中国科学院软件研究所学术年会’2019
暨计算机科学国家重点实验室开放周

学术论文

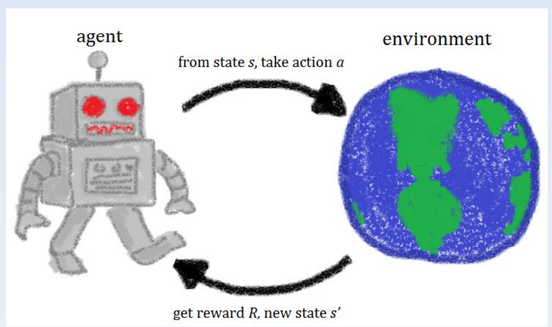
# Multi-Critic DDPG Method and Double Experience Replay
# 多critic的DDPG算法和双经验池结构

吴蛟* 王瑞 李瑞英 张慧 胡晓惠

*通讯方式: 15611537585 | wujiao2016@iscas.ac.cn

## 背景

强化学习方法(Deep Reinforcement Learning, DRL)在人工智能和自动控制等领域得到了广泛的关注和研究。在DRL过程中，智能体与环境进行交互来尽可能的获得更多的累计奖励，通常可建模成马尔科夫决策过程。




- 连续决策过程
- 无监督，只有奖励信号反馈
- 反馈通常是延迟的
- 数据相关，并不独立

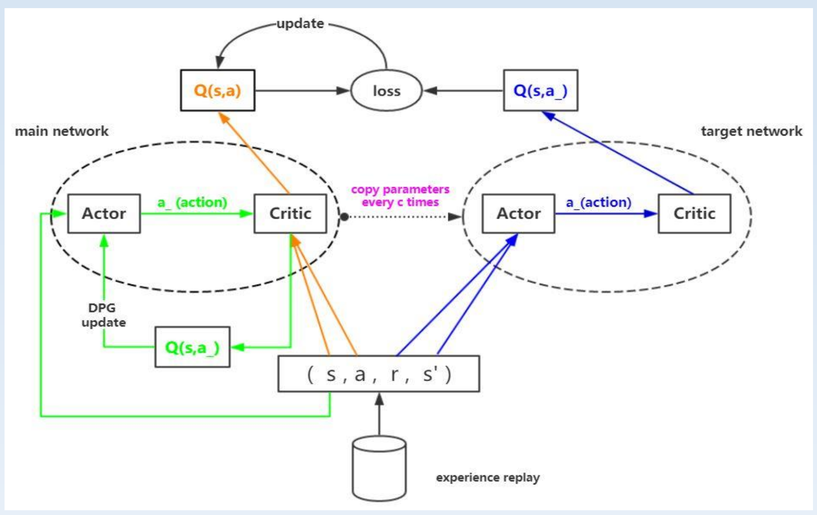
## DDPG算法

(Deep Deterministic Policy Gradient)

确定性动作 $a = \mu(s|\omega)$

确定性策略梯度：

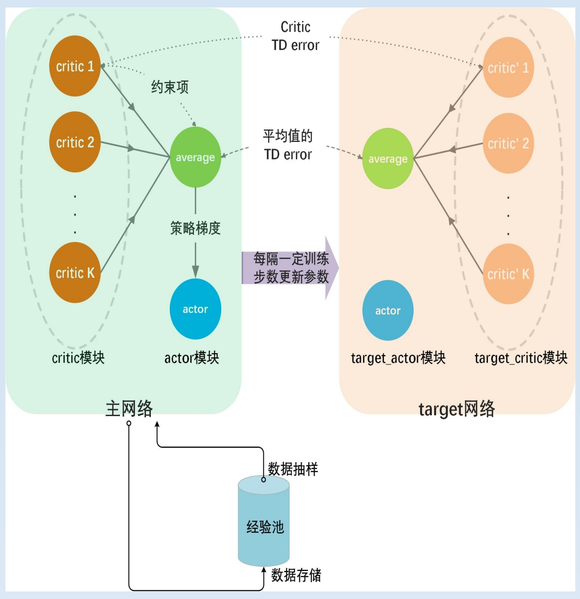
$$\nabla_a Q(s,a) = \nabla_a Q(s,a|\theta)|_{a=\mu(s|\omega)} \nabla_\omega \mu(s|\omega)$$




