ISSN 2186-5647

2-25

メモリーの動的圧縮による好奇心探索

日大生産工(院)　○種田 祥吾　　日大生産工　山内　ゆかり

1. まえがき

強化学習分野において深層強化学習のDeep Q Network(DQN)[1]は有力な手法として用いられている。またDQNは様々な手法と組合され、発展したモデルが生まれている。Burdaらは状態に対して内部報酬を与えることによってまだ訪れていない状態での報酬を大きくし、よく訪れる状態での内部報酬を小さくすることで未知の環境への探索を促す好奇心探索(Random Network Distillation)[2]を提案し、Atari2600で難しいモンテズマの復讐[3]において人間よりも優れたスコアを得ることができた。

深層強化学習モデルの学習において経験再生(Experience Replay)は必要な要素の一つである。経験再生では蓄えた経験を用いてミニバッチ学習を行っている。Tomらは経験を蓄積しても選択確率は一様であるという問題点から、経験に優先度を与え、経験再生する優先度付き経験再生(Prioritized Experience Replay)[4]を提案した。

本研究では，この問題において経験バッファを動的に小さくすることを提案する。学習初期では好奇心探索を行い、未知の状態を探索し続け、一定の学習が進んだ時メモリーサイズを小さくすることで新しく追加された経験を学習しやすくし、学習の収束を早めることを目指す。

2. 従来研究

2.1 Deep Q Network

Deep Q Network(DQN)はQ学習について深層ニューラルネットワーク(DNN)を用いて実現する手法である。DQNの構造をFigure1に示す。

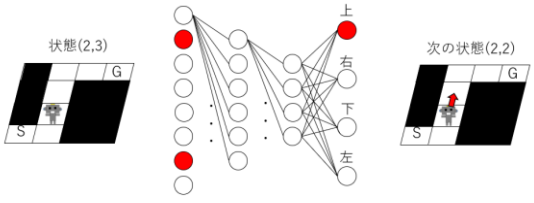

Figure 1　DQN の構造

DQNの入力には現在の状態(座標や画像)を使用する。出力値がQ値となり、出力値を元に方策を用いて行動を選択する。エージェントは環境における状態から最適な行動を選択できるように行動価値(Q値)を学習させる。Q学習の更新式は式(1)で表されるが、DQNでは出力値が式(2)になるように学習させる。

$$Q(s_t,a_t) = Q(s_t,a_t) + \eta \times (R_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)) \quad (1)$$

$$Q(s_t,a_t) = R_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') \quad (2)$$

2.2 Experience Replay

DQNで安定した学習を行うための手法として経験再生(Experience Replay)と呼ばれる手法がある。Q学習では1エピソード毎に学習を行うが、経験再生では各ステップの経験をメモリーに保存しておき、各ステップでメモリーから複数ステップを取り出し、ミニバッチ学習を行う。経験には現在の状態、その時選択した行動、次の状態、報酬が保存される。

2.3 Random Network Distillation

Random Network Distillation(RND) はOpenAIによって提案された手法である。エージェントはランダムに初期化されたTarget NetworkとTarget Networkの出力を予測するPredictor Networkを使用して学習する。Figure2にRNDの構造を示す。

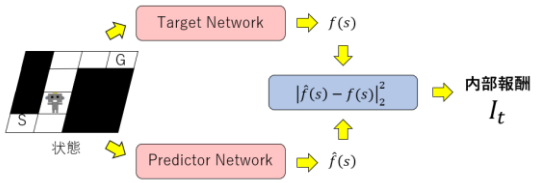

Figure 2　RND の構造

両ネットワークの二乗誤差を内部報酬I_tとして式(2)に加えることでエージェントは内部報酬の高い新しい状態や不確実な状態へ遷移する傾向がある。RNDによってエージェントは未知の状態に興味を持ち、幅広い環境を探索し、柔軟な行動方針を得る。

3. 提案手法

経験再生において事前に得た良い経験がミニバッチとして選択される確率は小さい。最初は経験が蓄積されていないため、直近のエピソードの経験が使用されているが、学習が進むにつれて直近のエピソードがミニバッチとして選択されない回数は少なくなる。

Curiosity Search by Dynamic Compression of Memory

Shogo TANEDA and Yukari YAMAUCHI

この問題を解決するために、本研究では動的に経験メモリーを圧縮することを提案する。経験再生における経験を記憶するメモリーを動的に小さくする。二つのメモリーの圧縮方法を示す。

