4ZJ-04

深層強化学習を用いた自律型ロボット教材の提案

和田 翔† 柳澤 一機‡
日本大学大学院生産工学研究科† 日本大学生産工学部機械工学科‡

1. 諸言

深層学習によるロボットの認識技術向上，およびロボティクスと関連が深い強化学習と組み合わせた深層強化学習の物理世界における適用が大きな役割を果たし，人工知能技術とロボティクスとの融合が進んでいる[1].

そのため，ロボットを作るための知識と人工知能の知識の両方を有する技術者の教育が重要である．しかし，人工知能教材の多くは，シミュレーションのみを取り扱っており，ほとんどの場合，物理世界への適用について考慮されていない．Mirko Kovac はこの数十年で人工知能や機械学習は格段に進歩したが，それに対してロボットの構造や材料の開発などは遅れていると指摘しており，人工知能に適応するロボット工学をフィジカルな AI と定義し，フィジカルな AI の学習・実践の必要性を説いている[2].

本研究では，自律型ロボットを深層強化学習で制御することで，深層強化学習の理論と物理世界への適用方法の両方を学ぶことのできる教材の開発を目指す．そのために，(1)シミュレーションを用いて，深層強化学習の基本的は理論とプログラミングの理解を促す教材と(2)シミュレーションで学習したデータを用いて実機を動かすことで，シミュレーションと実機の違いが学ぶことができる教材の開発を行う．
本稿では，(1)の教材の説明を行う．

2. 強化学習

強化学習とは，よい状態と悪い状態だけを決めておいてその過程を自動的に学習し，よりよい動作を獲得する問題などに用いられる機械学習の一手法である[3].

強化学習は，学習環境から状態s_tを取得し，学習対象であるエージェントが決められた行動a_t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が与えられ，悪い状態であれば負の報酬を与えられる．

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Q値と呼び，各状態における行動に対する Q 値をまとめた表を Q-table と呼ぶ．Q 学習とは，Q-table を用いて 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値を更新する手法である．Q 値は正の報酬が与えられると大きなり，負の報酬が与えられると小さくなるため，Q値の最も大きな行動が最適行動となる．

2.2 DQN

強化学習とニューラルネットワークを基にした深層学習を組み合わせた学習法を深層強化学習という．深層強化学習の中で最も一般的な手法を Deep Q-Network(DQN)という．DQN は Q 学習と深層学習を組み合わせてQ値を学習する手法である．Q 学習では Q-table で表現されていた Q 値を DQN ではニューラルネットワークを用いて表現することでより複雑な状態を用いて学習することが可能である．

3. 開発したロボット教材

本教材では，2 台のロボットで鬼ごっこを行うことを題材とする．この 2 台を Q 学習と DQN を用いてシミュレーション上で制御を行うことで，Q 学習と DQN の基礎とそのプログラミングの理解を促す教材となっている．

3.1 基本設定

Fig.1 は鬼ごっこのシミュレーションの座標を離散化した様子である．赤い丸を鬼(Agent)，青い丸を逃げるロボット(Target)とし，ロボットの半径は20px，フィールドの大きさ400×400pxとした．鬼ごっこは黒く覆われた壁の内側で行う．ロボットの行動は上下左右の 4 方向に 40px 移動する 4 種類とする．

次に状態は Q 学習と DQN によって変わる．Q 学習の場合は Q-table を使用するため状態の総数を小さくする必要がある．そのため，Fig.1 のようにAgentとTargetの座標を40pxずつ10分割し，離散化を行う．その離散化した Agent の絶対位置を(x_A, y_A)，Target の絶対位置を(x_T, y_T)とする．Q 学習の状態は，離散化したエージェントとターゲットの絶対位置(x_A, y_A, x_T, y_T)とした．DQN の場合はニューラルネットワークを用いて状態を

Suggestion of teaching materials for autonomous robots using deep reinforcement learning

†Sho Wada, Department of Mechanical Engineering, Graduate School of Industrial Technology, Nihon University
‡Kazuki Yanagisawa, Department of Mechanical Engineering, College of Industrial Technology, Nihon University

表現するため座標の離散化する必要がない．そのため，状態はエージェントとターゲットの各座標(X_A, Y_A, X_T, Y_T)とした（フィールドサイズが400px×400pxのため，0~400の値入力した）．

