The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

様々な学習戦略と学習環境における Hybrid Reward Architectureの性能の評価

Evaluation of Hybrid Reward Architecture on various learning policies and environments

藤村悠太朗 *1　　金子知適 *2
Yutaro Fujimura　　Tomoyuki Kaneko

*1東京大学教養学部学際科学科
Department of Interdisciplinary Sciences, College of Arts and Sciences, The University of Tokyo

*2東京大学大学院情報学環
Interfaculty Initiative in Information Studies, the University of Tokyo

Deep Q-Network (DQN) was able to achieve a level comparable to the performance of a professional human player. However, in large and complex domains (e.g. Ms. Pacman), learning can be very slow and unstable. In Hybrid Reward Architecture (HRA), a reward function is decomposed in advance to enhance learning in such domains, and then value functions are separately learned for decomposed reward functions. In this paper, we made some environments that made learning more difficult to evaluate the performance of HRA. The results indicated that HRA need more enhancements to learn environments where learning is difficult on the uniform random policy.

1. はじめに

人工知能研究の一分野として，AI エージェントの強化学習に関する研究が数多く行われてきた．囲碁において AlphaGo [Silver 16] が人間のトッププレイヤと対戦して勝ち越し，チェスや将棋においても自己対戦のみで学習した Alpha Zero が既存の人間より強いプログラムに勝ち越す [Silver 17] など，様々なゲームで人間より強い AI エージェントが登場するまでとなった．一方で，リアルタイムで迅速な意思決定をする必要がある状況を想定し，このようなゲームに関する研究も注目されている．強化学習の手法である Q 学習 [Sutton 98] と深層学習を組み合わせた Deep Q-Network (DQN) という手法によって，Atari 2600 のゲームの一部で人間より高いスコアを記録した [Mnih 15] が，人間のプレイヤのスコアに達することができないゲームも存在し，そのようなゲームの中の 1 つが Ms. Pacman である．Ms. Pacman の難しさは取りうる状態の数が非常に多く，状態と報酬の関係も複雑であるため，DQN を用いてもうまく学習が進まないという点である．van Seijen らは Hybrid Reward Architecture (HRA) を提案し，人間のスコアを上回るプレイヤの作成に成功した [Seijen 17]. HRA は，ゲームがエージェントに与える報酬関数を分割し，それぞれに対して行動価値関数を学習することで，非常に大きな状態空間に対しても学習を可能にしている．

本研究では，Ms. Pacman よりも難しい性質を持ったゲームにも HRA の手法を適用できるかを調査するため，HRA による分割を用いた学習が適するゲーム環境をいくつか作成し，DQN と HRA で性能の比較と考察を行った．

2. 関連研究

ここでは，本研究がベースとしている，DQN と HRA について述べる．

連絡先: 藤村悠太朗, 東京大学教養学部学際科学科, 東京都目黒区駒場 3-8-1 15 号館, yut-mak874@g.ecc.u-tokyo.ac.jp

2.1 Deep Q-Network (DQN)

Deep Q-Network (DQN) [Mnih 15] は，Q 学習の行動価値関数 Q をニューラルネットワークを用いて関数近似することにより，学習を行う手法である．ニューラルネットワークの重みを θ としたときの行動価値関数を $Q(s,a;\theta)$ と書く．DQN では，次の誤差関数 $\mathcal{L}_i(\theta_i)$ を最小化するように θ を繰り返し更新することによって学習を行う．

$$\mathcal{L}_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[(y_i^{\mathrm{DQN}} - Q(s,a;\theta_i))^2\right], \quad (1)$$

ただし，　$y_i^{\mathrm{DQN}} = r + \gamma \max_{a'} Q(s',a';\theta_{i-1})$　(2)

ここで，θ_i とは，i 回更新が行われた後の重みを指す．

2.2 Hybrid Reward Architecture (HRA)

Hybrid Reward Architecture [Seijen 17] では, 環境報酬関数 R_{env} が式 (3) のように n 個の報酬関数に w_i の重み付け和で与えられる環境を考える:

$$R_{env}(s,a) = \sum_{k=1}^{n} w_k R_k(s,a) \qquad \text{for all } s,a. \quad (3)$$

