6U-04

複数報酬型 DQN の提案

片岡　玄太 †　　顧　優輝 ‡　　真部　雄介 †　　菅原　研次 †

† 千葉工業大学情報科学部 ‡ 千葉工業大学大学院情報科学研究科

1　はじめに

近年，ディープラーニングと強化学習を組み合わせた手法の研究が盛んに行われおり，その 1 つに Google DeepMind 社が開発した Deep Q-network（以下，DQN と呼ぶ）が挙げられる [1]．DQN とは Convolutional Neural Network(以下，CNN と呼ぶ）によって Q 学習の行動価値関数を近似する学習モデルである．

本研究では DQN に対し，独自の報酬を与えて CNN の学習を行うことで特定の行動を取る行動価値関数を作成する手法，及びそれらを制御して行動を決定する手法の提案と評価を行う．

2　提案手法

2.1　複数の報酬戦略に基づく行動価値関数の学習手法

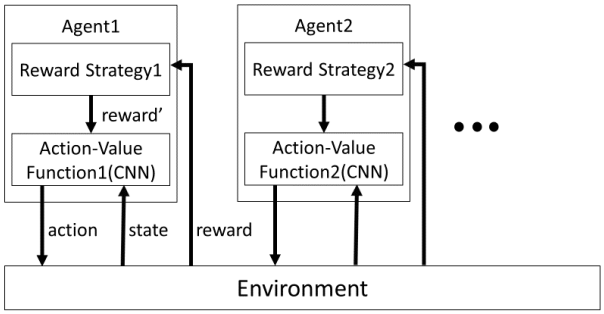

図1　行動価値関数の学習手法の概要図

図 1に行動価値関数の学習手法の概要を示す．通常，DQN は環境から与えられる報酬の値を使用して CNN の重みを更新するが，本研究では報酬因子と報酬戦略という考え方を導入し，図 1のように報酬戦略を用いて独自の報酬を与える．複数の報酬戦略を設定することで，特定の行動を取る行動価値関数を作成する．

図 2に報酬因子, 報酬戦略の例を示す．報酬因子とは

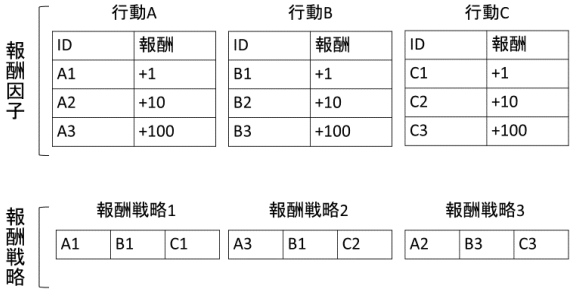

図2　報酬因子と報酬戦略の例

強化学習における行動とそれによって得られる報酬の組み合わせとする．一般的な強化学習では 1 つの行動に対して 1 つの報酬因子を持つが，本研究では 1 つの行動に複数の報酬因子を設定する．報酬戦略とは行動ごとに報酬因子を 1 つずつ選択し，それを組み合わせたものである．DQN の学習の際に，報酬戦略を用いることで独自の報酬を与える．

2.2　複数の行動価値関数を利用した行動の決定

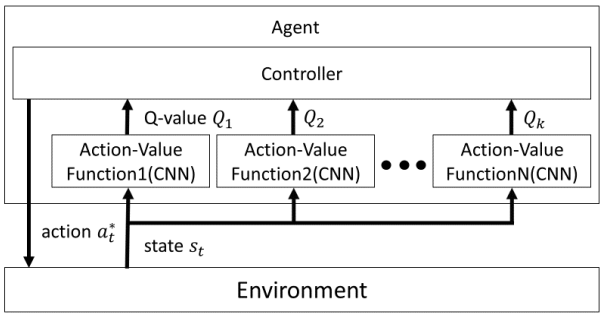

図3　行動決定手法の概要図

図 3に行動決定手法の概要を示す．行動の決定は複数の行動価値関数とコントローラを組み合わせたエージェントによって行う．最初に各行動価値関数は環境から状態を受け取る．次に各行動価関数はその状態で取り得ることが可能な行動の Q 値を全てコントローラに対して与える．最後にコントローラは方策に従い，Q 値を使用して行動を決定する．本研究では方策として，式 (1) で定義する，Q 値が最大の行動を採用する最大値採用方式

