7N-08

深層強化学習ロボットの実仮想環境統合フレームワークへ向けた検討

小林 賢一 † 辻 順平 † 能登 正人 †
神奈川大学大学院工学研究科電気電子情報工学専攻 †

1 はじめに

近年，深層強化学習という手法が注目を集めており，ロボティクスへの応用研究も始まっている [1]． 一般にロボットへ機械学習を応用する際には，学習のための試行を実施するための仮想環境が必要不可欠である．一方で，仮想環境上で学習されたモデルを実際のロボットに搭載し実行することは，両環境間の差異に伴う困難が生じると考えられるが，特に深層強化学習においてはその影響は十分明らかではない．

本研究では，レゴマインドストームを用いた実際のロボットタスクと OpenAI Gym に基づく仮想環境を連携するシステムの構築作業を通して，実仮想環境統合フレームワークに向けた課題を明らかにすることを目指す．

2 ロボットへの深層強化学習の応用

ロボットへの深層強化学習の応用においては，電池の使用時間や試行回数の課題があり，仮想環境とロボットを連携した学習法が検討されている．仮想環境の構築は，深層強化学習環境用ツールキットの OpenAI Gym や ChainerRL があり容易であるが，仮想環境での実験に特化している．そのため両環境を連携した学習法のためには，両環境のインタフェースの構築が必要である．また，両環境間の差異により，仮想環境の学習した学習モデルを実環境のロボットで用いることは容易ではない．両環境間の差異も明確でない．

3 実仮想環境統合フレームワークの課題調査

実仮想環境を連携した学習を行うために，両環境間の統合が可能な実仮想環境統合フレームワーク必要がある．

本研究では，ロボットを用いた実際のロボットタスクと仮想環境を連携したシステムの構築作業を通して，実仮想環境統合フレームワークに向けて両環境間にどのような差異があり課題となっているのか明らかにすることが目的である．そこで，ロボットタスクを一般的な楕円形型のコースによるライントレースの学習を行う．これは，問題設定を簡素化することで両環境間の差異をより把握しやすくするためである．

まず，仮想環境でエージェントによる事前学習を行う．次に，この仮想環境で事前学習させた学習済モデル（50 episode と 200 episode の 2 種類）を用いて実環境のロボットで周回行動を行う．両環境で 700 step 間の報酬の合計値と 1 step 当たりの黒線上にあるセンサ数（3 個と 5 個の 2 種類）を比較する．図 1 に概要図を示す．本研究のシステムは深層強化学習環境として OpenAI Gym を用い，深層強化学習アルゴリズムの実装にあたっては ChainerRL を用いる．以下に本研究の仮想環境および実環境，深層強化学習の問題設定についての詳細を示す．

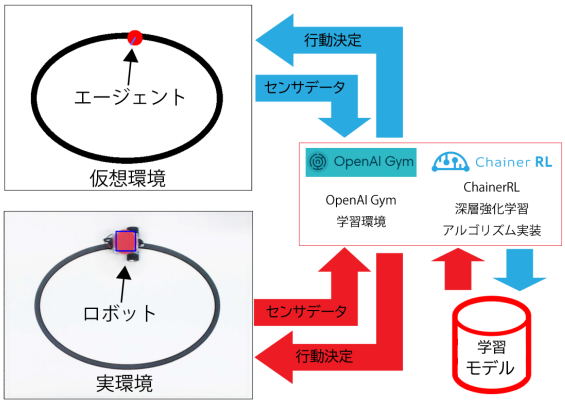

図 1: 仮想環境での事前学習（青矢印時）と実環境でのロボットの自律走行（赤矢印時）

Discussion About Virtual-to-real Integration Framework for Robotics Based on Deep Reinforcement Learning
†Kenichi Kobayashi, Junpei Tsuji and Masato Noto
†Graduate School of Electronical, Electronics and Information Engineering, Kanagawa University

