ISSN 2186-5647

7-1

自律型ロボットを用いたゲーム性のある強化学習教材の開発と評価方法の検討

日大生産工(院)　○和田 翔　　日大生産工　柳澤 一機

1. 緒言

近年, 人工知能技術は様々な実問題に適用されつつあり, 人工知能技術とロボティクスの融合が進んでいる[1]. そのため, ロボットと人工知能の知識の双方を有する技術者の教育が重要である. しかし, 人工知能教材のほとんどはシミュレーションのみを取り扱っており, 実環境への適用について考慮されていない. そのため, 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[2]. そこで, 人工知能の知識に加え, 実環境への適用方法についても学ぶことのできる教材が求められている.

この問題を解決するために, 諏訪はライントレースロボットの制御を通して, 人工知能の1つである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[3]. この教材により, シミュレーションだけでなく, 実環境へ適用させロボットを制御することで, 強化学習についてより理解できる可能性があることを示した.

また, 筆者らは対戦型のゲームを用いて, 人工知能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シミュレーション部分の開発を行った[4]. 2台のロボットで鬼ごっこをさせることで, 強化学習について理解を促すものとなっており, ゲーム性のある教材とすることで楽しみながら学ぶことのできるものとなっている. ここで, ゲーム性とはルールの中で自由度やリスク, 再現機会があることから, 個人差はあるものの, 面白さ, やりがいが感じられものとする[5]. しかし, どちらの教材も定量的な結果がなくシミュレーションの様子のみを見て学習がうまくいっているかを判断することしかできなかった. そこで, 他者あるいはそれまでの自分自身と競い合うことができるように, ロボットの動作を評価するスコアを追加し, 定量的な結果を得る必要があると考えた.

村川らはゲームに勝つための方略を考えることで, 学習の動機付けや継続的な学習意欲が向上する[6]ことを示しており, スコアを競い合うことで学習の意欲向上にも繋がる可能性がある. そのため, シミュレーションの結果をスコアで評価し, 他者と競い合うことでゲーム性を強化することを提案する.

著者らは強化学習の手法としてQ学習とDeep Q Network(DQN)という2つを用いて, ロボットをシミュレーションと実環境の両方で制御する教材の開発を行う[7]. ロボットは, 障害物を避け, 目標を捕まえるものとなっており, 目標を捕まえるまでの移動経路からスコアを算出し, 他者とそのスコアを競うゲームを題材とすることで強化学習の理論と実環境への適用方法の両方を学べる教材となっている.

本研究では, 開発した教材の内容とその教材の評価方法について記述する.

2. 強化学習

強化学習とは, エージェント(学習対象)が環境(学習空間)と相互作用しながら集めたデータを使い報酬を得る方法を学習する機械学習の手法である[8].

強化学習では, 環境から状態s_t(環境から得られるエージェントの情報)を取得し, エージェントが決められた行動a_t(環境内のエージェントの行動方法)の中からランダムに行動することで状態s_{t+1}に変化する. 状態s_{t+1}がよい状態であれば正の報酬$(+r)$が与えられ, 悪い状態であれば負の報酬$(-r)$を与えられる. この一連の流れを行う回数をステップ数といい, よりよい動作を獲得するまでの試行回数をエピソード数という. 強化学習は, 複数のステップ数とエピソード数から成り立っており, 学習を繰り返し行うことでよりよい動作が取れるようになる.

本研究では, 強化学習の手法の中で, 一般的なものであるQ学習とDQNの２つを対象とした教材の開発を行う.

2.1 Q学習

状態s_tに対するエージェントの行動a_tの価値を表す関数を行動価値関数$Q(s_t, a_t)$(Q値)と呼ぶ. Q学習とは, Q値が最も大きな行動を選択し, 目的を達成できるようにQ値を更新する

Development of and Consideration of evaluation methods of game-like reinforcement learning materials using an autonomous robot