そして，この分割された報酬関数それぞれに対して, 強化学習のエージェントを訓練する．それぞれのエージェント k が持つ行動価値関数を $Q_k(s,a)$ と書く．エージェントの意思決定には，分割された行動価値関数 Q_k の合計

$$Q_{\mathrm{HRA}}(s,a;\theta) := \sum_{k=1}^{n} w_k Q_k(s,a;\theta) \quad (4)$$

を用い，誤差関数

$$\mathcal{L}_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\sum_{k=1}^{n}(y_{k,i} - Q_k(s,a;\theta_i))^2\right], \quad (5)$$

ただし，　$y_{k,i} = R_k(s,a,s') + \gamma \sum_{a'\in\mathcal{A}} \frac{1}{|\mathcal{A}|} Q_k(s',a';\theta_{i-1})$　(6)

を最小化する．

問題固有の知識を用いた報酬関数の分割を行うことで, Ms. Pacman やそれに類似するゲームにおいて，DQN より行動価値関数の収束が早いことが文献 [Seijen 17] で示されている．

3. 問題設定

3.1 Fruits Collection Task

Fruits Collection Task は文献 [Seijen 17] で，Ms. Pacman の簡易版として提案されているゲームである．この環境を本研究では，FruitsCollection-v0 と呼称する．

このゲームでは，エージェントは 10×10 のグリッドの中で，できるだけ早くフルーツを集めることが目的である．グリッドのうち，10 箇所がフルーツ置き場となっており，各エピソードの開始時に，10 個のフルーツ置き場の中から，ランダムに 5 個が選ばれてフルーツが配置される．エージェントはランダムな位置から出発して，ターン毎に上下左右の 4 方向を選び，選ばれた方向に隣接するマスへ移動する．エージェントはフルーツを得た場合に 1 つあたり+1 の報酬を得て，それ以外の場合は 0 の報酬を得る．5 個のフルーツを全て集めるか，300 ターンが経過するとそのエピソードは終了する．

Fruits Collection Task に対して，いくつかの要素を加えたゲームを複数用意した．これは，行動の種類を増やしたり，報酬を得にくくするような要素を加えることによって，ゲームの設定を難しくすることが目的である．

3.2 エージェントの方向があるゲーム

FruitsCollection-v1 は，FruitsCollection-v0 に，次のルールを追加したゲームである．

- エージェントは上下左右の 4 つの方向を持つ．
- エージェントが行動するとき，
  - エージェントの方向と，選択した行動の方向が一致するとき，その方向に隣接するマスへ移動する．
  - そうでなければ，エージェントは，選択した行動の方向に向きを変える．

3.3 遠距離からフルーツを集める行動があるゲーム

FruitsCollection-v1 に更に次のルールを追加したゲームを作成し，これを FruitsCollection-v2 とする．

- エージェントは，上下左右へ 4 方向を選択する行動に加えて,「レーザーを撃つ」行動を選択することができる．
- レーザーを撃った場合，エージェントの向いている方向の半直線上に存在するフルーツすべてを回収する．

また，FruitsCollection-v1,2 を簡略化し，FruitsCollection-v0 に［上，下，左，右］の方向へレーザーを撃つという 4 つの行動を追加したゲーム FruitsCollection-v3 も作成した．

3.4 特定の行動によりペナルティが発生するゲーム

FruitsCollection-v0,1,2,3 は，ランダムに行動しても 300 ターンの間に報酬を得ることが可能な設定のゲームである．そこで，FruitsCollection-v2 に，次のルールを追加したゲームを作成した．

- 10×10 のグリッドのうち，中央の 8×8 の領域にエージェントが移動すると，-1.0 の報酬を得た上で，エピソード開始時の場所に移動する．

これは，[Sutton 98] で強化学習エージェントのための環境の例として提示されている Cliff Walking と FruitsCollection を組み合わせたものである．このルールが追加されることによって，正の点数でエピソードを終了するために取るべき行動の選択肢が少なくなり，ランダムに行動とっても 300 ターン以内にエピソードを終了することが難しくなる．

このゲームを本研究では，FruitsCollection-v4 と呼称する．

4. 比較対象とするエージェント

2. 節で述べた DQN, HRA によって学習を行うエージェントを作成した．

ニューラルネットワークの構造は，ゲームの状態を受け取る入力層，全結合の中間層となっていて，中間層から各ゲームで可能な行動数のユニットを持つ出力層へと繋がるという 3 層の構造で統一した．HRA では，出力層がフルーツ置き場の数だけ独立しており，それぞれの出力層は自分の担当するフルーツ置き場に関する行動価値関数を求める．また，Q_{HRA} を求める際は式 (4) を用いて計算する．

