步态, 达到理想的训练效果.

图2(b)中实线给出了四足机器人在训练过程中每回合行走的最大距离, 可以看出: 在训练初期, 四足机器人每回合走的距离为负值, 此时机器人由于未学习到策略而做出向后退的动作; 但随着训练的进

行,四足机器人的行走距离逐渐变长,最终当训练结果收敛时,机器人在600步内可以稳定行走10 m左右. 图2(b)中虚线为四足机器人在训练中每回合的平均速度,可以看出:在当训练初期,四足机器人速度为负值,四足机器人出现向后退的动作;但随着训练的进行,四足机器人的速度逐渐提高,最终达到最大速度0.7 m/s.

图2(c)展示了actor网络和critic网络参数更新过程的损失函数,可以发现随着训练步数的增加,两个损失函数逐渐减小,同样在时间步400 k时达到收敛.

图2(d)显示了四足机器人在训练过程中的位姿变化,从图中可以直观地看出:随着强化学习的进行,四足机器人在训练初期的3个俯仰角上下波动;训练中期角度逐渐稳定;最终收敛到0附近,此时代表四足机器人已经学习到稳定的步态.

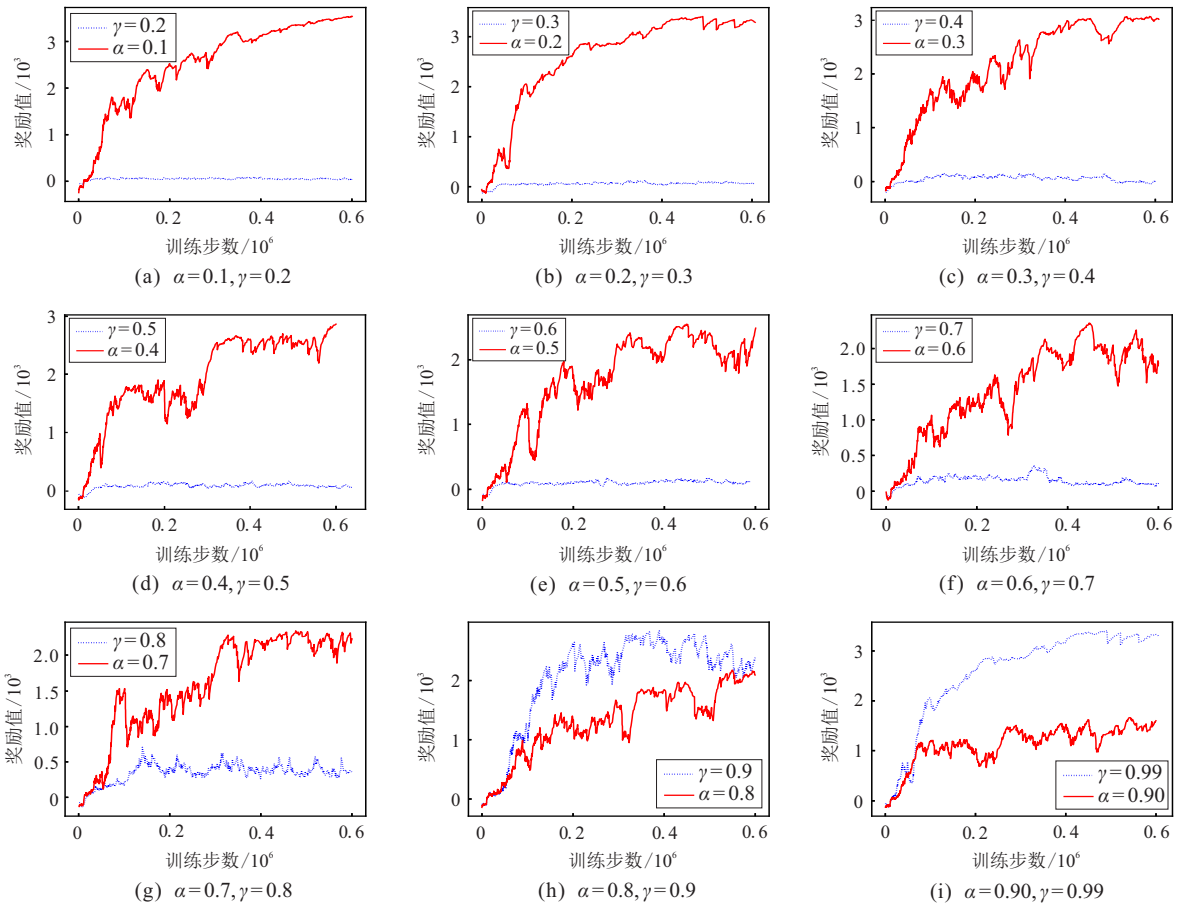
图2(e)为四足机器人在训练的3个时期中的足端接触图像,由图可得:在训练初期与中期,四足机器人的足端触地没有规律,步态较为杂乱;在训练后期行走过程中,忽略运动噪声,其足端接触稳定在数值2上,此时四足已经训练出对角小跑步态.

