3D1-OS-37a-4in1

深層強化学習を用いた動作制御への基礎的検討

A Basic Study on Action Control Using Deep Reinforcement Leraning

橋本　さゆり [*1]　　小林　一郎 [*2]
Sayuri Hashimoto　　Ichiro Kobayashi

[*1]お茶の水女子大学理学部情報科学科　　[*2]お茶の水女子大学基幹研究院自然科学系
Ochanomizu University　　Ochanomizu University

Deep reinforcement learning, combining reinfrocement learning and deep learning, enables Q-learning in the continuous state space. In this paper, we discuss a basic study on action control, e.g., robots' behaviours, based on deep reinforcement learning. In concrete, we use triple-inverted-pendulum as a target to be controlled, and discuss the feasibility of applying deep reinforcement learning to its action control – we conducted two experiments: one is to invert the triple-inverted-pendulum, and the other is to investigate the ability of action control by changing the size of Experience Replay and mini-batch.

1. はじめに

近年，ロボットや自律運転車の動作制御などに深層強化学習が盛んに用いられてきている．深層強化学習は，強化学習と深層学習を融合し，連続の状態空間における Q 学習を可能にした．本研究では，深層強化学習を用いた動作制御を行い，その手法について考察を行う．具体的には，三重倒立振子を対象にして実験を二つ行なう。一つ目は，深層強化学習を用いて三重倒立振子を倒立させる．二つ目は，Experience Replay のサイズとミニバッチのサイズを変更して動作制御の性能調査を行なう．これらの実験を通じて深層強化学習の基本的な特性を考察する．

2. 深層強化学習

強化学習の手法の一つである Q 学習では，エージェントが状態 s で行動 a をとった時の行動価値を Q 値という値で評価し，ある状態においてこの値が高い行動を選択する方策を学習していく．Q 値の更新は，以下の式で行う．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t + \alpha[r_{t+1} + \gamma max Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (1)$$

Q 学習は Q-table を用いて Q 値の更新を行なうが，この方法では状態が高次元や連続で表現される際に計算コストが高くなる，あるいは不可能になるという問題点がある．そこで，本研究では状態空間を連続として扱うことができる深層強化学習 [**?**, **?**] を用いる．深層強化学習では，深層ニューラルネットワークに状態を入力し，出力をそれぞれの行動の Q 値とする．報酬を環境から受け取ることで行動に関する方策を決定するため事前に正解データを与えておくことがないことから，教師信号として目標値 (r_{t+1}+$\gamma max Q(s, a)$) をある時刻での正解データとして与え，出力との誤差をとり乖離伝搬していくことで学習を行なう．また，状態が一時刻進むと同時に目標値も一時刻進む．深層ニューラルネットワークを用いることで高次元の状態に対しても Q 値を簡単に更新できるようになる．しかしエピソードの単位で学習を行なうと，連続する状態を対象とすることになるため学習に偏りが生じるという問題点がある．そこで Experience Replay という技術が用いられる．Experience Replay とは，過去の $\{s_t, a_t, r_t, s_{t+1}\}$ を全て保存し，そこからランダムに $\{s_t, a_t, r_t, s_{t+1}\}$ をとってきてミニバッチを作成することにより，学習データの偏りを無くす方法である．

連絡先: 橋本さゆり，お茶の水女子大学理学部情報科学科小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，g1320530@is.ocha.ac.jp

3. 深層学習を用いた三重倒立振子の制御

3.1 三重倒立振子

図**??**に制御対象となる三重倒立振子を示す．

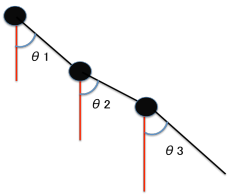


図 1: 三重振子

一般に，振子の制御には複雑な運動方程式を用いて制御を行なうが，本研究では三重振子のそれぞれの角度 θ_1，θ_2，θ_3 とし，状態となる角度の更新が，それぞれの振子の軸に与えられる力 power1，power2，power3 によって変更されるとした基で，強化学習を用いることで以下の単純な振子の角度の更新式を用いて制御を行なう．

$$\begin{aligned} \theta_1 &\leftarrow \theta_1 + power1 \\ \theta_2 &\leftarrow \theta_2 + power2 \\ \theta_3 &\leftarrow \theta_3 + power3 \end{aligned} \quad (2)$$

3.2 深層ニューラルネットワーク構成

深層ニューラルネットワークには，深層学習フレームワークである Chainer を用いてネットワークを構築する．ネットワークは入力層，出力層，中間層 4 層の全 6 層のネットワークである．状態 12 個 (θ_1 の連続する時間の角度 4 つ，同様に θ_2 の連続する時間の角度 4 つ，及び θ_3 の連続する時間の角度 4 つ) を入力とする．出力には振子の軸の動作となる左右の動き 2 つの Q 値とする．また，教師信号として Q 値の目標を設定し，誤差伝播法によりネットワークの結合荷重の学習を行なう．