仮想環境

OpenAI Gym と ChainerRL を用いて環境構築を行う．コース中の二値化画像上をエージェントが座標移動を行う．エージェントの位置からセンサの位置座標を設定し，その座標の色を入力として学習を行う．

実環境

OpenAI Gym と ChainerRL を用いて仮想環境の学習モデルを利用したロボットの自律走行を行う．ロボットはレゴマインドストーム EV3 を用いる．センサ値と行動制御信号の EV3 とサーバ間の送受信は，MQTT プロトコルを用いて Wi-Fi 経由で通信を行う．

ロボットタスクと深層強化学習環境の設定

両環境で解決可能な深層強化学習の設定とする．設定の変化による影響は考慮しない．

- 行動：7 行動
  （前進，左 30・45・60 度旋回，右 30・45・60 度旋回）
- 入力（状態）：白黒判別センサ [0 or 1]
- 報酬 (r)
  黒色判別：$0.1 < r < 0.8$, 白色判別：$r = -1$
  位置 $((20 > x > 620), (20 > y > 460))$: $r = -1$
  step 数 >20000 step: $r = 1.0$
- 深層強化学習の設定：隠れ層：50 ユニット × 2 枚

ロボットとエージェントの行動のキャリブレーション

エージェントとロボットの両方で一定区間を単一行動をさせる．この結果が始点と終点の座標が同様になるように調整を行う．これを先述の 7 つの行動で行う．

4 結果および考察

700 step 間の報酬の合計と各 step での黒線上の平均センサ数を表 1 に示す．また，周回行動の軌跡を図 2 に示す．図 2 より，実環境では仮想環境のような線上を走行できていない．さらに表 1 より報酬の合計と平均センサ数が実環境では少ないことからも両環境間の行動に差異があると言える．

構築作業を通して下記のような課題が明らかとなった．ロボット単体のみでの深層強化学習実装は厳しく，サーバを介したシステムの設計が必要であること．ロボットで搭載可能なセンサ位置や各センサ間の距離を考慮した上で，仮想環境のセンサ間の位置を設計する必要があること．また，用いる実際のセンサの出力値と仮想環境の模擬センサで実装可能なものをキャリブレーションが必要であること．ロボットは充電池消費量による電圧変化やモーターの加速度，路面摩擦を考慮する必要があり，それらを考慮した両環境の行動を調整する必要があるなどの課題が得られた．以上の課題は，ライントレースに限定した課題ではなくロボットと仮想環境を連携した際に発生する共通の課題であると考える．

表 1: 報酬の合計値と各 step の黒線上にあるセンサの個数（実：実環境，仮：仮想環境）

	報酬		センサ数/t	
実験条件	仮	実	仮	実
センサ 3 個（200ep）	549.0	451.3	1.41	0.35
センサ 5 個（50ep）	462.2	255.0	1.62	1.07
センサ 5 個（200ep）	602.8	430.4	1.31	0.37

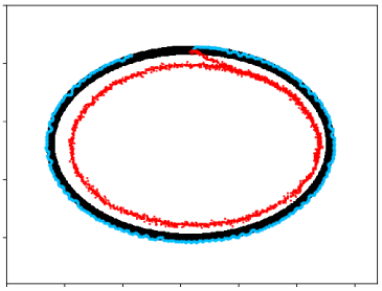

図 2: センサ 3 個・200 episode 学習後・700 step 間両環境の周回行動の軌跡 (青：仮想環境，赤：実環境)

5 おわりに

本研究では，実際のロボットタスクと仮想環境を連携するシステムの構築作業を通し，実仮想環境統合フレームワークに向けた課題を明らかにすることを目指した．結果として，ライントレースのようなロボットタスクでも生じる両環境の行動に差異が生じる多くの課題が得られた．また，本研究で得られた課題はライントレースに限られたものではなくロボットと仮想環境を連携した際の共通の課題であると考えられる．

参考文献

[1] Gu, S., Holly, E., Lillicrap, T. and Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3389–3396 (2017).