RANDOMIZED ENSEMBLED DOUBLE Q-LEARNING: LEARNING FAST WITHOUT A MODEL —— hoho

论文试图解决什么问题？

model-free RL中Sample efficiency问题

这是否是一个新的问题？

在model-free的RL方法中，可能是一个新的问题。作者说使用类似的方法可以媲美model-based的SOTA模型。

这篇文章要验证一个什么科学假设？

在高UTD的情况下，综合使用emsemble Q function和in-target minimization可以降低Q functions训练的偏差的方差到0。

另：UTD: Update-To-Data，$UTD = \frac{\text{智能体网络更新的次数}}{\text{智能体与环境交互的次数}}$

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？

hoho_todo

论文中提到的解决方案之关键是什么？

1. 使用高的UTD， UTD >> 1

本文使用UTD = 20，即G = 20

2. ensemble of Q functions

集成 N 个 Q 网络，本文使用N = 10

3. in-target miniization across a random subet of Q funcdtion from the ensemble

从N个ensemble of Q functions中随机采样M个Q function，计算其中的最小值，作为贝尔曼方差目标Q值的近似。

本文使用M = 2

总体算法过程如下：

Algorithm 1 Randomized Ensembled Double Q-learning (REDQ)

Initialize policy parameters θ, N Q-function parameters ϕ_i, $i = 1, \ldots, N$, empty replay buffer $\mathcal{D}$. Set target parameters $\phi_{\text{targ},i} \leftarrow \phi_i$, for $i = 1, 2, \ldots, N$
repeat
  Take one action $a_t \sim \pi_\theta(\cdot|s_t)$. Observe reward r_t, new state s_{t+1}.
  Add data to buffer: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1})\}$
  for G updates **do**
    Sample a mini-batch $B = \{(s, a, r, s')\}$ from $\mathcal{D}$
    Sample a set $\mathcal{M}$ of M distinct indices from $\{1, 2, \ldots, N\}$
    Compute the Q target y (same for all of the N Q-functions):

$$y = r + \gamma \left(\min_{i \in \mathcal{M}} Q_{\phi_{\text{targ},i}}\left(s', \tilde{a}'\right) - \alpha \log \pi_\theta\left(\tilde{a}' \mid s'\right) \right), \quad \tilde{a}' \sim \pi_\theta\left(\cdot \mid s'\right)$$

    for $i = 1, \ldots, N$ **do**
      Update ϕ_i with gradient descent using

$$\nabla_\phi \frac{1}{|B|} \sum_{(s,a,r,s') \in B} \left(Q_{\phi_i}(s, a) - y\right)^2$$

    Update target networks with $\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho)\phi_i$
  Update policy parameters θ with gradient ascent using

$$\nabla_\theta \frac{1}{|B|} \sum_{s \in B} \left(\frac{1}{N} \sum_{i=1}^{N} Q_{\phi_i}\left(s, \tilde{a}_\theta(s)\right) - \alpha \log \pi_\theta\left(\tilde{a}_\theta(s) | s\right) \right), \quad \tilde{a}_\theta(s) \sim \pi_\theta(\cdot \mid s)$$

本文提出的REDQ方法是基于SAC进行改造的，原则上可以使用在各种model-free RL方法上。

论文中的实验是如何设计的？

使用MuJoCo实验环境，具体分别有Hopper、Walker2d、Ant和Humanoid

主要进行两类实验：

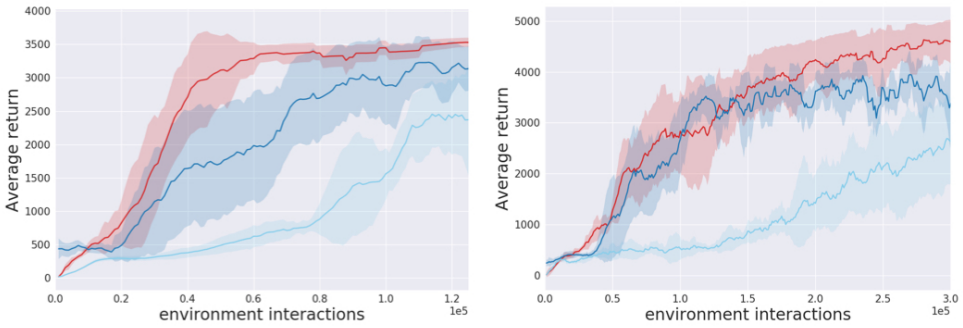
1. 与其他模型的对比：SAC-20（基于SAC，但UTD=20），MBPO（model-based）
2. 消融实验：改变参数N，M的对比

用于定量评估的数据集是什么？代码有没有开源？

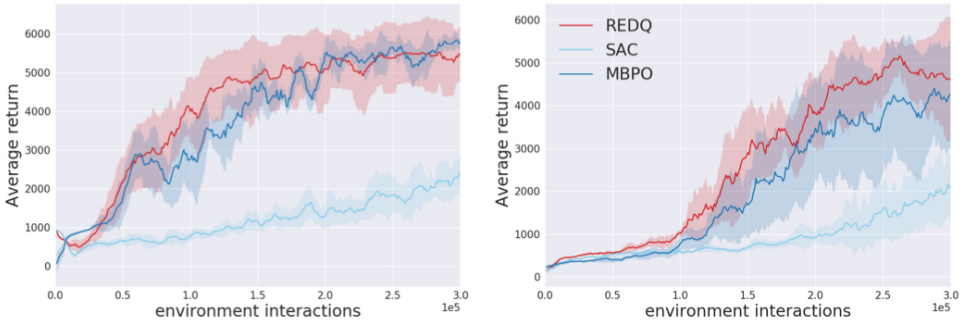
1. 没有数据集
2. 代码：https://github.com/watchernyu/REDQ