Sho WADA, Kazuki YANAGISAWA

手法である．Q学習におけるQ値は次の(1) 式で表される．

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

ここで，αは学習率，γは割引率を表す．αは学習の更新速度を表し，この値が大きい場合には学習が早くなり，小さい場合には最適解を導きやすくなる．$maxQ(s_{t+1}, a_t)$は将来の得られる可能性のある最大の報酬を表している．そのため，γは$maxQ(s_{t+1}, a_t)$は将来の報酬の価値を決める値である．この値が大きいほど遠回りでも大きな報酬を求めるようになり，小さいほど目先の報酬に飛びつくようになる．

Q学習ではQ値を図1のようにQ-tableと呼ばれる各状態に対する行動をまとめた表で表現することで学習を行っていく．しかし，Q-tableでは状態や行動の数が大きくなるほどQ値を更新し，学習するまでに時間がかかる．そのため，状態と行動の数が多い複雑な問題にQ学習を扱うのは現実的ではなく，図2のように深層学習を利用してQ値を更新していくDQNが用いられる．

2.2 DQN

DQNとは，人の脳の神経細胞を模倣した数理モデルであるニューラルネットワークを基にした深層学習とQ学習を組み合わせた学習法である．

深層学習は図2のように入力層，中間層，出力層の3層から成り立ち，入力層で外部からのデータを受け取り，中間層でそのデータの特徴量を抽出し，出力層でその結果を出力する．その出力値から自動で更新を行い，最適な出力値を求める．

DQNでは，入力層に状態$\boldsymbol{s_t}$を構成する成分$s_a, s_b \cdots s_n$，出力層に状態$\boldsymbol{s_t}$の各行動のQ値を当てはめ，学習の更新を行うことでQ値の最も大きな行動が最適な行動となる．全ての状態を考慮せずに現在の状態から最適なQ値を導くことができるため，状態と行動の数が多い複雑な問題でも学習を行うことができる[8]．

3. 開発したロボット教材

3.1 教材の概要

本教材は，Q学習とDQNの2つの学習法を用いて，障害物を避けながら目標を捕まえるロボットの制御を行い，その移動経路から算出したスコアを競い合うゲームを題材とする．スコアは(目標を捕まえるまでにかかったステップ数)×(1ステップの移動量)とする．そのためスコアが小さいほどスコアが小さいほど，無駄な

		Action				
		a_0	a_1	⋯	a_t	⋯
State	s_0	$Q(s_0, a_0)$	$Q(s_0, a_1)$	⋯	$Q(s_0, a_1)$	⋯
	s_1	$Q(s_1, a_0)$	$Q(s_1, a_1)$	⋯	$Q(s_1, a_t)$	⋯
	⋮	⋮	⋮	⋱	⋮	
	s_t	$Q(s_t, a_0)$	$Q(s_t, a_1)$	⋯	$Q(s_t, a_t)$	⋯
	⋮	⋮	⋮		⋮	⋱

Fig.1 Q-table[5]

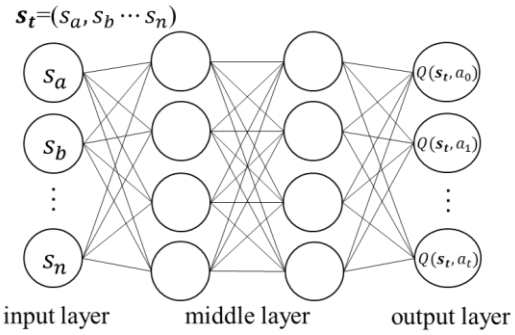

Fig.2 深層学習[5]

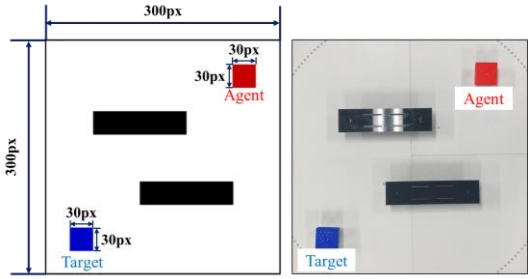

Fig.3 開発した教材の環境[5]

動きが少なく，短い経路で捕まえたことを意味する．

図3のような環境で，目標位置である青い正方形のロボット(ターゲット)が移動し，赤い正方形のロボット(エージェント)が黒い長方形の障害物を避けながらそのターゲットを捕まえる問題のスコアを改善させる内容となっている．

3.2 教材の内容

強化学習はその性質上トライアンドエラーを繰り返すシステムとなっており，全ての学習を実環境で行うということは現実的ではない．そのため，シミュレーションで学習してから実環境のロボットに適用させる必要がある．そのため，教材はシミュレーションを用いて，強化学習の理論の理解を促す要素(シミュレーションの要素)と実環境への適用方法の理解を促す要素(実環境の要素)の2つに分かれる．本教材

はシミュレーションで学習を行い, その学習した結果を基に実環境で制御を行う. 以下にシミュレーションの要素と実環境の教材の詳細について記述する.

