ISSN 2186-5647

P-17

Experience Replay の改良による DQN の性能向上

日大生産工（学部）　○室伏 玲耶　日大生産工　山内 ゆかり

1 まえがき

DQN（deep Q-network）とは、Googleの子会社のDeepMindが開発した人工知能である。DQNは深層強化学習アルゴリズムを利用していて、Atari2600の一部のゲームにおいて人間以上のスコアを獲得できている。[1]

深層強化学習とは、深層学習を用いた特徴抽出と強化学習の手法である Q 学習を組み合わせた技術である。Q 学習には、入力値を画像情報から Q 学習を行うと、エージェントが過去の画像から得られる高次元な情報から、特徴量を抽出し活用しなければならない問題がある。そこで、DeepMind は画像からの特徴抽出に深層学習の手法である Convolutional Neural Network(CNN)を用いて問題の解決を試みた。これにより、ゲームのルールを教えない場合でも視覚と試行錯誤のみで、高いスコアを獲得できる方策を学習できた。[2]

だが、Q 学習と CNN を組み合わせただけではスコアがあまり高くならなかった。そこで、DeepMind は連続値に対しても Q 学習を適用できるように Experience Replay という手法を導入した。これは、過去の経験を Replay Memory という配列に蓄積して、蓄積した経験からランダムに選び、教師データとして Q 値を更新する手法である。[3]この手法を加えたことがスコアの向上に大きく貢献した。

しかし、Experience Replay はランダムに経験を選ぶので、似たような経験を何度か選んでしまい、あまり学習に有用ではない結果になってしまう可能性がある。

そこで、本研究では、蓄積していく記憶を選ぶときに同じ経験を選ばないことを目的とした手法を提案する。

2 従来研究

DeepMind が提案した DQN について説明していく。まずネットワーク部分は以下の図 1 の通りである。

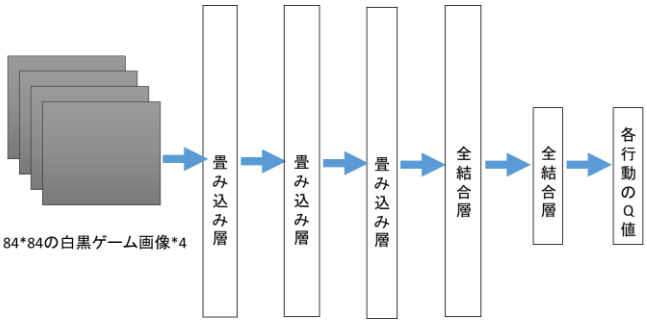

図1. DQN のネットワーク構成

図1は、DQN のネットワーク構成であり、畳み込み層が 3 層と全結合層が 2 層で構成されている。まず入力画像として、ゲーム画面の画像を 84*84*4(ピクセル)にリサイズし、RGB 値をグレースケールに変換したものを与える。4 枚のゲーム画像は直近の 4 フレームを 1 フレームあたり一枚に分けたものである。出力層のユニット数はプレイするゲームのアクションの数の分だけ Q 値を求める。

次にエージェント部分のアルゴリズムを示す。

1. CNN 部分から、Q 値を受け取り、4step に 1 度、ε -greedy 法により行動a_tを選択し、実行する。4step に 1 度ではない場合は前回の行動a_tを選択し、実行する。
2. 実行後、s_t, a_t, r_t, s_{t+1} の値を Replay Memory に保存する。保存量が限界の場合は古い順に上書きして、3 に進む。そうでない場合は 4 に進む。
3. Replay Memory の中からランダムに選択し、Q 値を更新する。
4. step を進めて 1 に戻る。

3 提案手法

本研究では、従来研究のランダム性により効率的に学習できない可能性がある問題を解決するべく、同じような記憶を選ばない方法を提案する。

そのため、基本的なアルゴリズムの流れは従来と変わらないが、Experience Replay内の記憶をランダムに選ぶとき、同じような結果を選

Performance Enhancement of DQN by Improving Experience Replay
Reiya MUROHUSHI and Yukari YAMAUCHI

ぶのではなく、s_t, a_t, r_t, s_{t+1}の値を比較して、同一でなければ学習に使うようにする。

4　実験環境

本研究では予備実験として以下の三目並べを行う。

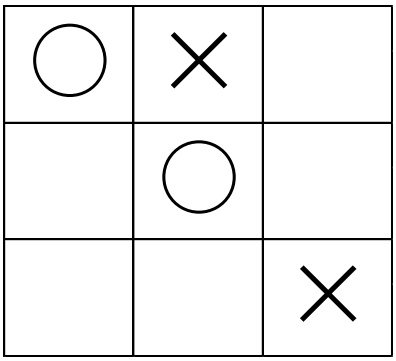

図 2.予備実験環境

実験環境は、図 2 の盤面のように○、×、空の 3 状態で、9 マスである。学習には、普段はランダムに選択し、勝てるならその場所を選択する CPU を使う。報酬は 1 回ゲームごとに勝つと+1、引き分けで 0.5、負けると 0 となる。また、1000 戦ごとに勝率を保存する。これを 100 回行い、通常の Q 学習と DQN、提案手法を比較する。また、方策を学習しているか確認するため、自分で実際に対戦して確認する。図 3 に Q 学習の勝率のグラフを示す。

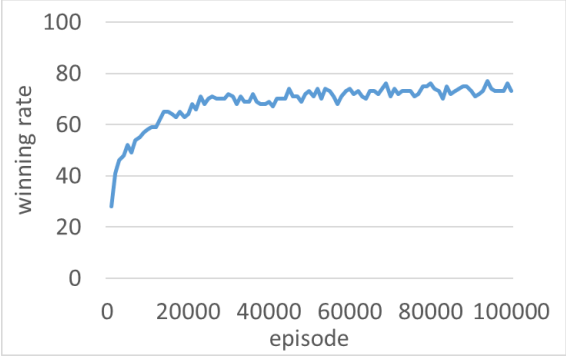

図 3.Q 学習の実験結果

3 目並べでは状態数が少ないため、通常の Q 学習でも学習が行えていることが確認できる。

本実験は同様の条件で、盤面を図 4 に示した 5 目並べで行う。なお、5 目並べでは、○×ではなく、○●を用いる。盤面のサイズが増大したことで、状態数が 3 の 100 乗となり、通常の Q 学習では学習が困難であると考えられるため、DQN による畳み込みが有効であると考えられる。

	0	1	2	3	4	5	6	7	8	9
0										
1		○	○	●						
2				○	●	●				
3					●					
4					○					
5										
6										
7										
8										
9										

図 4.実験環境

5　まとめ

現段階では提案手法が完成していないが、予想では、既存手法より、効率よく学習するはずのため、より少ないエピソード数で高い勝率が出ると考えられる。また、5 目並べによる実験では、状態数が三目並べと比べ多くなり、QL では方策が固まるのに時間がかかるが、DQN は、畳み込み処理により、学習数が QL より短い回数で方策が固まることが期待される。

「参考文献」

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver,Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312.5602, 2013.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver,Andrei A Rusu, Joel Veness, Marc G Bellemare, AlexGraves, Martin Riedmiller, Andreas K Fidjeland, GeorgOstrovski, et al. Human-level control through deep reinforcement learning. Nature, Vol. 518, No. 7540, pp.529–533, 2015.

[3] Long-Ji Lin. Reinforcement learning for robots using neural networks. Technical report, DTIC Document, 1993.