图2(f)为训练完成四足机器人的足端轨迹,从图中可以看出,此时其左前腿与右后腿、右前腿与左后腿的运动轨迹一致,表明四足机器人的步态为对角小跑步态. 因为足端轨迹的实验数据是通过脚步中心运动过程中的z轴位移来表示,这个过程中将足端抽象成一个质点,并非足端与地面接触点,所以四足机器人足端高度的最低点不为0. 另外,由于强化学习训练的随机性,最后得出的足端轨迹中,左右脚抬脚高度会存在2 cm～4 cm的误差,相对于整个四足机器人而言,误差在正常范围内,可以忽略不计.

4 结果讨论

4.1 超参数对比

SAC算法的优化目标如式(5)所示,其中奖励值的期望为未来时间的折扣奖励之和. 在优化目标中,α为策略网络熵的系数,其值决定了最大熵原理在SAC算法中的重要程度:当$\alpha = 0$时,SAC算法退化为普通策略梯度算法;当$\alpha = 1$时,SAC在每次对智能体进行动作估计时都要最大化考虑熵的影响. 因此,α的大小与算法性能密切相关. 为了探究α的大小对算法性能的影响,将α从0.1～0.9均匀取值进行对比实验. 在四足机器人运动技能学习中,当训练的时间步数达到500 k时奖励收敛,因此对比实验步数设为600 k,所有对比实验均设置$\gamma = 0.99$,实验结果如图3中红色实线所示. 从结果可以看出:



(a) $\alpha=0.1,\gamma=0.2$ (b) $\alpha=0.2,\gamma=0.3$ (c) $\alpha=0.3,\gamma=0.4$

(d) $\alpha=0.4,\gamma=0.5$ (e) $\alpha=0.5,\gamma=0.6$ (f) $\alpha=0.6,\gamma=0.7$

(g) $\alpha=0.7,\gamma=0.8$ (h) $\alpha=0.8,\gamma=0.9$ (i) $\alpha=0.90,\gamma=0.99$

图 3 算法关键超参数对比

当α为0.1与0.2时, 训练结束奖励达到3 500左右; 随着α的继续增大, 奖励值逐渐减小, 奖励曲线波动逐渐增大, 达不到收敛条件. 因此, 可以得出α取值为0.1～0.2时, 训练效果最好.

在折扣奖励中, 折扣因子γ是SAC算法中确定奖励类型的另一个重要参数, 它定义了未来奖励的重要性. $\gamma = 0$意味着只考虑当前奖励, $\gamma = 1$意味着更重视长期奖励. 为了探究γ的大小对训练效果的影响, 本实验将γ从0.2～0.99均匀取值进行对比实验, 实验步数设为600 k, 所有对比实验均设置$\alpha = 0.2$, 实验结果如图3中蓝色虚线所示. 从结果可以看出: 当γ取值较小时, 训练曲线上下波动, 奖励值较低; γ继续增大至0.9时, 奖励值逐渐提高; 当$\gamma = 0.99$时, 训练效果最佳.

4.2　对比实验

为了验证本文提出的DSG-SAC算法的优越性, 本节设计对比实验, 将DSG-SAC与传统的SAC、近端策略优化(PPO)、 深度确定性策略梯度(DDPG)进行比较. 上述4种算法执行相同的任务, 即训练四足机器人学习行走步态, 基本超参数取相同值, 其训练结果如图4所示. 从结果可以看出: 当训练到大约15 k步时, DSG-SAC算法的奖励值继续保持上升; 当训练步数达到24 k时, DSG-SAC算法奖励值的总体趋势继续上升, 并且上升到3 000左右. 而SAC算法在训练到15 k步时, 其奖励值趋于平缓; 当训练步数达到24 k时, 奖励值来回波动, 总体趋势保持不变, 最终收敛于2 300左右. 在PPO算法的仿真实验中, 因具有先验知识, 故设置一定的初始奖励, 随着训练步数的增加, 奖励值逐渐增加到3 100左右; 当训练步数超过100 k步时, 奖励逐渐减少, 最终收敛到2 500步. 在DDPG算法中, 虽然奖励值逐渐上升, 但其上升速度远小于DSG-SAC和SAC算法, 整个训练过程逐渐收敛到1 M步左右. 综上所述, 本文所提出的算法能够在前期减小训练产生的波动, 中期训练速度较快, 最终奖励值更高, 算法性能整体优于其他3种强化学习算法.

