## 推送文章标题：一文七问 | ICLR’16 ：利用深度强化学习进行连续动作控制——DDPG

论文简介

标题：CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING

荣誉：

作者：Timothy P. Lillicrap , Jonathan J. Hunt , Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver & Daan Wierstra

单位：Google DeepMind（全挂在了google下面）

**文章简介+论文摘要（150字）**

本篇推文将为大家介绍DeepMind团队于 2016 年在人工智能领域顶级会议 ICLR上发表的一篇论文：Continuous Control with Deep Reinforcement Learning。该论文介绍了一种用于解决连续动作空间的深度强化学习方法。具体为：基于DQN与DPG的思想，利用深度网络对高维连续动作策略进行逼近，构成一种无模型的Actor-Critic结构的off-policy算法。本文同时加入了软更新、经验回放和批标准化的技巧，用于提高训练稳定性。

**以下每个问题200-250字左右**  
**1. What is the motivation? 这篇论文的研究动机是什么？（研究背景）**

现实世界的问题通常有复杂、高维的特点。人们希望利用人工智能可以直接利用原始的高维感觉输入来处理现实世界的复杂问题，从而Deep Q Network(DQN)的方法被提出。该算法以智能体的观测值作为输入，利用深度神经网络来拟合动作价值估计函数。该方法可以以未经处理的高维数据作为输入，而且在许多游戏环境上表现出超越人类的水平。但是由于DQN的动作决策依赖于寻找使得动作价值函数最大的动作，它没有办法很好的适应连续动作空间。一种直观的解决方法是将动作空间离散化，但是这会带来严重的问题：随着操作自由度的增加，动作空间维度随自由度指数增长；并且在控制精度要求较高的场景这一问题更加严重 ，因为采样间隔减小会导致动作空间维度成比例增长；同时离散化的方法抛弃了动作域结构上的信息，会对解决问题不利。DDPG着眼于解决DQN无法处理连续动作的问题。

在DDPG之前，DPG也是着眼于处理连续动作空间问题的算法。其利用确定性策略代替随机策略，减小了动作价值估计网络的规模，为解决连续动作问题提供了可行的方向。

**2. What is the problem the paper wants to solve? 这篇论文主要解决了什么问题？（问题描述，前提假设）**

本文基于DQN与DPG算法，提出了一种新的、用于解决连续动作空间问题的算法，利用一些技巧解决了训练中一些不稳定问题。

DDPG的思想承DPG而来。DPG介绍了随机测略与确定策略两种策略类型，指出当策略动作的方差为零时即成为确定策略，并证明了确定策略下奖励对策略网络参数的可微性，使得奖励的期望只是状态的积分，提高了效率。但是DPG对动作估值网络有一定的要求，使用的不是神经网络拟合，而是线性函数。从而导致面对状态空间很大的场景表现欠佳。DDPG采用神经网络对动作价值函数进行拟合，指出其对复杂状态空间的重要性。

为了实现off-policy更新，增强训练鲁棒性，DDPG算法引入了DQN的思想：经验回放与软更新。这两个方法目的都是为了提高训练的稳定性。在获取数据的过程中，前后数据具有极大的时空相关性，这对深度网络的学习是十分不利的，因此使用经验回放法，每次随机抽取mini batch，以减轻数据相关性。另一方面，由于神经网络拟合动作价值函数不易拟合的特点，引入了软更新的方法。在DDPG中，同时对策略网络和动作价值估计网络采用了软更新，减缓更新速度，提高了训练的稳定性。

由于策略网络直接输出动作值，存在现实中物理单位不统一的问题，对训练造成一些难度。所以还采用了批标准化的方法，统一期望和方差。

最后，算法在训练过程中回为动作值增加噪声，以提高探索能力，避免次优策略。

**3. What is their approach? 这篇论文的方法是什么？（所提算法）**

该论文提出的算法如图 1所示，首先算法会初始化四个网络的参数与经验库，其中策略网络和估值网络各有两组，每组的初始参数是相同的。算法首先利用有噪声的动作进行探索，按照的方式选择动作，并将一组transition：填充到经验库中。当经验库中有足够的数据时再开始训练。其中动作估值网络的更新方式和DQN类似，利用贝尔曼方程和动作价值目标网络计算Q值得最佳估计，再用动作价值估计网络去进行拟合，做MSELoss

