
图 1 Atari 游戏图

1. 研究背景和动机：传统的强化学习方法在处理高维感知输入（如视频游戏画面）时面临困难。论文的目标是开发一种能够直接从原始像素输入学习控制策略的算法。
2. 深度 Q 网络（DQN）介绍：论文提出了深度 Q 网络，这是一种结合了卷积神经网络（CNN）和 Q 学习的算法。网络通过观察游戏的像素和分数作为输入，学习如何在游戏中做出决策以最大化总奖励。
3. 关键创新：
    - 经验回放：为了打破连续样本间的相关性并稳定学习过程，DQN 使用了一个经验回放机制，它随机采样过去的经验来更新网络。
    - 目标网络：为了进一步稳定学习过程，DQN 使用了两个网络：一个进行当前的预测，另一个用作固定的目标。
4. 实验结果：DQN 在多个 Atari 2600 游戏上进行了测试，结果表

明它能够超越之前的强化学习方法，并在某些游戏中达到甚至超越了人类专家的水平。

5. 意义和影响：这篇论文不仅展示了深度学习在处理复杂感知任务中的潜力，而且还促进了强化学习在实际应用中的发展，对后续的研究产生了深远的影响。

"Playing Atari with Deep Reinforcement Learning" 这篇论文是深度学习和强化学习结合的一个重要转折点，特别是在使用深度神经网络处理复杂决策问题方面。

### 1.2 论文主要技术创新点

该论文主要有两个重要的创新点。

经验回放：为了提高学习的稳定性和效率，DQN 使用了一种称为“经验回放”的技术。在这种方法中，代理的经历（即状态、动作、奖励和新状态）被存储在回放记忆中，然后随机采样以训练网络。这有助于打破数据的时间相关性并减少过拟合的风险。

目标 Q-网络：DQN 引入了一个单独的目标网络来稳定学习过程。在传统的 Q-learning 中，更新目标和当前估计值可能会导致不稳定。通过引入目标网络，可以减少这种不稳定性，提高学习过程的效果。

## 2 复现工作说明

### 2.1 选择理由

作为一名刚开始科研的学生，我选择复现 "Playing Atari with Deep Reinforcement Learning" 这篇论文，原因如下：

1. 技术学习和实践：作为一个初入科研领域的学生，我需要实际的项目来提升我的编程和数据分析技能。通过复现理解这篇论文，我将学习如何使用 Python 和深度学习库（如 TensorFlow 或 PyTorch）来构建和训练深度 Q 网络（DQN）。这不仅是对我的技术能力的直接提升，也是一个实践中学习理论的机会。
2. 理解强化学习的核心概念：强化学习是机器学习中的一个复杂分支，通过实际操作一个具体的案例，我可以更深入地理解强化学习的基本原理，例如奖励信号、决策过程和学习策略。
3. 提高项目能力：这篇论文展示了如何处理高维感知输入（如视频游戏的像素），这对于理解如何应用强化学习解决实际问题是非常重要的。复现这个项目将帮助我学习如何处理类似的复杂数据输入，并将理论应用于实际情况。这个复现项目为我的科研生涯积累经验，为我未来的研究项目奠定基础。通过这个项目，我可以探索我感兴趣的研究方向，并建立起一定的研究背景。

   通过这个复现项目，我不仅能够加深对深度强化学习的理解，还能够提升自己的实践能力和科研技能，这对于我的学术发展非常重要。

## 2.2 拟复现的具体内容

作为研一学生，不熟悉框架的使用，拟复现的具体内容将集中在"Playing Atari with Deep Reinforcement Learning" 论文的改进上。复现的主要内容包括：

对应该论文的创新，本次复现通过两个方向。

一个是对在原论文中采用的参数更新机制做出改变，原论文中在目标网络更新时，新参数直接替换旧参数。这有时可能导致训练不稳定，特别是在复杂或不稳定的环境中。而通过软更新，涉及逐渐将新参数值混合到旧参数中。这通过一个加权平均来实现，其中保留了旧参数的一部分，并添加了新参数的一部分。

