ISSN 2186-5647

7-7

オンライン学習とオフライン学習の違いを体験するロボット教材の開発と評価方法の提案

日大生産工(院)　○和田 将太　　日大生産工　柳澤　一機

1. 緒言

近年，情報系のエンジニアに限らず非専門家でも強化学習を扱うことが増えつつある．しかし，人工知能の理論は授業で学んでいても実際にその理論を使った実機による実験の経験がないという企業人は少なくない[1]．そこで，諏訪はライントレースロボットの制御を通して，強化学習の実環境への適用方法について学ぶことのできる教材の開発を行った[2]．この教材により，シミュレーションだけでなく，実環境においてロボットを制御するから強化学習についての理解をより深めることができる可能性を示した．

諏訪の教材のように，強化学習教材のほとんどが強化学習の方策を作成する際にオンライン学習を前提にしている．

しかし，オンライン学習は，仮想環境の場合はシミュレーションを構築するためのプログラミング技術，実環境の場合はプログラミング技術に加え，ロボットの故障やリアリティーギャップと呼ばれる仮想環境との誤差などへの対応などが必要であり，非専門家が利用することは難しい．

そこで，プログラミング技術の必要性が低く，過去のデータのみで強化学習の方策を作成できる，オフライン学習の一種であるデータ駆動型深層強化学習が注目されている．現在，この手法は仮想環境での活用が中心であるが，実環境での利用について学ぶことができる教材が開発できれば，今後様々な分野で実環境でのデータ駆動型深層強化学習の活用が期待できる．筆者らは，これまでにデータ駆動型深層強化学習を実環境に適用させるための方策に必要な学習方法やデータセットの特徴について体験する教材を提案した[3]．

本研究では，オンライン学習として一般的なQ学習と，オフライン学習としてConservative Q-Learning(CQL)を対象に，ライントレースロボットを題材とし，Q学習を用いて強化学習の基礎を学び，CQLを仮想環境と実環境の両方で体験できる教材を開発する．

教材の開発については，「学習者の動機づけ」に着目したARCSモデルを導入し，学習者の強化学習の学習意欲を高めるよう設計する．

2. オンライン学習とオフライン学習

2.1 オンライン学習とは

オンライン学習とは，エージェント(学習対象)が環境(学習環境)と相互作用し，報酬を得るための方策(エージェントの行動の選択基準)を学習する手法である[4]．本研究では最も一般的な手法であるQ学習を対象とした．

Q学習とは，エージェントがある「状態s_t」において「行動a_t」を決定する際に，常に行動の価値が最大になるような「行動a_t」を選択し，行動した結果から「状態s_t」における「行動a_t」の価値を評価することで「学習」を行う手法である[5]．Q学習では，行動の価値を最適行動価値関数で表し，「状態s_t」における「行動a_t」の価値を$Q(s_t,a_t)$と表現する．この値をQ値と呼び，式を(1)で求めることができる．

$$Q(s_t,a_t) \rightarrow (1-\alpha)Q(s_t,a_t) + \alpha\left(r+\gamma \times maxQ(s_{t+1},a_t)\right) \quad (1)$$

式(1)の右辺第一項は，ある「状態s_t」にてエージェントが 選択する「行動a_t」に対して見積もった価値である$Q(s_t,a_t)$である．式(1)の右辺第二項の先頭にあるαは学習率であり，$0<\alpha<1$の範囲で設定される．αが大きければ，$Q(s_t,a_t)$が更新される量が大きく「学習」は早くなるが，偶然報酬を獲得するなど，例外的な報酬の影響を大きく受けてしまう．γは割引率であり，この値も$0<\gamma<1$の範囲で設定される．割引率γが大きいほど最終的な報酬が大きくなるような行動を取るようになるが，目の前の小さい報酬に反応しない場合がある．

Q学習では「次の状態s_{t+1}」においても，常に最大の価値$maxQ(s_{t+1},a_{t+1})$をもつ「行動a_{t+1}」を選択するため，実際の価値における$Q(s_{t+1},a_{t+1})$は$maxQ(s_{t+1},a_{t+1})$と表現される．

2.2 オフライン学習とは

オフライン学習は，これまで集められた経験データ（状態，行動，報酬の記録）を使ってエージェントの学習を行う方法であり，環境と相互作用が発生しない特徴があるため[6]，プログ

Development of robot teaching materials to experience the difference between online and offline learning and proposal of evaluation method.

