深度 Q-學習與 Experience Replay Buffer 報告

May 7, 2025

Abstract

本報告首先總結 Deep Q-Learning（DQN）的核心機制與計算流程，接著介紹 Experience Replay Buffer 的設計理念與實作細節，最後🄴明兩者如何結合以提升訓練穩定性與樣本效率。

1 Deep Q-Learning 概述

Deep Q-Learning（DQN）是將經典 Q-Learning 中的表格函數 $Q(s,a)$ 替🄵🄵參數化的深度神經網路 $Q_\theta(s,a)$，以處理高維度、連續或🄴雜的狀態空間。

- **主網路（Online Network）** Q_θ：輸入狀態 s，輸出每個動作 a 的估計價值 $Q_\theta(s,a)$。
- **目標網路（Target Network）** Q_{θ^-}：🄵期性從 θ 更新🄵 θ^-，用於計算 TD 目標，降低目標波動。
- **-greedy 策略**：以概率 ε 探索（隨機選動作），否則選擇 $\arg\max_a Q_\theta(s,a)$。

1.1 TD 目標與損失函數

對於每個經驗組 (s,a,r,s',d)：

$$y = \begin{cases} r, & \text{若} d = 1(\text{終止}), \\ r + \gamma \max_{a'} Q_{\theta^-}(s',a'), & \text{否則}, \end{cases}$$

當前估計值 $\hat{Q} = Q_\theta(s,a)$，🄵用均方誤差作🄵損失：

$$\mathcal{L}(\theta) = \mathbb{E}\big[(\hat{Q} - y)^2\big].$$

利用隨機梯度下降更新 θ：$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$。

1.2 主要演算法流程

Algorithm 1 Deep Q-Learning with Target Network

Initialize online network Q_θ and target network $Q_{\theta^-} \leftarrow Q_\theta$
Initialize replay buffer $\mathcal{D}$
for episode = 1 to M **do**
  Reset environment, get initial state s
  while not done **do**
    Select a via -greedy from $Q_\theta(s, \cdot)$
    Execute a, observe r, s', d
    Store (s, a, r, s', d) into $\mathcal{D}$
    Sample minibatch from $\mathcal{D}$, compute loss $\mathcal{L}(\theta)$
    Update θ by gradient descent
    if step mod $C = 0$ **then**
      $\theta^- \leftarrow \theta$
    end if
    $s \leftarrow s'$
  end while
end for

2 Experience Replay Buffer 介紹

Experience Replay Buffer 是一種緩衝區,用於存儲代理與環境互動🄴生的經驗:$(s_t, a_t, r_t, s_{t+1}, d_t)$。

- **動機**：打破資料時間相關性，提升訓練穩定性；重🄴利用樣本，提高樣本效率。
- **結構**：常用固定長度環形緩衝或 `deque(maxlen)`，滿時覆蓋最舊資料。
- **操作**：
  1. `push(state,action,reward,next_state,done)`
  2. `sample(batch_size)` 隨機抽取 minibatch。
  3. 轉🄵🄵張量，送入網路訓練。
- **可擴充**：優先級回放（Prioritized Experience Replay）、多步返回（n-step returns）、重要性抽樣等。

3 DQN 與 Replay Buffer 的結合

將 Replay Buffer 與 DQN 結合的步驟如下：

1. **互動階段**：每步🄴 -greedy 選動作，執行後將經驗 (s, a, r, s', d) 存入緩衝區。
2. **訓練觸發**：當緩衝區樣本量 $\geq$ batch_size，即可抽樣🄵訓練。
3. **抽樣與更新**：
   - 隨機抽取 minibatch，計算當前 Q 預測與 TD 目標 y。
   - 最小化 MSE 損失 $\mathcal{L}(\theta)$，更新 online network 參數。
4. **目標網路同步**：每 C 步將 θ 🄵🄵至 θ^-，穩定目標估計。

整合後，Replay Buffer 能有效去相關、重🄴利用經驗，🄴在 DQN 中提供穩定且高效的訓練機制。