また，先行研究 [Seijen 17] での実装では，全結合層はそれぞれの出力層で独立に存在しているが，論文中にアイデアとして示されていた，全結合層を共有するネットワークでの学習も行った．これを，本研究では HRA-Share と呼称する．

エージェントに与える報酬については，DQN ではゲームから受け取った報酬をそのまま与え，HRA, HRA-Share ではどの場所のフルーツから得た報酬かに応じて，適切な出力層に報酬を与えた．

5. 実験と考察

本研究では，プログラムの実装には python 3.5 を利用した．提案した環境を実装するために Open AI Gym [Brockman 16] を利用し，深層学習のフレームワークとして tensorflow を利用した．

5.1 実験条件

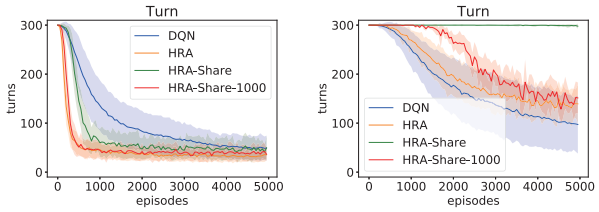
方向の要素がない FruitsCollection-v0,v3 がエージェントに与えるゲームの状態は，[Seijen 17] と同様に，10×10 のマップ内でエージェントがいる場所を表す one-hot ベクトルと，10 個のフルーツ置き場のそれぞれでフルーツがあるかどうかを表現するベクトルを結合した，要素数 110 のバイナリ型のベクトルとした．方向の要素がある FruitsCollection-v1,2 については，方向を表す one-hot ベクトルをこれに加えた，要素数 114 のバイナリ型のベクトルをエージェントに与えるゲームの状態とした．

先行研究 [Seijen 17] に従い，DQN における式 (2) の γ は 0.85, ネットワークの学習率は 0.0005，HRA における式 (6) の γ は 0.99, ネットワークの学習率は 0.001 とした．DQN，HRA-Share における全結合層のユニット数，及び，HRA で分割されている全結合層のそれぞれのユニット数は，全て 250 個とした．また，HRA-Share における全結合層のユニット数を 1000 個としたエージェントを以下の実験では，HRA-Share-1000 とする．

DQN では，訓練時は ϵ-greedy 方策で行動し，評価時は greedy 方策で行動する．この時の ϵ の値は 0.01 とする．

5.2 ランダム方策による訓練

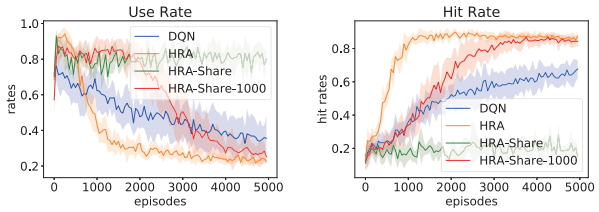
先行研究 [Seijen 17] と同様に，HRA における訓練例を一様ランダム方策で生成し，行動価値関数を式 (6) で更新する際の方策は，一様ランダム方策とした．



(a) FruitsCollection-v0 **(b)** FruitsCollection-v1

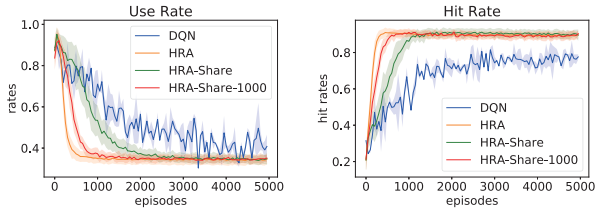
(c) FruitsCollection-v2 **(d)** FruitsCollection-v3

図 1: 各ゲームでのゲーム終了までに要したターン数



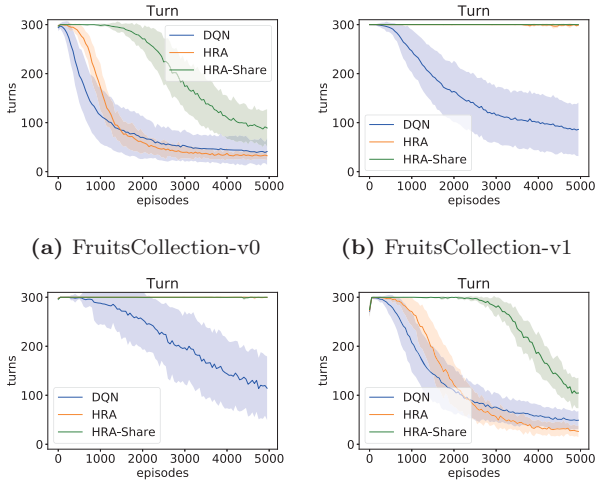
(a) レーザーを使用した割合 **(b)** 獲得した報酬のうちレーザーによって得た報酬の割合

