A Walk in the Park: Learning to Walk in 20 Minutes With Model-Free Reinforcement Learning —— hoho

论文试图解决什么问题？

如何更加高效使用model-free 强化学习方法进行训练？本文试图用训练四足机器人行走的实验进行说明，解决计算低效，采样低效，训练时间长的问题。

（how efficiently can we implement fully model-free deep RL algorithms? Our goal is to train a robot to walk in the real world as efficiently as possible, and efficiency includes computational complexity, sample complexity, and total wall-clock time）

这是否是一个新的问题？

不是。解决强化学习训练效率一直是业界的难题。

这篇文章要验证一个什么科学假设？

作者认为当前的很多方法训练不高效不在于选的模型不对，而在于没有仔细设计算法的实现。

导致一个重要问题是agent进行策略提升的效率跟agent与环境交互的效率之间有延迟。

（add a delay between the agent interacting with the environment and learning from the samples, which slows down training.）

作者认为精心设计正则化方法有助于改善这种训练不高效的问题。

(These results suggest that the key is not any one specific critical choice, but the general prniciple of augmenting actor-critic RL with regularization or normalization.)

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

一些增加了正则化方法的强化学习模型：

譬如REDQ，利用集成了大量的critic网络来最小化的它们当中的子集作为正则化方法；

（by utilizing a large ensemble of critics and computing target values by minimizing over a random subset of them）

还有DroQ，使用dropout和layer normalization，来提高数据的利用率（？）

（allows for a higher update to data ratio by regularizing the critic networks with dropout and layer normalization）

论文中提到的解决方案之关键是什么？

本文使用的实时同步的训练方式（our choice of algorithm and implementation is aimed at enabling real-time synchronous training.）

主要基于REDQ进行的改进。

一些定义：

- 状态空间：包括机器人行进的方向，角速度，线速度，关节的角度，上一时刻的动作空间等

(contain the root orientation, root angular velocity, root linear velocity, joint angles, joint velocities, binary foot contacts, and previous action.)

- 动作空间：每只脚关节上的角度

（each of the 12 joints for every leg as [p - o, p + o]， where p corresponds to default motor angles and o is an action offset）

- 奖励函数：

$$r(s,a) = r_v(s,a) - 0.1v_{yaw}^2$$

v_{yaw}是angular yaw velocity

$$r_v(s,a)\begin{cases}1 & \text{if } v_x \in [v_t, 2v_t] \\ 0 & \text{if } v_x \in (-\infty, -v_t] \cup [4v_t, \infty] \\ 1 - \frac{|v_x - v_t|}{2v_t} & \text{otherwise.}\end{cases}$$

v_t是目标速度（targe velocity）

- 使用基于REDQ的实现方式，对其中SAC的方法进行改进，将UTD增加到20

1. UTD 比率 (Update-To-Data ratio) ： 智能体进行更新的轮数（gradient step, network parameter updating step）和实际与环境交互轮数(data collection step)的比值，它是衡量算法采样效率的重要指标。基于模型的 sota 算法，例如 MBPO 拥有 20-40 的 UTD 值，而无模型的算法，例如 SAC 的 UTD 值仅为 1 。理论上，在无模型的方法中，通过直接增大 UTD 比率也可以达到与基于模型的方法相近的高采样效率，但是实际中，直接采取这种做法会导致价值函数的估计不准确，严重影响算法的训练稳定性及表现。

