# 事前知識を導入した深層強化学習による行動獲得に関する研究

TP17004 稲盛有那　　指導教授：藤吉弘亘

## 1. はじめに

強化学習は，環境とエージェントのやりとりで得られる報酬を手掛かりに，報酬を最大化する行動を獲得する学習手法である．Deep Q-Network(DQN)[1] をはじめとする深層強化学習法が提案され，ゲームや制御問題を解決することが可能となっている．深層強化学習は，複雑な問題に対する制御を獲得できるが，学習パラメータ数の増加や勾配消失により学習が困難となる場合がある．また，現在の状態に対する行動が過去の行動に依存するような連続的な行動の獲得は，離散的な行動に比べて困難である．本研究では，人が操作した行動を事前知識として学習に導入することで，望ましい行動を効率的に獲得する手法を提案する．本研究では自動運転車両の車線変更を対象とし，シミュレータを用いた実験より提案手法が追従や追い越しといった人のような行動を獲得できることを示す．

## 2. Experience Replay

Experience Replay[2] とは，DQN において学習を安定化させるために用いられている手法の 1 つである．エージェントが獲得した状態，行動，報酬のデータを行動遷移として Replay Buffer と呼ぶメモリに蓄積する．そして，学習する際に蓄積したデータを繰り返しランダムサンプリングして利用する．通常の強化学習では，学習開始時，Replay Buffer の中身は空である．そのため，一定のエピソードを経て Replay Buffer に行動遷移を蓄積してから学習を始める．また，Replay Buffer の中身は古い経験から新しい経験に入れ替わっていく．しかし，初期の行動遷移は殆どランダムな行動をした時のものである．よって，膨大な状態と行動の組み合わせから最適な行動を学習するのに時間がかかる．

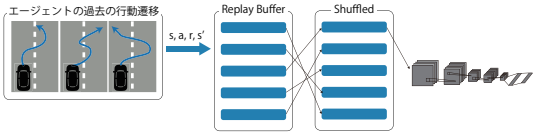



図 1: Experience Replay の流れ

## 3. 提案手法

提案手法では，事前知識として望ましい行動遷移を人が操作して作成し，Replay Buffer にあらかじめ蓄積してから学習を開始する．これにより，学習初期のランダムな行動選択を抑制する．

### 3.1 事前知識

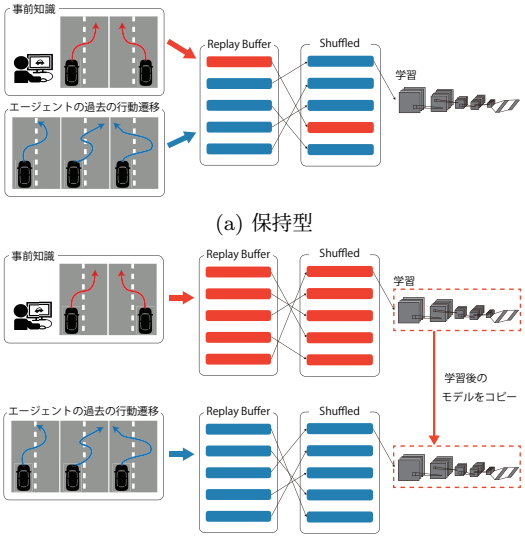
提案手法では学習初期のランダムな行動選択を抑制するために，人がエージェントを操作したときの行動遷移を Replay Buffer に入る形にして保存しておく．これを本研究では事前知識と呼ぶ．この事前知識を，Replay Buffer に蓄積してから学習を行う．

### 3.2 事前知識の与え方

本研究では，事前知識の与え方として，保持型と予行型を提案する．図 2 に，事前知識の与え方を示す．図 2 の Replay Buffer および Shuffled 内の赤色は事前知識，青色は学習中にエージェント自身が経験で得た行動遷移を表す．

**保持型**　図 2(a) に示すように，Replay Buffer 内の 1 割に事前知識を常に蓄積する．残りの 9 割は従来の Experience Replay と同様に，学習が進むにつれて古い経験が新しい経験に入れ替わる．これにより，Replay Buffer 内には常に望ましい行動遷移が蓄積され，一定の割合で事前知識を用いて学習を行い，ランダムな行動選択を抑制することが可能となる．

**予行型**　図 2(b) に示すように，学習前に Replay Buffer 内を事前知識で満たした状態で学習する．このとき，エージェント自身が経験した行動遷移は Replay Buffer には蓄積せず，事前知識のみで学習する．その後，学習したネットワークモデルを初期値にして，再度学習を行う．これにより，事前知識をもとに獲得した行動から，さらに最適な行動を獲得することが期待できる．




