ISSN 2186-5647



2-12

# 優先度付き経験再生を行う利益分配深層強化学習

日大生産工　○林　航輝　　日大生産工　山内　ゆかり

## 1. まえがき

近年，強化学習が注目されている．強化学習は，エージェントと環境が相互作用を起こし，得る報酬の総和を最大にする行動を身に付けることが目的である．その代表的な手法として，Q学習(Q-Learning:QL)がよく知られている．QLは，行動数×状態数分のQTableを用いて，行動価値関数を更新する．そのため，囲碁や将棋などの膨大な状態数の環境下での学習は不向きである．そこで，QLに深層学習を組み合わせた深層強化学習(Deep-Q-Network:DQN)[1]が2013年に提案された．DQNでは，学習時に状態間の強い相関を解消するため，学習のサンプルデータをメモリに保存してランダムに抜き取る手法を用いた．これを，経験再生と呼ぶ．ネットワークで行動価値関数を表現することにより，膨大な量のQTableを必要とせず，Atari2600と呼ばれるゲーム環境下での学習に成功した．

宮崎ら[2]はDQNにXoL手法の一つである利益分配法(PrrofitSharing:PS)を組み合わせた”DQNwithPS”を提案した．この手法は，エージェントが報酬を得ると，エピソードの経験に対して報酬を分配し，利益分配に基づいた経験再生を行うことが特徴である．そのため，DQNのみよりも成功体験を強く意識しており，少ないエピソード数で安定した学習を行えたことが報告されている．さらなる学習の高速化及び利益分配の恩恵を受けるには，経験再生に何かしらの指標を付け，効率的に必要なサンプルデータの抜き取りができるように拡張する必要があると考える．

本研究では，利益分配後の完全ランダムな経験再生に優先度を与えた”PS優先度付き経験再生”を提案し，迷路問題における計算機実験により提案手法と従来手法を比較し，さらなる学習の高速化を目指す．

## 2. 従来手法

### 2.1 Deep-Q-Network

DQNのアルゴリズムをフローチャートとして図１に示し，概要を説明する．まず，マップ状態$s_t$の入力に対して，方策に従い出力から行動$a_t$を選択する．状態遷移先の$s_{t+1}$と報酬$r_t$を加え，4つの情報を経験としてメモリに保存する．その後，Q-NetworkとTarget-Networkにそれぞれ$s_t$と$s_{t+1}$を入力として与えて，行動価値関数を出力する．式(1)に，TDターゲットの求め方を示す．これは，教師あり学習における正解ラベルに相当する．

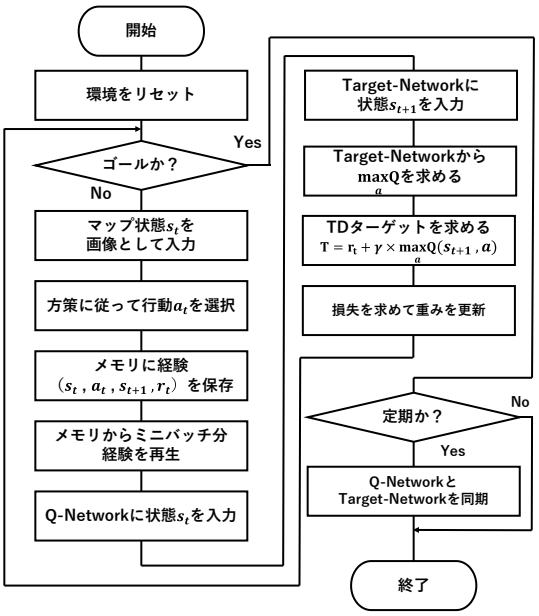


図１．DQNアルゴリズム

$$\boldsymbol{T = r_t + \gamma \max_{a} Q(s_{t+1}, a)} \qquad (1)$$

Q-Networkの出力とTDターゲットの損失を求め，重みを更新する．2つのネットワークで役割を分け，定期的に重みを更新することにより，TDターゲットの変動を抑え，安定した学習を行うことができると考えられている．

### 2.2 Deep-Q-Network with Profit Sharing

DQNwithPSはエージェントが報酬を得るたびにエピソード単位で報酬を伝搬させ，それに基づいた経験をメモリから再生する．そのため，ネットワークの主要部分はDQNと同じであり，唯一の違いは”PSに基づいた経験再生”を実行することだけである．DQNにPSを加えるのみのため，本質を変える必要がなく，他の拡張したDQNの手法に組み合わせることが容易である．

報酬を伝搬させるため，減衰関数を使用する．時刻tにおける行動を$a_t$，状態を$s_t$とした場合の報酬を$Q(s_t, a_t)$とし，報酬の分配式を式(2)に示す．$R$は報酬値，$N$は1エピソードの総ステップ数，$t$はステップ数，λは割引率である．

$$\boldsymbol{Q(s_t, a_t) = \lambda^{N-t} R} \qquad (2)$$
$$\boldsymbol{t = 1, 2, \ldots, N \ , 0 < \lambda < 1}$$

Profit Sharing Deep Reinforcement Learning with Prioritized Experience Replay
Kouki HAYASHI and Yukari YAMAUCHI