，同时优化参数。

策略网络方面，利用确定策略动作梯度进行随机梯度上升进行更新：。为了保证寻来训练的稳定性，目标网络始终不参与实际动作决策。

最后，设置更新幅度，对两个目标网络的参数进行更新

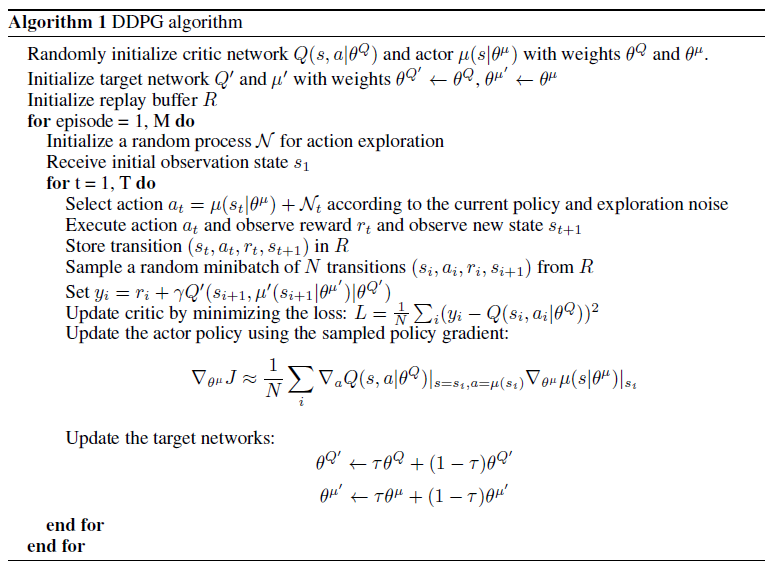


图1 本文提出的DDPG算法

**4. What are the key conclusions and key results? (e.g. SOTA? new bound? etc) 这篇论文的关键结果是什么？（实验结果，论文结论）**

为了证明算法的有效性和稳定性，论文用相同的网络结构与参数设置，在20多个环境下进行了实验。以动作空间平均采样作为下基线，以iLQG方法作为上基线，发现在一些场景下DDPG的得分可以超过iLQG方法，即便使用的是原始像素作为输入。

为了证明DDPG相较于DPG改进点的作用，论文在10个环境下进行了对照试验，如图2所示。加入了批标准化的DPG算法效果由浅灰色线所示，加入了目标网络的DPG算法效果由深灰色线所示，同时加入了目标网络与批标准化的算法效果由绿色线所示，使用原始像素作为输入的算法效果由蓝色线所示。其中绿线，也即DDPG的效果最好，证明了目标网络的重要性。

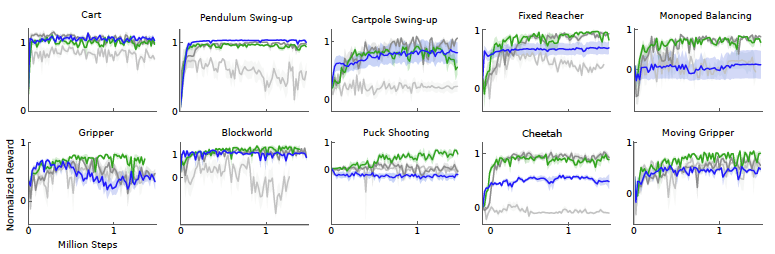


图2 DDPG实验表现

同时，论文对倒立摆、车杆和Cheetah三种环境下的动作价值估计值与真实值进行了比较，如图3所示。发现在简单环境下，可以准确估计真实值，但是复杂环境却出现了明显的过估计。相关解决办法在DDQN、TD3等方法中有所讨论。

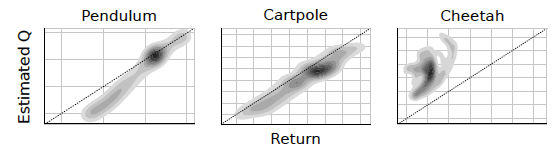


图3 DDPG的Q估计函数准确度密度图

**5. What is the novelty? 这篇论文的创新点在哪里？（论文贡献）**

该论文以解决连续动作空间问题为起点，以DPG的确定动作梯度更新为基础，结合了DQN中软更新与经验回放的技巧，，比较有效的解决了连续动作空间问题，并保证了训练的稳定性，解决了收敛难的问题。应用较多，成为一种比较成熟的Actor-Critic算法。

**6. Any related work worth of following? 有什么值得阅读的相关文献吗？（相关工作）**

1. 作为该论文基础的 DQN算法

Mnih V , Kavukcuoglu K , Silver D , et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science, 2013.

2. 作为该论文基础的DPG算法

Silver D , Lever G , Heess N , et al. Deterministic Policy Gradient Algorithms. JMLR.org, 2014.

**7. Any comment? 任何评价？**

该论文对已有的确定性策略方法进行了改进，解决了连续动作空间问题，并且相较于随即动作策略网络规模更小，训练更加容易。同时，采用经验回放与软更新的策略，使得动作价值估计网络更容易收敛，也成为后续算法中的基本操作步骤。但是文中也指出，虽然效果好于随机梯度方法，但Q值过估计的情况确实存在，为后续解决Q过估计的方法进行铺垫。

**作者简介：**

陈烁，北京航空航天大学2021届自动化专业本科毕业生，中科院自动化所实习生。兴趣方向为强化学习。

联系邮箱：labyrinthcs@outlook.com