(a) 保持型

(b) 予行型

図 2: 事前知識の与え方

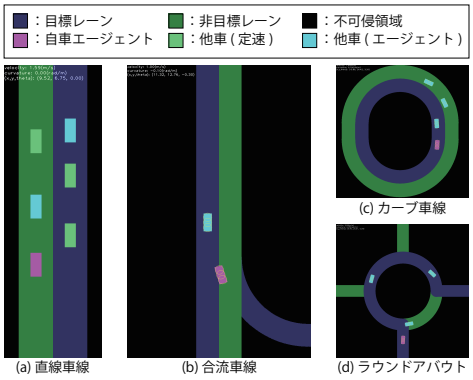



図 3: 各シミュレータ環境の例

### 3.3 シミュレータ環境

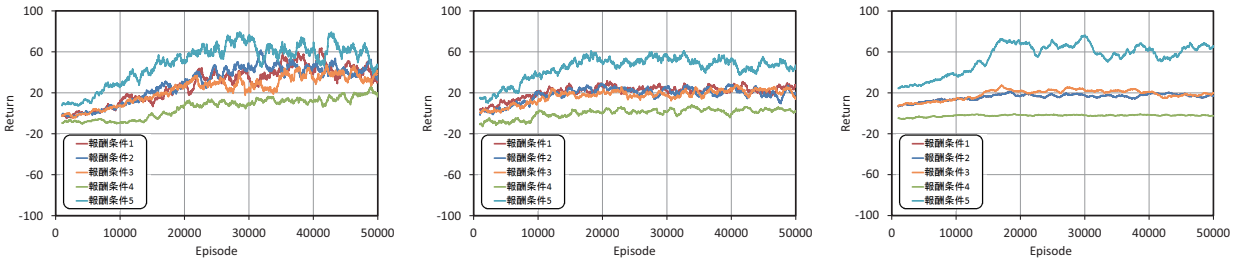
本研究では，自動運転の車両の車線変更を対象とする．図 3 に，本研究で扱うシミュレータ環境を示す．青色は目標レーン，緑色は非目標レーン，黒色は不可侵領域であり，ピンク色は自車エージェント，黄緑色は固定速度で直進する他車，水色は自車エージェントのモデルをコピーした他車である．エージェントが可能な行動は，速度 3 種，曲率 3 種を組み合わせた 9 つである．速度は，±0m/s(速度維持)，+0.1m/s(加速)，−0.1m/s(減速) の 3 種である．曲率は，0rad/m(車体角度維持)，+0.01rad/m(左回転)，−0.01rad/m(右回転) の 3 種である．

## 4. 報酬設計の検証

図 3(b)，(c)，(d) の環境において，どのような報酬設計が適しているかを検証する．表 1 に，検証する報酬条件を示す．全検証で共通して，エピソードの終了条件は衝突か 100 ステップ到達である．本実験では，強化学習手法として Double DQN[3] を用いる．

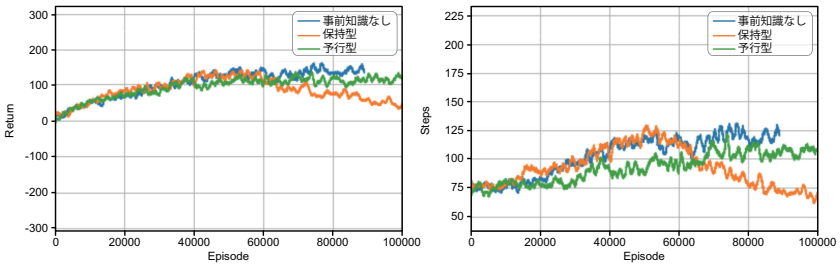
### 4.1 検証結果

図 4 に，エピソード毎の収益の移動平均を示す．報酬条件 1，2，3 の収益はほぼ同じ推移をしていることから，この 3 つの条件には有効な変化が見られない．さらに，報酬条件 4 の収益は，他の条件よりも低く，向上していない．一方，報酬条件 5 の収益は，他の条件よりも推移が変動していることから，一定の行動パターンにかたよらずに様々な行動を獲得できていると考えられる．