另一个是使用 Double dqn 代替 dqn 使得更新更加稳定。

1. 作选择： 在 DQN 中，目标 Q 值是通过下一个状态（next state）的目标网络计算的，但是它仍然使用了同一个网络的选择动作策略，这可能导致高估问题。Double DQN 解决了这个问题，它使用本地网络选择下一个状态的动作，然后使用目标网络来评估这些动作的 Q 值，从而减小高估的概率。
2. 提高训练稳定性： Double DQN 通过将选择和评估分离到不同的神经网络中，降低了高估的风险，从而提高了训练的稳定性。这使得 Double DQN 在许多强化学习任务中能够更快地收敛到更好的策略。

图 2 是 DQN 和 DoubleDQN 的不同更新公式。

$$Y_t^{\mathrm{DQN}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t^-) .$$

$$Y_t^{\mathrm{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t') .$$

图 2 具体参数更新

## 2.3 预期结果与演示

图 3 是 dqn 和 double dqn 的运行代码绘制的曲线图。 左边是 dqn，右边是 double_dqn，x 轴表示迭代的 epoch，y 轴表示平均 reward，可见两者在进行多次迭代的结果，double_dqn 的性能更好，rewad 值更大，如果进行更多 epoch 的迭代，会有更明显的效果。

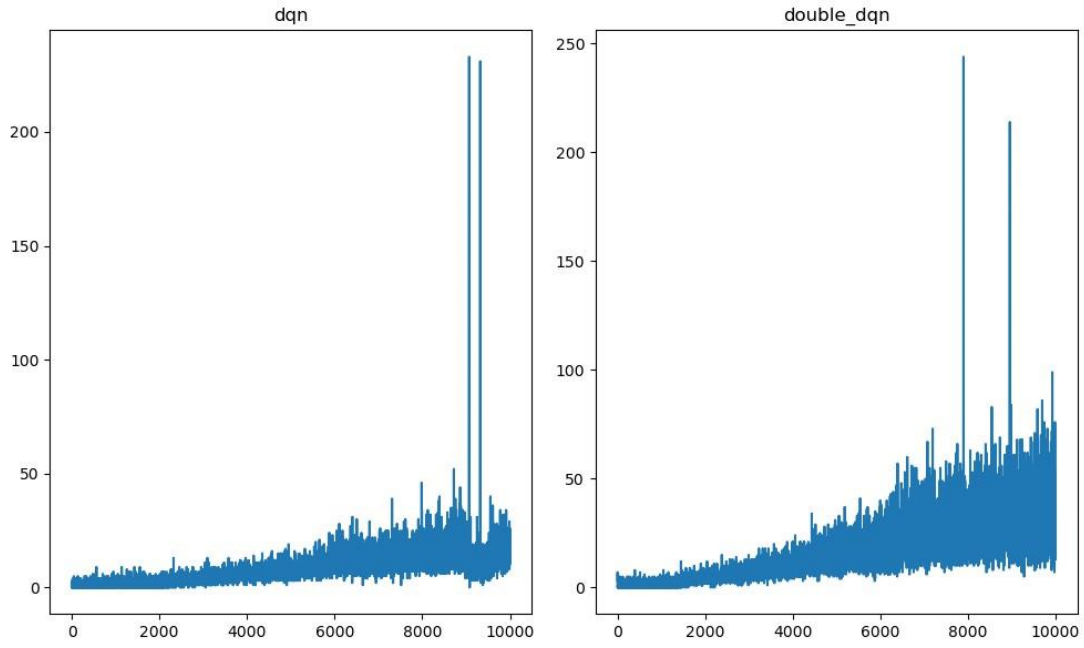


图 3 dqn 和 double dqn 对比

## 3. 复现工作计划进度

表 1 预期复现计划进度安排

| 时间安排 | 预计进度 |
|---|---|
| 2023 年 10 月 10 日-2023 年 11 月 20 日 | 论文阅读整理 |
| 2023 年 11 月 21 日-2023 年 12 月 20 日 | 论文代码阅读 |
| 2023 年 12 月 20 日-2024 年 1 月 10 日 | 论文代码改进及工作整理 |

# 参考文献


[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. arXiv e-prints. https://doi.org/10.48550/arXiv.1312.5602

[2] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In Thirtieth AAAI Conference on Artificial Intelligence.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[4] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

[5] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT press.

[6] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In International conference on machine learning (pp. 387-395).