人工知能学会研究会資料
SIG-AGI-004-06

実世界で動作する強化学習ロボットを目指して
-学習環境 Re:ROS の試作-

Toward Deep Reinforcement Learning of Robots in the Real World.
-Prototyping of a Simulation Environment “Re:ROS”-

上乃 聖 [1]　大澤 正彦 [2]　今井 倫太 [2]　加藤恒夫 [1]

Sei Ueno[1], Masahiko Osawa[2], Michita Imai[2], and Tsuneo Kato[1]

[1] 同志社大学 理工学部
[1]Doshisha University Faculty of Science and Engineering
[2] 慶應義塾大学 大学院 理工学研究科
[2] Graduate School of Science and Technology, Keio University

Abstract: Deep reinforcement learning has achieved great success in learning to play video games. In contrast to the video games in which the status changes discretely in space and time, robots in the real world move continuously and asynchronously following physical rules. To apply deep reinforcement learning to robot control, we prototyped a robot simulation environment "Re:ROS" with asynchronous system architecture based on Gazebo simulator and Robot Operating System (ROS).

1 はじめに

近年の Deep Neural Network(DNN)の発展により,DNN と強化学習を組み合わせた深層強化学習モデルの適用はビデオゲームのクリアタイム,あるいはビデオゲーム中のスコアにおいて高い性能を示している.深層強化学習モデルの代表的な手法として Deep Q-Learning(DQN), Asynchronous Advantage Actor-Critic(A3C)が挙げられる.文献[1]ではホッケーゲームの Pong,ブロック崩しの Breakout, Space Invaders などの複数のビデオゲームを指す Atari 2600 games において DQN の適用により,従来手法を上回るスコア,ゲームによっては人間を上回るスコアを獲得している.また,文献[2]では A3C の適用により Atari 2600 games において,DQN よりも高いゲームスコアを獲得したことを報告している.

強化学習は学習と意思決定を行うエージェント(Agent)とエージェントの外部に当たる環境(Environment)から構成されている.エージェントと環境は相互作用があり,数値化された報酬を最大化するように学習し環境に適応する学習制御の枠組みを指す.エージェントは環境の状態(State),報酬(Reward)をもとに可能な行動を選択する政策(policy)に従って，行動(Action) を選択する．環境はその行動に応答し,エージェントに新しい状況,報酬を提示する[3](図 1).

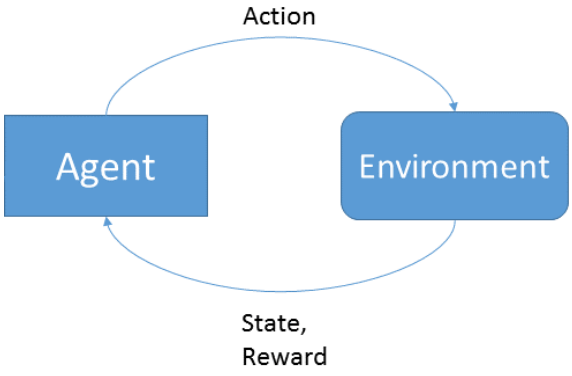

図 1 強化学習の概念図

強化学習の特徴として試行錯誤的な探索(trial-and-error search)が挙げられる.エージェントは過去試みた行動の中で報酬を得るために効果的なものを優先的に選ばなくてはならない.しかし,このような行動を発見するためには,過去に試みたことがない行動も選択してみなくてはならない.つまりエージェントはより報酬を多く得るために,一定の小さな確率で学習結果とは関係なくランダムな行動を選択することがありえる.

強化学習の適用例としてロボットの制御を行う研究もあり,文献[4]では 4 足ロボットの歩行動作を獲得した報告している.また文献[5]では 2 台のアームロボットが協調して荷揚げ作業を行う問題で協調行動を獲得したと報告している.近年では自動で動く災害時の救護用ロボットといったような,より複雑なタスクの解決を行うためにビデオゲームで成果をあげている深層強化学習モデルの適用が期待されている.