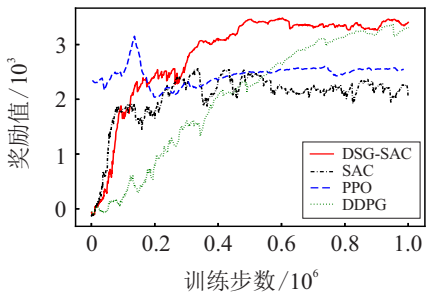


图 4　不同算法训练效果对比

5　结　论

具备学习能力是高等生物的典型特征和显著优势, 因为有了学习能力, 高等生物可以在与环境交互过程中不断进化自身行为. 为了使四足机器人具备学习能力, 本文受四足动物运动技能形成机理的启发, 使用SAC算法模拟其运动技能从零开始学习的过程. 为了提高SAC算法在四足机器人步态学习任务的学习效果, 本文提出了一种基于期望状态奖励引导的SAC算法, 在pybullet环境中对四足机器人进行仿真. 从实验结果可以看出, 训练最初四足机器人步态不稳定, 无法稳定行走, 到后期可以学习到标准的对角小跑步态, 4条腿的运动轨迹呈现节律信号. 实验表明: 所提出算法可以使四足机器人步态学习任务的奖励快速收敛, 使其快速学会对角小跑步态; 验证了算法中关键超参数对算法性能的影响, 得出算法效果最优的超参数范围; 与现有的强化学习算法进行对比, 得出所提出算法能够有效减小训练过程中的误差, 加快收敛速度, 并且有更高的奖励值, 验证了DSG-SAC算法性能整体优于传统强化学习算法. 在后续研究中将引入地形因素, 使四足机器人能够完成在复杂地形下的步态学习, 同时计划在实物机器人中部署所提出的算法, 完成实物实验.

参考文献(References)

[1] 贾庆轩, 袁博楠, 陈钢, 等. 关节锁定空间机械臂负载操作能力评估与轨迹规划[J]. 控制与决策, 2020, 35(1): 243-249.
(Jia Q X, Yuan B N, Chen G, et al. Load carrying capacity evaluation and task trajectory planning of space manipulator with the locked joint[J]. Control and Decision, 2020, 35(1): 243-249.)

[2] 尤波, 曲伟健, 李佳钰. 面向双操作者的六足机器人共享遥操作[J]. 控制与决策, 2022, 37(11): 2769-2778.
(You B, Qu W J, Li J Y. Shared teleoperation of hexapod robot for dual operators[J]. Control and Decision, 2022, 37(11): 2769-2778.)

[3] 郭非, 汪首坤, 王军政. 轮足复合移动机器人运动规划发展现状及关键技术分析[J]. 控制与决策, 2022, 37(6): 1433-1444.
(Guo F, Wang S K, Wang J Z. Development status and key technology analysis for motion planning of wheel-legged hybrid mobile robot[J]. Control and Decision, 2022, 37(6): 1433-1444.)

[4] McGhee R B. Finite state control of quadruped locomotion[J]. Simulation, 1967, 9(3): 135-140.

[5] Di Carlo J, Wensing P M, Katz B, et al. Dynamic locomotion in the MIT cheetah 3 through convex model-predictive control[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, 2019: 1-9.

[6] Bledt G, Powell M J, Katz B, et al. MIT cheetah 3: Design and control of a robust, dynamic quadruped robot[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, 2019: 2245-2252.

[7] Lee J, Hwangbo J, Wellhausen L, et al. Learning quadrupedal locomotion over challenging terrain[J]. Science Robotics, 2020, 5(47): eabc5986.

[8] Shi H J, Zhou B, Zeng H S, et al. Reinforcement learning with evolutionary trajectory generator: A general approach for quadrupedal locomotion[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3085-3092.

[9] 韩连强, 陈学超, 余张国, 等. 面向离散地形的欠驱动双足机器人平衡控制方法[J]. 自动化学报, 2022, 48(9): 2164-2174.
(Han L Q, Chen X C, Yu Z G, et al. Balance control of underactuated biped robot for discrete terrain[J]. Acta Automatica Sinica, 2022, 48(9): 2164-2174.)