Shota WADA, Kazuki YANAGISAWA

ラミングの必要性が低く，非専門家でも強化学習を利用できる方法である．

オフライン学習の1種にデータ駆動型深層強化学習がある．データ駆動型深層強化学習はD3RL(Data Driven Deep Reinforcement Learning)[7]と呼ばれ，事前に集められたデータセットを用い，オフラインで強化学習を行うことで，方策を作成する方法である．今回はCQLという学習方法を採用する．

CQLとは，データセットに存在しない状態に対する行動のQ値にペナルティを与えることで，データセットに前例のない行動が選択されることを防ぐオフライン学習を前提としたQ学習である[8]．

3. ARCSモデルとは

ARCSモデルは1984年にケラー(John M. Keller)によって提案されたインストラクショナルデザインである[9]．ARCSモデルのコンセプトは「学習意欲のデザイン」である．動機づけ理論とシステム的な問題解決プロセスを統合し，学習者に「もっと勉強したい」と思わせるため，学習者をいかに動機づけ，学習意欲を高めてそれを維持するかに注力したモデルである．ARCSモデルは，学習意欲に関する概念が，注意(Attention)，関連性(Relevance)，自信(Confidence)，満足感(Satisfaction)の4つの要素から構成される．それらの概念ごとに方略を設定することで，学習者が学習意欲を刺激・保持することを目指している．

4. 開発したロボット教材

今回開発した教材は，ライントレースを題材とし，①仮想環境上でオンライン学習のQ学習を用いて強化学習の重要な3要素である「状態」「行動」「報酬」について理解を深めること，②オフライン学習のCQLを利用する際に，どのようなデータセットが必要なのか体験から学習することの2点を目的とした教材である．

②については，仮想環境において作成した方策を実環境へ適応した時に，汎用性の高い方策を作成するために必要なデータセットの特徴を学ぶことができるよう設計した．

本研究では表１のようにARCSモデルとの対応を意識して教材を開発した．動機づけを強くするため，注意(Attention)，関連性(Relevance)に相当する要素として，ゲームコントローラを用いロボットを操作することで教材に関する注意を引き，プログラミング能力が高くない学習者でも強化学習に挑戦できるので，関連性(Relevance)から自信(Confidence)に発展させる．

Table１ ARCS モデルと教材の関連

項目	教材との対応
注意 Attention	ゲームコントローラで操作可能なロボット （ロボットを操作できることの面白さで教材に関する注意を引く）
関連性 Relevance	プログラミング能力が高くない学習者でも 強化学習を使ってロボットを制御に挑戦できる
自信 Confidence	汎用性が高い方策を作るためのデータセットの 作り方について、体験から理解できた
満足感 Satisfaction	学んだことを他のことに応用してみたい、実践してみたいと思った

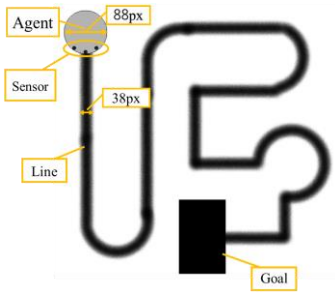

Fig.1 仮想環境上のライントレースコース

学習者が自発的に教材を使用することで自信(Confidence)から満足感(Satisfaction)への達成が期待できる．

4.1 仮想環境の設定条件

本研究では使用したコースは図1に示す．ラインの太さを38pxとし，エージェントの大きさは88pxとなっている．オンライン学習とオフライン学習で同じライントレースコースを設定した．本教材で使用する「エピソード」「状態」「行動」「報酬」は以下のように定義した．

・エピソード

エピソードとは，エージェントが行うタスクの開始から終了までの期間である．本研究では，図１のエージェントの初期位置からゴールに到達するまでを1エピソードと定義した．なお，学習者は任意のエピソード数で方策を作成するための学習を行うことができる．

・状態

状態は，エージェントのセンサの個数と読み取った値から求められる．初期設定では，センサの数を3個とした．センサで読み取った値は0から1024の範囲に設定されており，閾値を設け，測定値が512以下の場合は黒，513以上の場合は白の判定する．学習者は，この閾値を変更し，段階的に設定することで状態を増やすことが可能である．

・行動

行動は，直進，右折，左折の3パターンを基本とし，直進で進む場合は4px，右折は10°回転してから4px，左折は-10°回転してから4px直進とした．これらの1回の行動を1ステップと定義した．学習者は，1ステップ毎の移動量