3.2.1　シミュレーションの要素

シミュレーションの要素では, 障害物の有無とターゲットの動作を変えていくレベル1～4の問題に取り組む内容となっている. レベル4が最終問題の問題となっており, レベル3までで学んだ知識を使い, スコアを改善し, 最適な移動経路をとることができるように行動・状態・報酬などを変えていく. スコアを他者と競いながら学ぶことでQ学習とDQNの違いや強化学習の行動･状態･報酬について理解を促す. レベル1～4の内容は以下のようにする.

レベル1の問題は障害物がなく, ターゲットが移動しない環境で学習を行う. 行動は上下左右の4方向に30px移動する4種類とし, 報酬はエージェントとターゲットが捕まえた時に正の報酬(+1)を与える. Q学習の状態である離散化した相対座標と絶対座標とDQNの状態である離散化していない絶対座標の3種類で学習を行う[5]. まず, Q学習の相対座標と絶対座標の学習に必要なエピソード数の比較を行う事で強化学習の状態の概念について理解を促し, さらにDQNに学習法を変え, スコアの比較することで, Q学習とDQNの学習法の違いについて理解を促す.

レベル2の問題は障害物がなく, ターゲットが移動する環境で学習を行う. レベル1のDQNで学習したデータを使用し, 動いているターゲットも捕まえることができることを確認し, 強化学習の汎用性について理解を促す.

レベル3の問題は障害物があり, ターゲットが移動しない環境で学習を行う. 障害物にぶつかった時に負の報酬(-0.2)を与えるようにし, 学習を行う. 壁を避けながらターゲットを捕まえられることを確認し, 報酬について理解を促す.

レベル4の最終問題は, 障害物があり, ターゲットが移動する環境で学習を行う. レベル3の学習データを使い, レベル2と同様に捕まえられることを確認し, さらに最適な行動を取る方法について考察し, より良いスコアを出せる方法を検討してもらう.

レベル1～4の問題について3時間の授業で説明を行い, 強化学習の理論について理解を促す. その後, レベル4の問題をこれまでに学んだことを活かし, 2週間, さらに学習者が考えたことを試しながらスコアを改善し, 他者と競い合う事で強化学習の知識をより深めてもらう.

3.2.2　実環境の要素

実環境のロボットはソニー株式会社のtoio[9]を使用する. toioは, コアキューブと専用のマットを使用する. 専用のマットには目に見えない特殊な印刷が施されており, コアキューブの裏側の読み取りセンサーを使用し, マット上でキューブを動かすことで自己位置推定を可能としている.

実環境の要素では, シミュレーションで学習した最終問題のスコアが最も良かった方策を用いてtoioを動かし, 実環境への適用方法について理解を促す. シミュレーションと実環境のスコアを比較することで, 見た目の違いだけでなく定量的に評価を行えるようにした. 実環境でスコアをシミュレーションと同じ値に近づけるために, 環境をどう変えればいいのかを考えてもらい, シミュレーションと実環境が可能な限り同じ条件になるように設定することでスコアもシミュレーションと同程度の値になることを確認してもらう.

また, 実環境では, エージェットがターゲットを捕まえた後に次のエピソードに移行する際にターゲットをランダムの位置に移動させる必要がある. その際, シミュレーションの場合は座標指定のみで設定は可能だが, 実環境では障害物やエージェントを避け座標指定した場所に移動する必要がある. そういったtoioの移動方法を見ることでシミュレーションとの違いを体験してもらい, 実環境へ適応するためにどういった工夫を行う必要があるのか学んでもらう.