しかし,実世界のロボットの制御と,現在深層強化学習モデルが成果を挙げているビデオゲームのタスクには環境の差異が 2 点ある.1 つは物理条件である.ビデオゲームに設定された物理条件は動作のために単純化したものが多い.また操作に対して即時に応答するものも多い.一方,実世界のロボットにおいては重力による加速,摩擦や物体同士の衝突によって生じる複雑な反射や破壊といった物理条件が無視できない.またモータの応答についても即時に応答するわけではなく,命令通りの制御を安定的に行うには待ち時間が存在する.

次に動作の非同期性である. 強化学習が適用されているビデオゲームにおいては 1 動作,あるいは 1 手が明確なものが多く,1 動作,1 手後の環境の状態の変化,報酬の変化は離散的な値として処理される.その際処理にかかる時間や状態の変化にかかる時間も離散的な値として扱われる.一方,実世界においては 1 動作に明確な規定があるわけではなく,動作や応答は非同期的に行われる.

つまり実世界の環境はエージェントの行動に当たるロボットの操作や摩擦等,さまざまな要素が非同期に働くことになる.しかし,既存の強化学習手法は状態や報酬を同期的に返すことを前提に設計されており,ロボットの制御にそのまま適用することは難しい.

ロボット制御に強化学習を採り入れるには実世界の物理的な法則とロボット制御の非同期性を備えたシミュレータの整備が必要である.

実世界に近い強化学習の適用として文献[6]では Unity を用いて,視覚的により現実的な環境でゲーム中のタスクを解決するフレームワークを提供することでより汎用人工知能の設計,テストを行いやすくことを目標としている.ベンチマークとしてエージェントがリンゴを取得するという課題の解決を行った.

また,文献[7]では Open AI の提供している強化学習を設計するためのプラットフォーム gym[8]と ROS, 物理シミュレータ Gazebo を用いて強化学習を適用し,Gazebo 上でロボットが障害物を避けつつ,決められたコースを周回する課題の解決を実現している.

しかし先行研究には強化学習に非同期分散を適用していない。そこで本研究では強化学習にも非同期分散を適用することで実世界と強化学習の間の不整合性を解消するシステムを提案する.ロボット制御のためのミドルウェア Robot Operating System(ROS) と物理シミュレータの Gazebo を用いた非同期分散の強化学習プラットフォームを用いることにより,実世界上の動作を想定したシミュレーションを行うためのフレームワーク Re:ROS を提案する.また,本研究ではフレームワーク Re:ROS を用いてシミュレータで作成したサッカーの課題を例に試作を行う.

2 Re:ROS の構成

2.1 システム構成

本研究のフレームワーク中では実世界上の非同期性に適合するために強化学習の枠組みに非同期分散型のシステムを適用する.

従来の手法(図 1)の場合は環境とエージェントを定義し,最終的に同期をして環境は状態,報酬をエージェントへ送り,エージェントは行動を環境へ送ることになる.

提案手法(図 2)ではロボットが達成するべき目標を達成した,あるいは達成に近いと判断されるような行動をしたとする報酬や,環境中のオブジェクトの移動の検出やロボットのセンサデータの取得する状態があり,それらをまとめて環境とする.報酬や状態は計算に必要なデータをそれぞれ独立して取得する.報酬と状態は取得したデータから計算,処理を行った後,同期を取ることなく発信を続ける.

また行動についても,環境に発信する行動は 1 つに限るわけではなく,ランダムな行動や, DQN などにより決定した行動など,複数存在する場合があり,その場合の各行動もそれぞれ独立して報酬や状態からデータを取得し,計算を行い,同期を取ることなく発信し続ける.

これらの行動の中から政策に従ってエージェントが実際に取る行動を選択する.

このように Re:ROS では強化学習に必要な枠組みのうちで環境,エージェントの処理を非同期分散に行うようなシステム構成をとって,強化学習にかかる処理を行う.