報酬を各経験に分配した後は，ミニバッチサイズ分の経験をメモリから完全ランダムに再生してQ-Networkの重みを更新する．

### 2.3 Prioritized Experience Replay

優先度付き経験再生(Prioritized Experience Replay:PER)[3]とは，DQNにおける経験再生に優先度の指標を付け，学ぶべき経験を積極的に再生することを目的とした手法である．優先度の指標には，式(3)に示したTD誤差を用いる．

$$\delta = r_t + \gamma \times \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \quad (3)$$

この値の大きい経験は，意外性があると考えられるため積極的に学ぶことが求められる．そのため，経験をメモリに保存する際にその経験のTD誤差も同時に格納し，式(4)に沿って確率$P(i)$を求めて経験を抜き取るようにする．

$$P(i) = \frac{(|\delta_i| + \varepsilon)^\alpha}{\sum_k^N (|\delta_k| + \varepsilon)^\alpha} \quad (4)$$

ここでのαは，優先的に経験を抜き出す度合いであり，値が大きいほど意外性のある経験を貪欲に再生できるよう調節できる．

## 3. 提案手法

提案手法の”PS優先度付き経験再生”のアルゴリズムをフローチャートとして図２に示す．

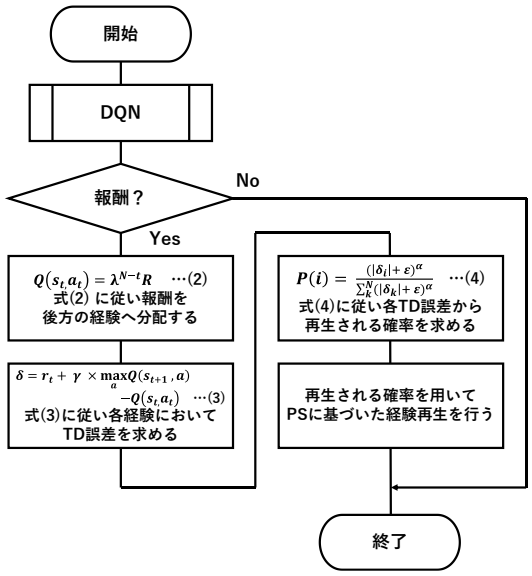


図２．PS 優先度付き経験再生アルゴリズム

まず，式(2)のPSの手法を用いて，報酬にたどり着くまでの経験に対して報酬を分配する．分配後は先ほどの経験において式(3)を用い，TD誤差を求める．次に式(4)に従い，TD誤差から各経験の再生される確率を求める．TD誤差の大きい経験は意外性があると考えられるため，再生される確率がそれに応じて高くなる．その後，確率を基にメモリからミニバッチ分の経験を抜き出して，PSに基づいた経験再生を行う．

## 4. 実験および検討

”DQN”,”DQNwithPS”,”DQN(PER)”の３つの従来手法と提案手法の比較を図３の環境で行う．

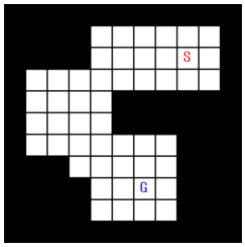


図３．迷路探索の実験環境

4つの手法を比較する指標には，S地点(スタート)からG地点(ゴール)までにかかる行動数を1000回分用いる．この実験を5回行い，平均として結果を表１に示す．

表 1．従来手法と提案手法の比較

| | 1回目 | 2回目 | 3回目 | 4回目 | 5回目 | 平均 |
|---|---|---|---|---|---|---|
| DQN | 74626 | 82013 | 76570 | 73935 | 71060 | 75640.8 |
| DQNwithPS | 64057 | 68242 | 63478 | 64662 | 68826 | 65853.0 |
| DQN(PER) | 64339 | 66260 | 61614 | 69938 | 64465 | 65323.2 |
| 提案手法 | 65179 | 59710 | 59343 | 59693 | 61847 | 61154.4 |

上記の表から，3つの従来手法と提案手法を比較する．それぞれの手法の行動回数を約15000回,4500回,4000回の削減をすることができた．“PSに基づいた経験再生”は完全ランダムに経験を抜き出していたが，TD誤差の指標を用いたことで学びたい経験を積極的に再生することができたため行動回数の削減に繋がったと考える．

## 5. まとめ

提案手法”PS優先度付き経験再生”により，エージェントの行動回数が減少した．今後は，他の複雑な学習環境下での実験を行い，提案手法の有効性をさらに確認していきたいと思う．

### 【参考文献】

[1]V.Mnih,K.Kavukcuoglu,D.Silver,A. Graves, I.Antonoglou,D.Wierstra, and M.Riedmiller, “Playing Atari with DeepReinforce-ment Learning”NIPS Deep Learning Workshop 2013, (2013).

[2]Kazuteru Miyazaki,”Exploitation-Oriented Learning with Deep Learning - Introducing Profit Sharing to a Deep Q-Network –”, Journal of Advanced Computational Intelligence and Intelligent Informatics 21 (5), 849-855, 2017-09-20, 富士技術出版株式会社 (2017)

[3]Tom.Schaul,John.Quan,Ioannis.Antonoglou,David.Silver,“PrioritizedExperienceReplay”, arXiv:1511.05952 [cs.LG],25 Feb 2016,(2016)