や旋回量，行動のパターン数を任意に設定することが可能である．

・報酬

報酬は，初期設定として，エージェントの中央センサがライン上にある場合+1の報酬を与え，左右のセンサがライン上にある場合-0.1報酬を与え，完全にラインから外れている場合-1の報酬を与えることとした．学習者はこの値を自由に設定することが可能である．

4.2 仮想環境上のオンライン学習とオフライン学習の条件

本研究では，オンライン学習の初期設定では「状態」が閾値512以下を黒，513以上を白の2値とし，センサが3つあるので$2 \times 2 \times 2 = 8$通りになる．学習者は初期条件でオンライン学習を行った後，例題に沿って「状態」「行動」「報酬」を変化させ，それぞれ変更した際にライントレースロボットが動作する様子がどう変わるのか体験させ，強化学習の基礎を学ぶ．

オフライン学習では，CQLを用いた方策の作成をd3rlpyというPython用オフライン深層強化学習ライブラリを使用して行う[6]．オフライン学習では，学習者はゲームコントローラの「ワイヤレスホリパッド for Nintendo Switch」のアナログスティック操作によって1ステップごとにエージェントの「行動」を手動操作により決定した．

エージェントの目的を達成するまでの一連の「エピソード」「状態」「行動」「報酬」のデータをCSVファイルとして収集する．このファイルの情報からCQLを用いてライントレースロボットの制御のための方策を作成する．その後，仮想環境上で学習した方策によって制御されるロボットの挙動を観察することで，用意するデータセットの「状態」「行動」「報酬」にどのような工夫が必要なのかを学ぶ．

4.3 実環境の設定条件

本研究で使用した実環境のコースを図2，製作したエージェントを図3に示す．実環境の設定条件を仮想環境と合わせるために，ラインの太さを65.0mmエージェントの直径を150mmとした．「状態」は，図1のエージェントと同様の位置にフォトリフレクタを設置することで条件をそろえた．仮想環境と同様，ラインの中心から外側に掛けて黒を0，白を1023とした1024段階のグラデーションになっている．「行動」については，エージェントに設置したDCモータの出力をPWM制御で調整することで，シミュレーションと同じように移動できる．

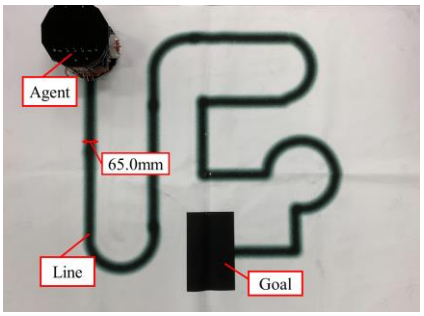

Fig.2 実環境でのライントレースコースの様子

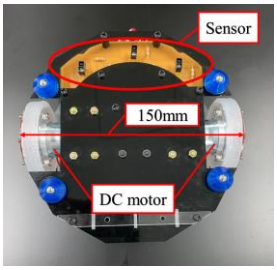

Fig.3 実環境のエージェントの様子

4.4 実環境上のオフライン学習の設定

オフライン学習で作成した方策の実環境への適用方法の理解を促す要素では，実環境でコースを完走するために，どのような特徴を持つデータセットを用意する必要があるか，例えば正の報酬が多いデータセットが望ましいのか，反対に負の報酬が多いデータセットのほうが，汎化性が高い方策を作成できるのかなど，様々なパターンのデータセットを作成してもらい，実環境でのロボットの挙動を観察することで，オフライン学習の実環境への適用方法の理解を深めてもらう．

5. 開発したロボット教材の評価

本研究で開発したライントレースロボット教材を用いて評価実験を行う．実験参加者は機械工学科３年生を対象に授業を行う予定である．

1コマ目は強化学習の基礎である「状態」「行動」「報酬」の説明やQ学習の基礎を説明した後，仮想環境でオンライン学習を体験してもらう．その後，センサの閾値を変更し「状態」の変更によるQ値の変化や，「行動」「報酬」の変化によるエージェントの学習の変化を確認してもらう．

2コマ目はデータ駆動型深層強化学習の基礎を説明し，仮想環境でエージェントを手動操作し「エピソード」「状態」「行動」「報酬」のデータセットを収集し，オフライン学習により方策を作成する．作成した方策でライントレー

スが可能か仮想環境上で学習結果を確認し, 成功させるにはどのようなデータセットが必要なのか学ばせる.