シミュレータに Gazebo を採用し,制御には ROS を採用している.今回採用したソフトウェア ROS, Gazebo についてそれぞれの特徴,また非同期分散型の強化学習にどのような利点があるかについて説明する.

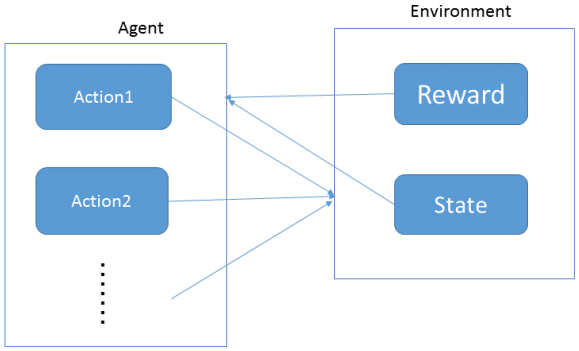

図 2　提案構成

2.2 物理シミュレータ Gazebo

Gazebo とはロボットに関する研究を推進する Open Source Robotics Foundation(OSRF)が開発している物理シミュレータである.

Gazebo の特徴としては以下の 5 点が挙げられる.

(1) 複数の物理エンジン(ODE, Bullet, Simbody, DART)を選択できる.
(2) 3 次元レンダリングエンジン OGRE によって,高品質なグラフィックを提供できる.
(3) カメラ,レーザレンジファインダなど多くのセンサをシミュレートできる.
(4) 全方向に移動できる胴体に 2 本のアームを持ち,カメラやセンサを搭載した自律型ロボット PR2 や Turtlebot(後述)など,ロボットのモデルが多数準備されている.
(5) ROS と開発元が同じ(OSRF)なため,ROS との連携が容易である.

2.3 ロボット制御用ミドルウェア ROS

ROS は Gazebo と同じく OSRF が開発を行っているロボット向けのメタ・オペレーティングシステムである.

ROS の特徴は以下の 3 点にある.

(1) 分散型システムを簡単に構築できることにより,再利用性が高い.
(2) Gazebo 同様,多数のロボットのモデルが準備されている.
(3) Gazebo のみではなく,可視化ツール rviz など多くのツールと連携できる.

2.4 ROS, Gazebo 使用の利点

以上のような特徴を持つ ROS と Gazebo であるが,非同期分散型の強化学習においても以下のような 4 つの利点がある.

(1) 実世界に近い物理条件が定義されている.
(2) 連続的な空間で動く.
(3) ライブラリが豊富である.
(4) 非同期分散システムが作成しやすい.

Gazebo には前述の通り物理エンジンにより現実に近い物理条件で実験を行うことができる.実世界上の動作を考慮するうえでは連続的な環境の変化を取得できることが望ましい.そこで,Gazebo を使用することによって連続的な空間でシミュレートを行い,ライブラリが豊富な ROS を使用できることで細かい制御や多数のロボットをシミュレーション中に採用することができる.

ROS 自体が非同期分散システムの採用をしているため,図 2 のような状態や報酬,行動の処理を非同期分散に設計することができる.

3 試作例

本研究では Re:ROS を用いて,ロボットによるサッカーの課題のうち PK を行うタスクで実験を行う[10].

3.1 PK ロボット

ロボットは Turtlebot を採用する.Turtlebot とは対向 2 輪型の移動台車である Kobuki とモーションセンサ Kinect を搭載した自律走行ロボットであり追加機能の開発が可能である.Turtlebot をエージェントとして制御を行い,ボールを転がして正面にあるゴールを目指す(図 3). 初期位置として Turtlebot はゴールまで約 3m の位置に,サッカーボールはゴールまで約 1.25m の位置に配置をする.リセット時には初期位置を全く同じにならないように設定する.