## 创新点

### ① 多critic的DDPG算法

多个独立且不同的critic的平均值代替DDPG算法中的Q值，提高训练过程的稳定性和智能体的表现。




误差：

$$L(\theta_i) = \left(r(s_t,a_t) + \gamma Q'(s_{t+1},a_{t+1}|\theta_i^-) - Q(s_t,a_t|\theta_i)\right)^2$$

$$L_{avg}(\theta) = \left(r(s,a) + \gamma Q_{avg}^{target}(s,a|\theta^-) - Q_{avg}(s,a|\theta)\right)^2$$

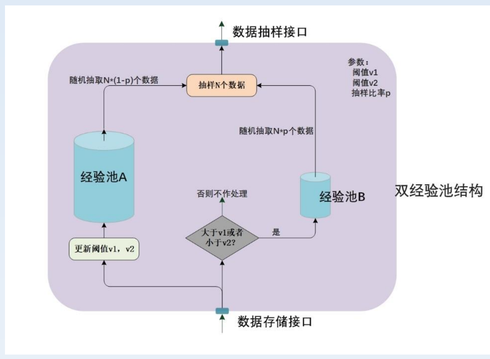
$$C(\theta_i) = \left(Q_i(s,a|\theta_i) - Q_{avg}(s,a|\theta)\right)^2$$

损失函数：

$$Loss(\theta_i) = \tau L(\theta_i) + (1-\tau) L_{avg}(\theta) + \beta C(\theta_i),$$

$$\tau, \beta \in (0,1)$$

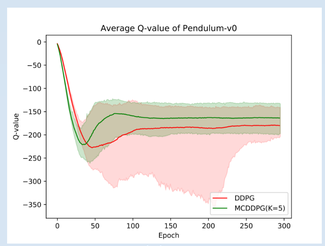
### ② 双经验池结构



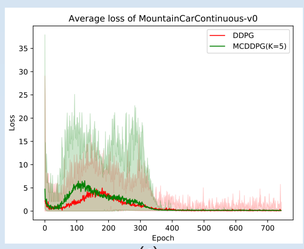

双经验池结构

- 额外的经验池用来保存特别好或特别差的经验信息。
- 抽样时依照比例，分别从两个经验池抽取一批数据。
- 目的在于加快收敛过程。

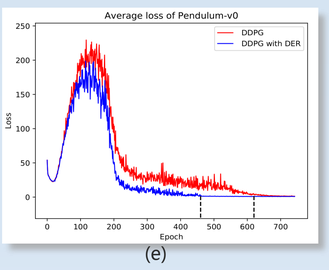
## 实验结果



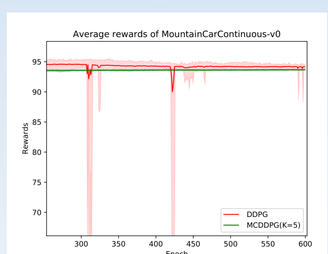

(a)



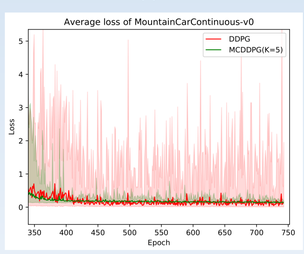

(c)




(e)




(b)




(d)

图(a)表明更高更稳定的Q值；图(b)表明完美的波动性对比；图(c)和图(d)表明损失函数的收敛范围更小更稳定；图(e)表明收敛过程的加速。

## 总结与展望

- 证明多critic的DDPG算法具有更好的稳定性和性能提升。
- 双经验池结构的加速收敛作用得到验证。
- 研究展望：超参数 $\tau, \beta$ 随着训练自动调整；经验数据的汰换方式的改进。

2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC2018)