① エピソードが進むにつれて小さくする。

② 1エピソードで得られた行動数に対して、得られた行動数×n回分を学習に使用する経験メモリーとする。Figure3に提案②のイメージ図を示す。

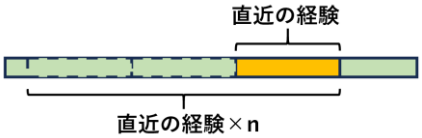

Figure 3　提案手法②

提案手法①では学習を考慮せず、線形に減らしていくが、提案手法②では、学習を考慮して得られた行動を基に経験再生を行う。

4.　実験および検討

好奇心探索を従来手法として実験し、従来手法に提案手法を加え、実験を行う。実験では行動回数及び総行動回数について記録をとる。また、実験に使用するバッチ数は20であるためバッチ数を基にミニバッチ平均選択回数について調査する。実験環境としてグリッドワールドの迷路問題を使用する。9×9の迷路に障害物を設置し、開始位置からゴールまで行動し、ゴールにたどり着いたとき、報酬として1が得られる。提案手法②に関しては直近の経験×3回分の経験を用いる。

従来手法と提案手法の10回平均の実験結果を示す。Figure4に各エピソードの行動回数を示す。

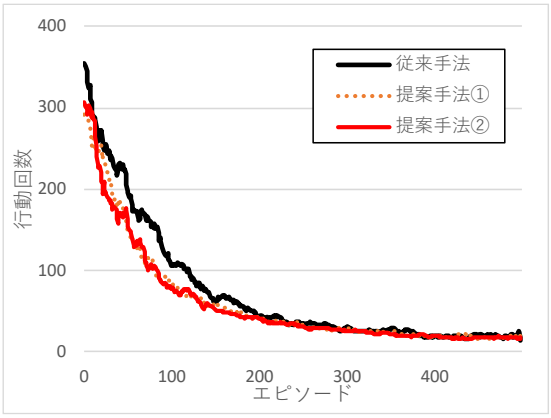

Figure 4　各エピソードの行動回数

それぞれ横軸がエピソード、縦軸が行動回数である。各提案手法は従来手法より早く収束していることがわかる。Table 1に総行動回数の表を示す。

Table 1　総行動回数

	総行動回数
従来手法	37631.3
提案手法①	31098.9
提案手法②	30406.0

従来手法よりも提案手法の総行動回数を6000回ほど減らすことができている。提案手法②が最も総行動回数が小さくなった。このことから学習を考慮した提案手法②が最も有効だといえる。次に各エピソードの平均選択回数をFigure5に示す。

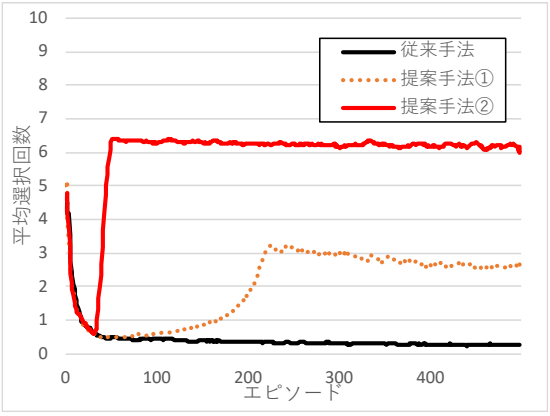

Figure 5　各エピソードの平均選択回数

提案手法①、②ともに選択回数が増加した。その中で提案手法②が最も選択回数が増加していることがわかる。

5.　まとめ

本研究では、好奇心探索に対して2つのメモリー圧縮の提案を行った。提案手法によって直近の経験の選択回数が増えたことにより、収束が早くなることを示し、また、総行動回数は従来手法より小さくなることを示した。今後、学習を考慮した場合の他の手法について調査し、実験を行いたい。

参考文献

[1] Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602 (2013).

[2] Burda, Yuri, et al. "Exploration by random network distillation." arXiv preprint arXiv:1810.12894 (2018).

[3] Montezuma's Revenge-Atari2600 https://www.retrogames.cz/play_124-Atari2600.php?language=EN (参照 2023-09-05)

[4] Schaul, Tom, et al. "Prioritized experience replay." arXiv preprint arXiv:1511.05952 (2015).