図 3 設定環境

5Hz で制御を行い,状態は Turtlebot に搭載された Kinect から取得した 60×60 の depth image とする(図 4).行動の出力は, 1m/s で前進,後退と 2rad/s で左に回転,右に回転と停止の 5 種類である.

報酬は

1. ボールがゴールに近づく:0 ~ 1.0 点
2. ゴール: 1.0 点
3. 10 秒以内にゴールできなかった:-1.0 点

の 3 種類とする.

この条件の下で深層強化学習モデルである DQN を適用して,ロボットがゴールを目指してボールを運ぶことができるのかをシミュレーション実験を行った.

ただし,以上の条件のうち,ランダムで出力された行動と DQN により出力された行動,報酬,状態となる depth image をそれぞれ非同期分散に発信する.

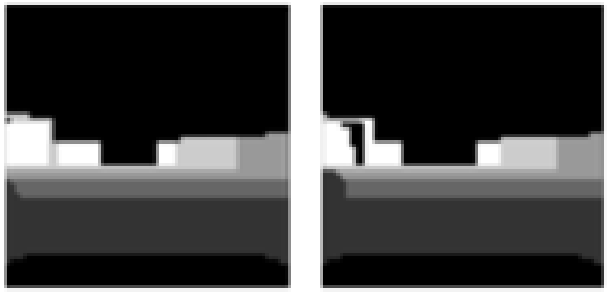

図 4　Kinect から取得する depth image

この学習の結果,学習初期では Turtlebot がボールから離れるような行動も見られ,10 秒間にボールがゴールに入ることは少なかった.しかし,学習後期には Turtlebot がボールから離れる行動は初期ほど見られなくなり,ボールがゴールに入ることもわずかに増えた.

5. まとめと今後の課題

本研究では,実世界のロボット制御の同期処理と強化学習の同期処理の不整合を強化学習の枠組みに非同期分散型のシステムを採用することで対応した,ロボットシミュレーション環境 Re:ROS を試作した.

今後検証する点としては,環境上でゴールキーパーを配置すること,また複数体のロボットの連携を考慮するような実験を行い,実世界上でも同等の実験を行い比較することで,本プラットフォームの実世界の再現性を評価する点である.

参考文献

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller.:"Playing Atari with Deep Reinforcement Learning".In arXiv preprint arXiv:1312.5602 (2013).

[2] Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." arXiv preprint arXiv:1602.01783, (2016).

[3] Richard S. Sutton and Andrew G.Barto 著, 三上貞芳, 皆川雅章訳: “強化学習”.森北出版株式会社, (2000)

[4] 木村元, 山下透,小林重信: "強化学習による 4 足ロボットの歩行動作獲得.", 電気学会論文誌 122 , pp. 330-337, (2002)

[5] 山田和明, 大倉和博, 上田完次: "強化学習による自律型アームロボットの協調行動獲得.", 計測自動制御学会論文集 39.3, pp.266-275, (2003)

[6] Masayoshi Nakamura and Hiroshi Yamakawa.:” A Game-Engine-Based Learning Environment Framework for Artificial General Intelligence Toward Democratic AGI”, Neural Information Processing, Volume 9947, pp 351-356, (2016)

[7] Iker Zamora, Nestor Gonzalez Lopez, Victor Mayoral Vilches, Alejandro Hernandez Cordero.” Extending the OpenAI Gym for robotics: a toolkit for reinforcement learning using ROS and Gazebo”. In arXiv preprint arXiv:1608.05742, (2016).

[8] Greg Brockman et al. “OpenAI Gym”. In: arXiv preprint arXiv:1606.01540, (2016).

[9] 木内祐介: “ROS/Gazebo を用いたロボットシミュレーション”, システム/制御/情報 : システム制御情報学会誌, Vol. 59, No. 2, pp. 65-70, (2015)

[10] 大澤 正彦, 芦原 佑太, 島田 大樹, 倉重 広樹:” 前頭前野 Accumulator を用いた複数の機械学習手法の調停”, 第 4 回汎用人工知能研究会, (2016)