最後に，報酬はAgentとTargetが接触した時にAgentに正の報酬，Targetに負の報酬を与え，Targetは壁に接触した時にも小さな負の報酬を与える．

3.2 教材内容

Q学習，DQNいずれも3.1で定義した行動，状態，報酬の条件でシミュレーションを行う．

・1章　Q学習を用いた鬼ごっこの制御

Q学習を用いる際，Q-tableが大きくなるほど最適行動を取るまでに，時間を要することを説明し，状態をAgentとTargetの絶対位置(x_A, y_A, x_T, y_T)にしてシミュレーションを行う．最適行動を取るまでの時間を確認し，行動，状態，報酬とQ-tableの関係の理解を促す．

・2章　DQNを用いた鬼ごっこの制御

Q-tableの代わりにニューラルネットワークを用いて学習するDQNの説明し，DQNを用いてロボットの制御を行う．状態は，エージェントとターゲットの各座標(X_A, Y_A, X_T, Y_T)にする．シミュレーションを通じて，Q学習とDQNを比べることで，複雑な状態を入力する場合は，DQNの方が適していることを理解させる．

4. 評価実験

作成した教材は，機械工学科の大学3年生10人を対象に評価を行った．参加者はPythonによるプログラミングの基本知識とQ学習の基礎の授業をあらかじめ受講している．授業終了後にアンケートを実施し，開発した教材の評価を行った．

Fig.2はAgentを止まっているTargetに対してQ学習とDQNで制御した場合の行動数を比較した結果である．Q学習とDQNそれぞれの学習後のAgentがTargetを捕まえるまでの行動数の平均はQ学習が10.1回，DQNが5.0回であった．DQNのほうが絶対座標という直感的な状態を設定していても，無駄な行動が少なく，効率的にTargetを捕まえることが可能である．このように，Q学習とDQNを比較することで，二つの手法の理解を促す．

アンケート結果については，3月の発表時に紹介する．

5. 結言

本研究では，深層強化学習の理論と物理世界への適用方法学ぶことができる教材の開発を目指した．そのため，シミュレーションを用いて，深層強化学習の基本的は理論とプログラミングの理解を促す教材の開発を行い，Q学習とDQNの基礎とその違いの理解を促した．

今後は，シミュレーションで学習したデータを使い実機を動かすことで，シミュレーションと実機の違いが学ぶことができる教材の開発を行う．

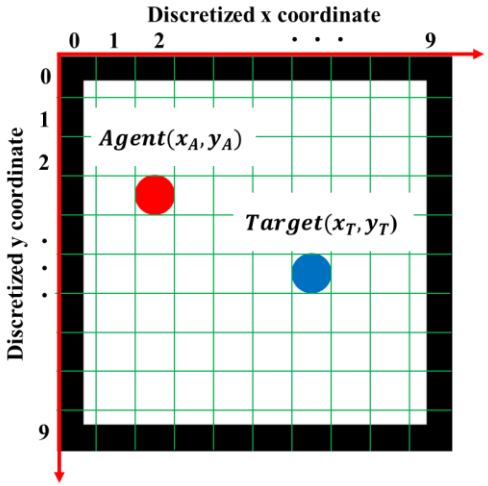

Fig.1　Discrete coordinates of the Tag Simulation

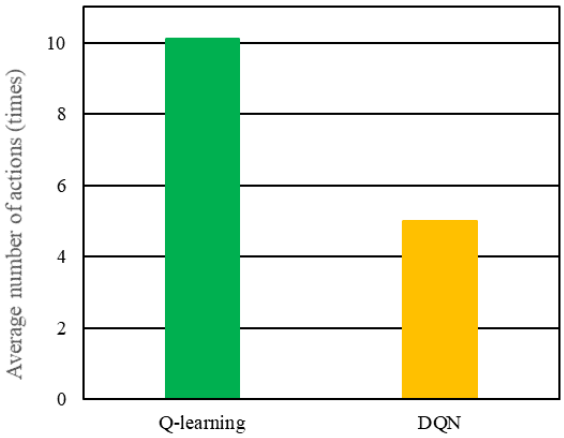

Fig.2　Comparison of the average number of actions between Q-learning and DQN

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，Vol35，No.3(2017)，pp.186-190.

[2] Miriyev，A.，Kovač，M.，Skills for physical artificial intelligence, Nature machine intelligence, Vol.2，(2020)，pp.658-660.

[3] 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)