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？

实验表明REDQ能达到MBPO的性能，比SAC-20占优。



(a) Hopper (b) Walker2d



(c) Ant (d) Humanoid

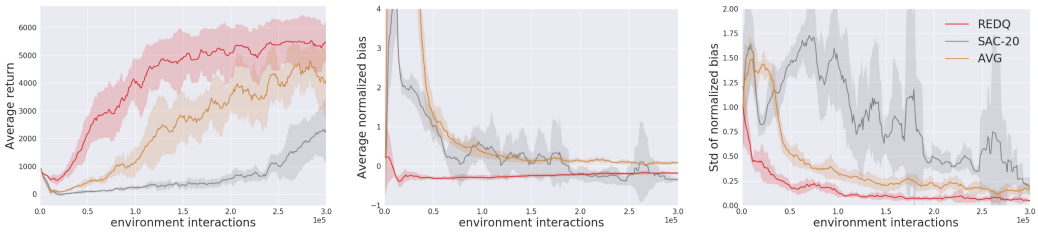
作者还分析了为什么REDQ比SAC-20性能要好的原因。

首先定义模型与groud true $Q^{\pi}(s,a)$的偏差 $bia = Q_{\phi}(s,a) - Q^{\pi}(s,a)$

其中$Q^{\pi}(s,a)$用ensemble的每个Q函数$Q_{\phi_i}(s,a)$的平均来估计。

将这个偏差做归一化：$\frac{Q_{\phi}(s,a)-Q^{\pi}(s,a)}{E_{\overline{s},\overline{a}\sim\pi}(\overline{s},\overline{a})}$，接着使用Monte Carlo方法来考察偏差bia的均值与方差。

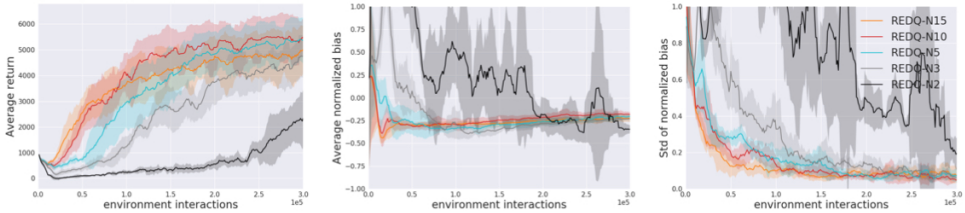
结果如下：



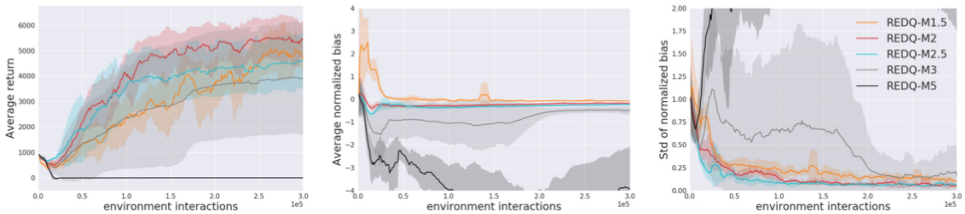
(a) Performance, Ant (b) Average bias, Ant (c) Std of bias, Ant

发现REDQ产生的偏差bia的均值与方差是最小的，虽然SAC也是使用较高的UTD，但其产生的偏差的方差较高，使得网络效果不稳定，由此证明了REDQ的可靠性。（上图AVG是使用全部ensemble of Q functions计算Q值，并不是随机采样其中的几个Q functions）

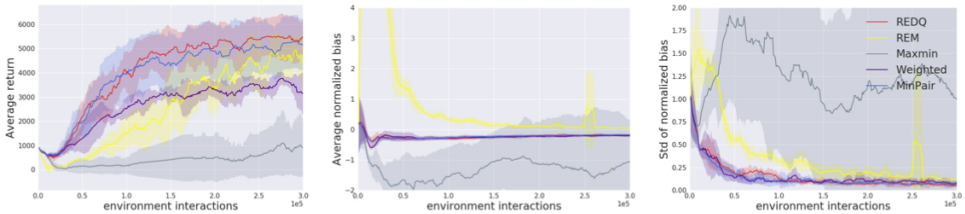
作者还通过N与M的改变对比验证了各种REDQ的效果：



(a) Performance, Ant (b) Average bias, Ant (c) std of bias, Ant



(d) Performance, Ant (e) Average Q bias, Ant (f) std of Q Bias, Ant



(g) Performance, Ant (h) Average bias, Ant (i) std of bias, Ant

综合来看，N=10， M=2效果较好。

（M=1.5为有一半的概率为M=1，另一半的概率为M=2）

这篇论文到底有什么贡献？

1. 提出了一种实现比较简单的model-free的RL sample efficient方法，其性能可媲美model-base SOTA RL方法
2. 仔细验证了在高UTD情况下REDQ为何可以胜过其他model-free RL方法
3. 与OFE结合（online feature extract），显示了在某些场景下REDQ-OFE可以训练得相当快。

下一步呢？有什么工作可以继续深入？

与representation learning结合，通过online feature extractor network， 从环境数据中学习表示向量，或许可以进一步提高REDQ的性能。