図 2: FruitsCollection-v2 でのレーザーの使用について



(a) レーザーを使用した割合 **(b)** 獲得した報酬のうちレーザーによって得た報酬の割合

図 3: FruitsCollection-v3 でのレーザーの使用について



(a) FruitsCollection-v0 **(b)** FruitsCollection-v1

(c) FruitsCollection-v2 **(d)** FruitsCollection-v3

図 4: 各ゲームでのゲーム終了までに要したターン数

FruitsCollection-v0,1,2,3 で DQN, HRA, HRA-Share, HRA-Share-1000 を 5,000 エピソード学習させた．学習の過程での，エピソードを終了するまでに要したターン数の変化を示したグラフを図 1 に示した．横軸が学習したエピソード数，縦軸はゲーム終了に要したターン数であり，ゲーム終了に要したターン数が少ない程性能が良い．

図 1a の DQN, HRA を比較すると，先行研究 [Seijen 17] の実験と同等の結果が得られていることがわかる．また，HRA-Share は DQN と HRA の中間の性能を示しているが，全結合層のユニット数を増やした HRA-Share-1000 は HRA と同等の性能を示すことが読み取れる．図 1d より，FruitsCollection-v3 においても，この関係性は同様である．

図 1b, 図 1c を見ると，DQN, HRA は学習が進んでいる様子が読み取れるが，HRA-Share は学習が非常に遅く，HRA-Share-1000 は HRA-Share 程ではないが，学習が HRA より少し遅く進んでいることが読み取れる． FruitsCollection-v2 では，特に HRA は学習が早く，最終的に獲得する報酬，ゲーム終了までに要したターン数の値も，DQN より優れていることがわかる．

また，FruitsCollection-v2,3 で，エージェントがエピソード中にとった行動のうち，レーザーを使用した割合，及び，獲得した報酬のうちレーザーによって得た報酬の割合を示したグラフを図 2, 図 3 に示した．

図 1c, 1d と図 2, 3 を比較すると，どのエージェントも学習が進むと，レーザーを使用する割合は小さくなり，フルーツをレーザーによって回収する割合が大きくなることが読み取れる．

5.3 ϵ-greedy 方策による訓練

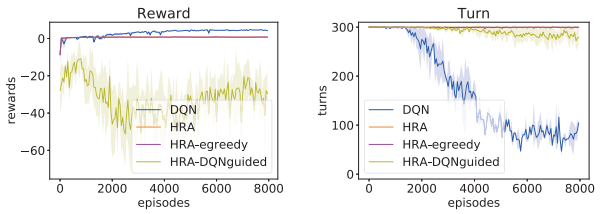
HRA における訓練例を，一様ランダム方策から，式 (4) を評価した ϵ-greedy 方策で行動したものに変更して学習を行う実験を行った．

HRA が行動を選択する際は，式 (4) を評価した ϵ-greedy 方策で行動し，行動価値関数を式 (6) で更新する際の方策は，一様ランダム方策とした．この時の ϵ の値は 0.01 とした．

FruitsCollection-v0,1,2,3 で DQN, HRA, HRA-Share を 5000 エピソード学習させた．HRA-Share-1000 は，学習にかかるコストが大きいため，今回は実験を行わなかった．学習の過程でゲームから得た報酬と，エピソードを終了するまでに要したターン数の変化を示したグラフを図 4 に示した．

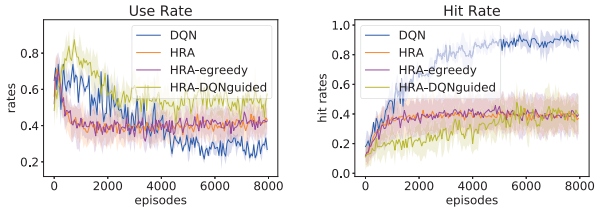
図 1 と図 4 を比較すると，ランダム方策で訓練例を作成する場合より，ϵ-greedy 方策で訓練例を作成する方法の方が，HRA の性能が悪化することが読み取れる．また，環境によって，ランダム方策で訓練例を作成した時は学習が進む環境の中で，ϵ-greedy 方策で訓練例を作成しても学習が進む場合と，学習が進まなくなる場合の両方があることがわかる．