(a) 合流車線　(b) カーブ車線　(c) ラウンドアバウト

図 4: エピソード毎の収益の移動平均




(a) 平均収益　(b) 平均生存ステップ

図 5: エピソード毎の平均収益および平均生存ステップ数の推移

表 1: 検証する報酬条件

| | | |
|---|---|---|
| 1 | 報酬 | 他車に衝突または不可侵領域に進入：-10<br>前進：+0.5<br>停止：-0.5<br>目標レーンに存在：+1.5<br>非目標レーンに存在：-1 |
| | 探索法 | Linaer decay epsilon greedy<br>$\epsilon$ 初期値：1.0<br>$\epsilon$ 最終値：0.3 |
| 2 | 報酬 | 他車に衝突または不可侵領域に進入：-10<br>前進：+0.5<br>停止：-1.5<br>目標レーンに存在：+1<br>非目標レーンに存在：-1 |
| | 探索法 | 1 と同様 |
| 3 | 報酬 | 2 と同様 |
| | 探索法 | Linaer decay epsilon greedy<br>$\epsilon$ 初期値：1.0<br>$\epsilon$ 最終値：0.1 |
| 4 | 報酬 | 他車に衝突または不可侵領域に進入：-10<br>前進：+0<br>停止：-0.5<br>目標レーンに存在：+1.5<br>非目標レーンに存在：-1 |
| | 探索法 | 3 と同様 |
| 5 | 報酬 | 他車に衝突または不可侵領域に進入：-5<br>前進：+1<br>停止：-1<br>目標レーンに存在：+1<br>非目標レーンに存在：-2 |
| | 探索法 | 3 と同様 |

## 5. 評価実験

本研究では，車線変更タスクを用いて提案手法の有効性を評価する．

### 5.1 実験概要

本実験では，図 3(a) の直線車線での車線変更タスクを扱う．本タスクは，目標レーンを設定し，他車との衝突を避けながら車線変更を行う．他車が複数台いるため，車線変更のタイミングを見極める必要がある．環境を表した画像から，自車周辺を 100 × 100 ピクセルに切り出したグレースケール画像 4 フレームを入力とする．報酬設計は，検証実験の 5 つ目の報酬条件とする．従来の Experience Replay を使用する Double DQN で学習したものを事前知識なしとし，保持型，予行型と比較を行う．エピソード内で獲得した報酬の総和を収益，エージェントが衝突するまでのステップを生存ステップとする．

表 2: テストスコアの比較

| | 初期レーン | 生存率 [%] | 達成率 [%] |
|---|---|---|---|
| 事前知識なし | ランダム | 53 | 50 |
| | 非目標 | 33 | 29 |
| 保持型 | ランダム | 49 | 37 |
| | 非目標 | 37 | 10 |
| 予行型 | ランダム | **67** | **66** |
| | 非目標 | **51** | **48** |
| 人 | ランダム | 50.7 | 50.3 |

### 5.2 実験結果

図 5 に，エピソード毎の平均収益および平均生存ステップ数の推移を示す．保持型は，平均収益，平均生存ステップの両方が 5 万エピソードをピークに途中から低下した．一方，予行型は，平均収益，平均ステップともに低下することなく学習できた．また，表 2 に，テストスコアを示す．生存率は，300 ステップまでエージェントが衝突しなかった割合である．達成率は，300 ステップ到達時にエージェントが目標レーンに存在した割合である．保持型のテストスコアは事前知識なしと比べて，生存率は同程度だが達成率が低い．これは，事前知識を常に一定の割合で学習に用いるため，仮に事前知識より良い行動を獲得していても事前知識があるため追加できなかった可能性が考えられる．予行型は，従来手法，保持型および人のスコアより生存率，達成率ともに高い結果を得た．先に事前知識のみを用いて学習するため，タスクで有効な基本行動を獲得し，それをもとにさらにタスクに適した行動を獲得できたと考えられる．

## 6. おわりに

本研究では，深層強化学習において安定した制御を獲得するための手法を提案した．提案手法では，事前に人が操作した行動遷移を Replay Buffer に保存し学習を行う．また，直線車線環境で提案手法の有効性を試した．今後の展望として，実機を使用した走行テスト，他環境での実験などが挙げられる．

## 参考文献

[1] V. Mnih, *et al.*, “Human-level control through deep reinforcement learning”, Nature 518, pp. 529–533, 2015.

[2] L. J. Lin., “Reinforcement Learning for Robots Using Neural Networks”, Technical report, DTIC Document, 1993.

[3] H. Hassel, *et al.*, “Deep Reinforcement Learning with Double Q-Learning”, AAAI, 2016.

## 研究業績

[1] 稲盛有那 等, “事前知識を活用した Memory Reinforcement Learning による行動獲得”, 人工知能学会, 2018.