[10] Kim J, Ba D X, Yeom H, et al. Gait optimization of a quadruped robot using evolutionary computation[J]. Journal of Bionic Engineering, 2021, 18(2): 306-318.

[11] Fahmy T A, Maged S A. Teaching quadruped to walk using fault adaptive deep reinforcement learning algorithm[C]. 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference. Cairo, 2021: 129-134.

[12] Yin F L, Tang A N, Xu L W, et al. Run like a dog: Learning based whole-body control framework for quadruped gait style transfer[C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, 2021: 8508-8514.

[13] Sheng J P, Chen Y Y, Fang X, et al. Bio-inspired rhythmic locomotion for quadruped robots[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6782-6789.

[14] Gao Y, Li Y K, Guo Z Q. A Q-learning based UAV path planning method with awareness of risk avoidance[C]. 2021 China Automation Congress. Beijing, 2022: 669-673.

[15] Zou Q J, Liu S H, Zhang Y, et al. Rapidly-exploring random tree algorithm for path re-planning based on reinforcement learning under the peculiar environment[J]. Control Theory & Applications, 2020, 37(8): 1737-1748.

[16] Liu Y H, Xu Y Z. Free gait planning of hexapod robot based on improved DQN algorithm[C]. 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT). Weihai, 2021: 488-491.

[17] Wang M C, Ruan X G, Zhu X Q. Heuristic gait learning of quadruped robot based on deep deterministic policy gradient algorithm[C]. 2020 Chinese Automation Congress. Shanghai, 2020: 1046-1049.

[18] Zhang W Y, Jiang Y C, Din Farrukh F U, et al. A portable accelerator of proximal policy optimization for robots[C]. 2021 IEEE International Conference on Integrated Circuits, Technologies and Applications. Zhuhai, 2021: 171-172.

[19] Dong H, Yang J, Li S B, et al. Research progress of robot motion control based on deep reinforcement learning[J]. Control and Decision, 2022, 37(2): 278-292.

[20] Zhao M H, Zhang X B, Guo X, et al. A general assembly sequence planning algorithm based on hierarchical reinforcement learning[J]. Control and Decision, 2022, 37(4): 861-870.

[21] Shao Y C, Jin Y B, Liu X W, et al. Learning free gait transition for quadruped robots via phase-guided controller[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 1230-1237.

[22] Rudin N, Kolvenbach H, Tsounis V, et al. Cat-like jumping and landing of legged robots in low gravity using deep reinforcement learning[J]. IEEE Transactions on Robotics, 2022, 38(1): 317-328.

[23] Sun H H, Hu C H, Zhang J G. Deep reinforcement learning for motion planning of mobile robots[J]. Control and Decision, 2021, 36(6): 1281-1292.

[24] Wong C C, Chien S Y, Feng H M, et al. Motion planning for dual-arm robot based on soft actor-critic[J]. IEEE Access, 2021, 9: 26871-26885.

[25] Tang H L, Wang A Q, Xue F, et al. A novel hierarchical soft actor-critic algorithm for multi-logistics robots task allocation[J]. IEEE Access, 2021, 9: 42568-42582.

[26] Sharma K, Singh B, Herman E, et al. Maximum information measure policies in reinforcement learning with deep energy-based model[C]. 2021 International Conference on Computational Intelligence and Knowledge Economy. Dubai, 2021: 19-24.

[27] Yu C, Rosendo A. Multi-modal legged locomotion framework with automated residual reinforcement learning[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10312-10319.

[28] Tsounis V, Alge M, Lee J, et al. DeepGait: Planning and control of quadrupedal gaits using deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3699-3706.

[29] Shi H J, Zhou B, Zeng H S, et al. Reinforcement learning with evolutionary trajectory generator: A general approach for quadrupedal locomotion[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3085-3092.

作者简介

张思远 (1996–), 男, 硕士生, 从事强化学习、机器人技能学习等研究, E-mail: zhsy@emails.bjut.edu.cn;

朱晓庆 (1987–), 男, 讲师, 博士, 从事人工智能与机器人等研究, E-mail: alex.zhuxq@bjut.edu.cn;

阮晓钢 (1958–), 男, 教授, 博士生导师, 从事人工智能与机器人等研究, E-mail: adrxgbjut@163.com;

李春阳 (1997–), 男, 硕士生, 从事机器人步态学习的研究, E-mail: 1378272571@qq.com;

刘鑫源 (1999–), 男, 硕士生, 从事机器人步态学习的研究, E-mail: 1063022135@qq.com.