本研究では, 作成した方策を定量的に評価するためのスコアを設けた. ライントレースロボットが走行する際に, 最も望ましい状態は常にロボットの中央のセンサが黒を検出する状態であるため, 中央のセンサが黒を検知した際に加点されるスコアを設定した．ラインを50ステップ見失った場合スコアの加算を中断した. このスコアの累計値から学習結果を定量的に評価することができる.

3コマ目は，2コマ目までに使用したライントレースを実環境で取り組んでもらう．2コマ目に仮想環境で作成した「方策」を図3のエージェントに適用し, 実環境でライントレースを完走するために, どのようなデータが必要なのか確認してもらう.

授業終了後にはアンケートによる授業内容の理解度評価を行う．アンケート内容を表2に示す. アンケートはARCSモデルに基づいて作成し，回答は5件法とした．各設問とARCS概念は表3の通りである．併せて仮想環境，実環境のそれぞれの理解度について自由記述の解答も行ってもらう.

6. 結言

本研究では，ライントレースを題材に，オンライン学習のQ学習を用いて強化学習の基礎を学び，オフライン学習のCQLを仮想環境と実環境の両方で体験できる教材を開発した. ARCSモデルに基づき，学習者が強化学習の学習意欲を高めるよう設計した.

具体的には, 仮想環境でオンライン学習であるQ学習を用いて強化学習の基礎である「状態」「行動」「報酬」の概念を学ぶ内容とした. オフライン学習では, 仮想環境で手動操作により取得したデータセットを用い, d3rlpyを用いてCQLで方策を作成した．仮想環境・実環境の双方において, 汎化性が高い方策を作成するためには, どのようなデータセットが必要なのか体験から学ぶことができる教材を開発した.

今後は，本教材を使用し，ARCSモデルに基づいたアンケートを基に，本教材の教育効果，学習者の意欲について検証する予定である.

Table 2 授業アンケートの項目

	アンケート項目
Q1	この講習会を受けてQ学習に興味を持ち, 使ってみたいと思いましたか？
Q2	この講習会を受けてデータ駆動型深層強化学習に興味を持ち，使ってみたいと思いましたか？
Q3	教材の内容は、あなたの期待や目的に合っていましたか？
Q4	この教材はご自身にとって役に立つと思いますか？
Q5	データ駆動型深層強化学に必要なデータを理解出来ました
Q6	正の報酬の割合を意識しながらデータ収集をうまくこなせたと思いますか？
Q7	データ駆動型深層強化学習を実環境で上手くこなせたと思いますか？
Q8	データ駆動型深層強化学習の理解度は？
Q9	この授業への満足度は？
Q10	授業を受けて人工知能技術が好きになりましたか？
Q11	本教材について意見・要望などがあればよろしくお願いします。(自由記述)

Table 3 各設問とARCSの対応

注意 Attention	関連性 Relevance	自信 Confidence	満足感 Satisfaction
Q1,Q2	Q3,Q4	Q6,Q8	Q7,Q9,Q10

参考文献

1) 中川友紀子, 企業における AI／ロボット教育・開発としての競技会活用のすすめ, 日本ロボット学会誌, Vol.38 No.9 (2020) pp.825-828

2) 諏訪晃也, 仮想環境と実環境を融合した強化学習ロボット教材の開発, 日本大学大学院生産工学研究科修士論文（2021）.

3) 和田将太, 柳澤一機, データ駆動型深層強化学習を活用したロボット教材の提案-実環境でロボットを制御するための検討-, ヒューマンインタフェースシンポジウム (2023), 7T-D12

4) 斎藤康毅, ゼロから作るDeep Learning4-強化学習編 ; 株式会社オライリー・ジャパン, (2022) p.313.

5) 牧野浩二, 西崎博光, Python による深層強化学習入門 -Chainer と OpenAI Gym ではじめる強化学習-, 株式会社オーム社, (2018) p.88

6) 斎藤康毅, ゼロから作るDeep Learning4-強化学習編 ; 株式会社オライリー・ジャパン, (2022) p.312.

7) Seno Takuma, and Michita Imai. "d3rlpy: An offline deep reinforcement learning library." (2022) The Journal of Machine Learning Research Vol.23, No.315, pp.1-20 .

8) A Kumar, A Zhou, G Tucker, S Levine " Conservative q-learning for offline reinforcement learning." Advances in Neural Information Processing Systems Vol.33 (2020) pp.1179-1191.

9) J.M.ケラー, 学習意欲をデザインする : ARCSモデルによるインストラクショナルデザイン. 北大路書房, (2010) p.351