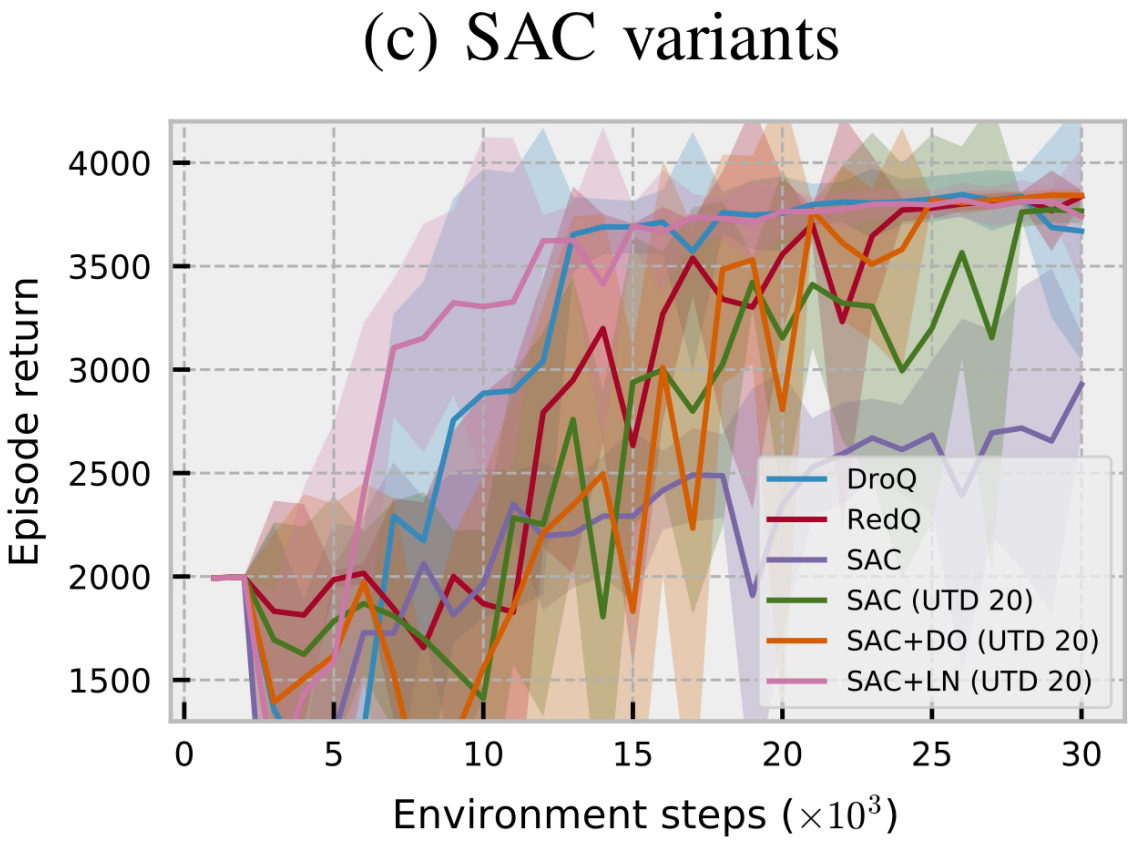
后来发现使用DroQ可以通过layer normalization和dropout的组合，达到REDQ的几乎相似的效果，而且计算效率更加高，于是最终使用了DroQ的变体作为四足机器人在真实环境中的训练方法。

（As noted in DroQ [60], REDQ is more computationally expensive due to the large ensemble, and regularizing the critic with a combination of layer normalization and dropout can lead to similar benefits at lower compute cost (blue)）

论文中的实验是如何设计的？

1. 模拟环境 MuJoCo
2. 真实环境：在五种不同的地面进行试验

进行了多种消融实验（ablation experiments）：



用于定量评估的数据集是什么？代码有没有开源？

- 没数据集
- 代码可参考：https://github.com/ikostrikov/walk_in_the_park

论文中的实验及结果有没有很好地支持需要验证的科学假设？

实验验证了：在训练中加入正则化的方法，都是同样的提高训练的效率。重要不是是否使用正则化方法本身（应用本文也单独将layer normalization和dropout分开进行单独验证，发现使用LN的效果要好于dropout），而是是否适当的使用了正则化方法使得SAC能够高效的搭配高的UTD比率进行使用。

（conclude that a variety of regularization or normalization methods, if implemented and applied carefully, can all achieve a similar level of improvement in performance over their underlying algorithm in our setup. That is, the important thing is not any

single specific regularization technique, but the use of any suitable regularization so as to enable SAC to effectively use higher UTD ratios.）

这篇论文到底有什么贡献？

验证了强化学习中加入正则化方法可以有效提高学习效率。

下一步呢？有什么工作可以继续深入？

hoho_todo