A Proposal of Multi-Reward Strategy Based DQN
†Genta KATAOKA　†Yuki KAERI　†Yusuke MANABE　†Kenji SUGAWARA
†Graduate School of Information and Computer Science, Chiba Institute of Technology

を提案する．s_t を状態，a を行動，A を行動の候補としたとき，式 (1) を満たす行動 a_t^* を求めることで行動決定を行う．

$$a_t^* = \underset{a \in A}{argmax} Q^{max}(s_t, a) \quad (1)$$

ここで用いる，状態 s_t における行動 a の Q 値が最も大きい行動価値関数を表す関数 $Q^{max}(s_t, a)$ は，式 (2) で表される．Q_k は報酬戦略に基づき作成された行動価値関数とする．

$$Q^{max}(s_t, a) = \max_k Q_k(s_t, a) \quad (2)$$

3　評価実験

提案手法の有効性を示すために Open AI が提供する AI シミュレーションプラットフォーム Open AI Gym を利用し，Atari 社のクラシックゲームである Ms.Pac-Man を対象に評価実験を行う．

3.1　実験 1　エージェントの行動傾向の分析

報酬因子

Pac-dot獲得行動	
ID	報酬
A1	+1
A2	+100

Power Pellet獲得行動	
ID	報酬
B1	+1
B2	+100

Ghost撃退行動	
ID	報酬
C1	+1
C2	+100

報酬戦略

Agent ID	Pac-dot獲得	Power Pellet獲得	Ghost撃退
Normal	A1	B1	C1
GetDot	A2	B1	C1
PowerUp	A1	B2	C1
Attack	A1	B1	C2

図4　実験で使用した報酬因子と報酬戦略

図 4に実験で使用した報酬因子と報酬戦略について示す．実験では Ms.Pac-Man の行動を 3 種類に分類し，それぞれを優先的に行う行動価値関数と優先度を付けない行動価値関数を作成する．そして，各行動価値関数を使用したエージェントの行動傾向を，3 つの評価項目を使用して分析を行う．評価項目は 1 ステップあたりの Pac-dot 獲得数，Power Pellet 獲得数,1 回のパワーアップあたりの Ghost 撃退数の 3 つである．

図 5に実験結果を示す．図 5は各項目の 1000 エピソード分の平均値を，Normal エージェントを 1 としたときの値に直したものである．全ての項目において，それぞれの項目の報酬を高く設定した報酬戦略を使用したエージェントが最大の値となった．このことから，報酬戦略を使用した学習を行うことで，行動価値関数の特定の

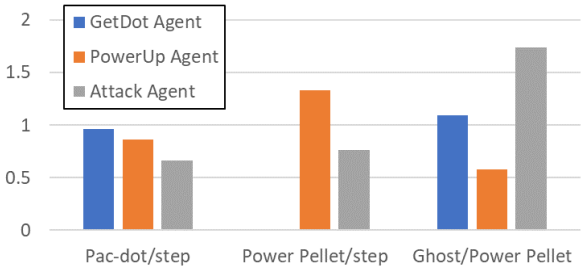

図5　Normal エージェントを基準としたときの行動傾向

行動を取る傾向を強くすることが可能であると考えられる．

3.2　実験 2　複数の行動価値関数を利用した行動の決定

表1　エージェント単体と組み合わせた場合のスコア

agent	max	min	mean
Attack	1060	190	709.25 ± 246.43
PowerUp	1940	450	816.67 ± 249.63
最大値採用方式	3670	40	853.59 ± 482.79

Attack エージェントと PowerUp エージェント単体のスコアと，最大値採用方式で組み合わせた場合のスコアを比較した．表 1に実験結果を示す．エージェントを組み合わせることで，最大スコア，平均スコアの記録が向上した．その一方で最小スコアが減少したこと，標準偏差が大きくなったことからスコアの幅が大きくなったことがわかる．

4　まとめ

報酬戦略を用いた行動価値関数の学習手法と，行動価値関数群を利用した行動決定の手法の提案，評価を行った．その結果特定の行動をとる傾向の強い行動価値関数の作成が可能であること，それらを組み合わせる手法によるスコアの向上を確認した．

今後の課題としては，スコアの幅が小さい，安定感のある行動価値関数群の制御方法の検討が挙げられる．

参考文献

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M. A.: Playing Atari with Deep Reinforcement Learning, CoRR, Vol. abs/1312.5602 (2013).