FruitsCollection-v1 における実験では，エージェントが同じ状態にとどまり続ける様子が見られ，これが原因で中々エージェントが正の報酬を観測することができず，学習を進めることができないものと考えられる．FruitsCollection-v1 のように方向の要素がある場合，エージェントが異なる座標へ移動するためには，移動したい方向を向き，その後移動するために 2



(a) 獲得した報酬　　**(b)** ゲーム終了までに要したターン数

図 5: FruitsCollection-v4 での学習結果



(a) レーザーを使用した割合　　**(b)** 獲得した報酬のうちレーザーによって得た報酬の割合

図 6: FruitsCollection-v4 でのレーザーの使用について

回同じ行動を取るという特定の手順を行う必要がある．しかし，エージェントが行動価値関数を更新する方策が一様ランダム方策であるため，なかなか新しい状態に推移することができないと考えられる．つまり，より新しい状態に遷移させるための工夫によって，性能が改善する可能性がある．

5.4　報酬を得にくい環境の学習

5.2 節の実験から，FruitsCollection-v0,1,2,3 のように，ランダム方策でも報酬を得ることが可能なゲームでの HRA の性能を確認することはできた．しかし，5.3 節の実験を見ると，ランダム方策で報酬を得ることが難しい ϵ-greedy 方策での学習が必要になるようなゲームでは，HRA での学習が難しいことが予想される．

FruitsCollection-v4 で，DQN，5.2 節と同様にランダム方策で訓練例を作成する HRA，5.3 節と同様に ϵ-greedy 方策で訓練例を作成する HRA-egreedy，更に DQN で学習を行ったときの行動を訓練例として用いる HRA-DQNguided の 4 つのエージェントで 8000 エピソードの学習を行う実験を行った．

学習の過程でゲームから得た報酬と，エピソードを終了するまでに要したターン数の変化を示したグラフを図 5 に示した．また，レーザーを使用した割合，及び，獲得した報酬のうちレーザーによって得た報酬の割合を示したグラフを図 6 に示した．

図 5 を見ると，DQN ではこのゲームでも学習が進むが，HRA, HRA-egreedy, HRA-DQNguided はどれも学習がうまく進んでいないことがわかる．図 5b を見ると，HRA, HRA-egreedy と異なり，HRA-DQNguided は少しだけゲームをクリアしているエピソードもあるという差異はあるが，負の報酬を得ている割合も HRA, HRA-egreedy と比べて非常に高い．また，図 6 を見ると，DQN は図 2, 3 と同様に，学習が進むにつれて，レーザーを使用する頻度は小さくなり，フルーツをレーザーによって回収する割合が大きくなるが，HRA, HRA-egreedy, HRA-DQNguided はどれも途中で変化が止まってしまっていることがわかり，報酬だけでなくレーザー使用率からも，HRA で学習したエージェントが学習が進んでいないことを示している．

6.　おわりに

本研究では，HRA で学習するエージェントを，複数のゲーム環境と学習戦略で学習を行い，それぞれの性能を比較した．その結果，HRA はランダム方策での訓練と相性が良く，DQN よりも優れる学習効率と性能を発揮するが，ランダム方策で報酬を得ることが困難なゲーム環境においては，ϵ-greedy 方策や DQN で学習を行ったときの行動を訓練例とするといった手法では，学習が困難であることを示した．

今後の課題としては，5.3 節の実験で見られた，ϵ-greedy 方策における同じ状態に留まりつづける行動を抑制し，新たな状態に遷移させるような工夫を加えることが挙げられる．

謝辞

この研究の一部は，JSPS 科研費 16H02927 と JST さきがけの支援を受けています。

参考文献

[Brockman 16] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W.: OpenAI Gym, pp. 1–4 (2016)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Seijen 17] Seijen, van H., Fatemi, M., Romoff, J., Laroche, R., Barnes, T., and Tsang, J.: Hybrid Reward Architecture for Reinforcement Learning, No. Nips 2017, pp. 1–16 (2017)

[Silver 16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, van den G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, pp. 484 EP – (2016), Article

[Silver 17] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (2017)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning : An Introduction*, MIT Press (1998)