4.　検証実験の方法

開発した教材を用いて機械工学科の3～4年生, 大学院生を対象に授業を行う. 授業はグループAとグループBに分けて行う. グループAでは4.1節で記述した通り, スコアを示し, 定量的な結果の変化を示しながら授業を行う. グループBはスコアを表示せず, これまで作成してきた教材同様にシミュレーションの様子のみで結果を判断し, 授業を行う. また, 最終問題について取り組んでもらう間, グループAにはスコアを競い合ってもらうため条件とスコアを共有し合える環境を用意し, スコアをより良くしてもらい, グループBは個々で最終問題に取り組んでもらうようにした. 2つのグループの比較を行う事で評価を行う. すべての問題終了後に行う事後アンケートと最終問題に2週間で取り組んだ時間とそのスコアとする.

事後アンケートの内容を表1に示す．シミュレーションの要素と実環境の要素のそれぞれの理解度についてそれぞれ3項目(Q1～Q3，Q5～Q7）と強化学習について興味を持ったか(Q9)を4件法の選択肢で解答してもらう．また，2つの要素の理解度には自由記述(Q4，Q8)の解答も行ってもらう．また，レベル4の問題に2週間で取り組んだ時間についても1日あたりの平均時間(Q10)を解答してもらい，学習時間の把握を行う．

レベル4のスコアは，最終問題の学習成果の中でグループAには最も良いスコアが出た学習データを提出してもらい，グループBにはスコアについての説明がないため，シミュレーションの様子から最も良い動きができていた学習データを提出してもらう．そのデータからスコアを求め，比較を行う．

スコアを用いるグループAとスコアを用いないグループBを事後アンケートと結果の比較を行う事で，定量的な結果が得られることによる教材の内容への理解向上や，ゲーム性を用いることによる教材への意欲が得られることを示すことができると考える．これらの結果から教材の有効性を示す．

検証実験の結果については，口頭発表時に説明する予定である．

表１　事後アンケートの内容

シミュレーションの要素の理解度	
Q1	強化学習の行動・状態・報酬について理解できた(レベル1･レベル3)
Q2	Q学習とDQNの違いについて理解できた(レベル1)
Q3	強化学習の汎用性について理解できた(レベル 2)
Q4	自由記述
実環境の要素の理解度	
Q5	シミュレーションを実機への適用方法について理解できた
Q6	二つの環境を同条件にする必要があることを理解できた
Q7	シミュレーションと実環境の違いについて理解できた
Q8	自由記述
Q9	強化学習について興味を持てたか
Q10	1 日あたり何時間ぐらいシミュレーションを行ったか(レベル 4)

5. 結言

本研究では，障害物を避け，目標を捕まえるロボットを用いてスコアを他者と競うことでQ学習とDQNの理論と実環境への適用方法の両方を学べる教材の開発を行った．シミュレーションで強化学習の理論の理解を促し，実環境でtoioを動かすことで実環境への適用方法の理解を促す教材となっている．

スコアを表示して授業を行うグループとスコアを表示しないで授業を行うグループに分かれ，事後アンケートとスコアの結果を比較することで教材の評価を行った．

参考文献

1） 比戸将平，人工知能技術のロボット産業応用，日本ロボット学会誌，Vol.35 No.3（2017） pp.186-190.

2） 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

3） 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）．

4） 和田翔，柳澤一機，深層強化学習を用いた自律型ロボット教材の提案，情報処理学会第84回全国大会（2022）4ZJ-04

5） 天野秀樹，水元千秋，北臺 如法，主体性を伸ばす箱ひげ図教材の開発 ―ゲーム性のある教材に着目して―，中学教育第51集 pp35-40

6） 村川弘城，白水始，鈴木航平，ゲームにおける方略の振り返りが動機付けに及ぼす効果―カードゲーム型学習教材「マスピード」を例に―， 日本教育工学会論文誌，Vol.37（2013）pp.109-112.

7） 和田翔，柳澤一機，自立型ロボットを用いたゲーム性のある強化学習教材の開発，ヒューマンインターフェースシンポジウム（2022）2T-D4

8） 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）

9） 小さなキューブ型ロボットトイ･toio(トイオ)，ソニー株式会社，https://toio.io/