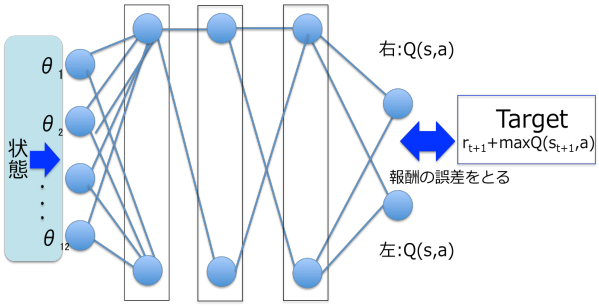


図 2: 深層ニューラルネットワーク

4. 実験

4.1 実験 1

実験 1 では，上述した深層強化学習を用いて，三重振子を倒立させることを目的とする．

4.1.1 実験設定

1 エピソードを 300 回の試行とし，30,000 回エピソード学習させる．高さ 0 は 1 つ目の振子の支点の位置を指す．それぞれの振子の棒の長さを 1 とするため高さは最小で −3，最大で 3 の値をとる．報酬については，高さが 0 より大きい時は高さの絶対値に対して 5 倍の報酬を与え，高さが 0 より小さい時は高さの絶対値に対して −1 倍の報酬を与えた．上述した Experience Replay は 30,000 エピソードのうち上位 100 エピソードのみを保存し，それからミニバッチを生成する．また，本実験では上述した振子のそれぞれの power は power1=± 0.005，power2=∓ 0.005，power3=± 0.005 に設定した．

4.1.2 実験結果

- 1 エピソード学習時（図 3 参照）

1 エピソード学習済みのモデルは以下のようにになっている．高さは −3 から −1 を行き来する状態にあることがわかる．トルクでは，常に 1 の力をかけているが，上手く振子を振り上げられていないことがわかる．

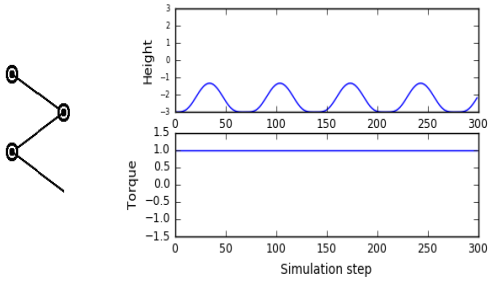


図 3: 1 エピソード学習した結果

- 940 エピソード学習時（図 4 参照）

940 エピソード学習済みのモデルでは，高さが −3 から 3 まで振子を高く上げることができている．しかし，3 まで高さを上げた後その状態を維持することができず，振子が回転してしまっている．

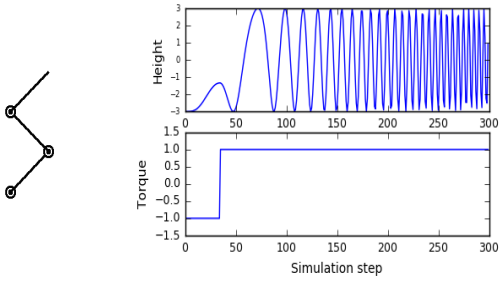


図 4: 940 エピソード学習した結果

- 25,590 エピソード学習時（図 5 参照）

25,590 エピソード学習済みのモデルでは，高さが −3 から 3 まで振子を高く上げてから，倒立した状態を維持できている．

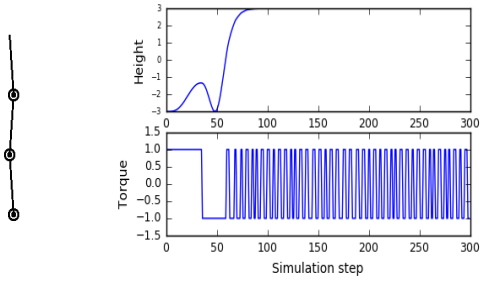


図 5: 25,590 エピソード学習した結果

0 エピソードから 30,000 エピソードまでの総報酬の遷移は図 6 のようになった．初期の総報酬は −1000 に設定している．

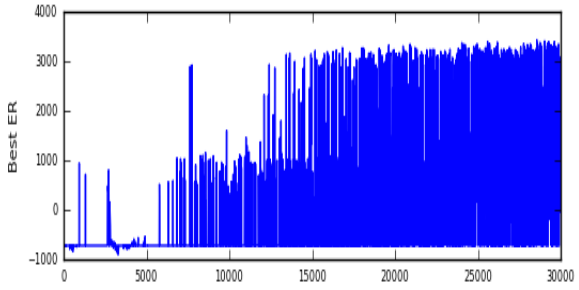


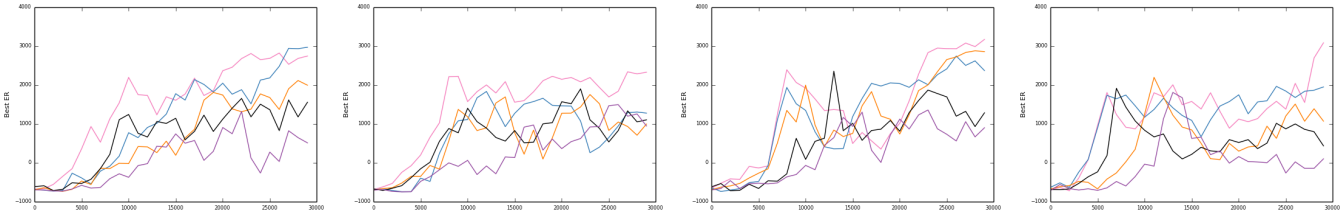
図 6: 30,000 エピソードまでの累積報酬の遷移

4.1.3 考察

実験結果から，30,000 エピソード学習することで，振子を高く上げ倒立することを学習している様子がわかる．強化学習における特定の状態下での適切な行動探索の成否により，良いモデルと悪いモデルの獲得を繰り返す局面も見られた．しかし，最終的に 30,000 エピソードまでの間に振子がかなり高い位置で倒立状態を維持できるモデルができたため，深層強化学習での三重倒立振子の倒立はうまくいったと考えられる．

4.2 実験 2

実験 2 では，三重倒立振子を Experience Replay のサイズ及びバッチ数を変更することによる動作制御の性能を調査する．



サイズが 100 の場合　サイズが 80 の場合　サイズが 60 の場合　サイズが 40 の場合

図 7: 実験結果 2:ExperienceReplay のサイズにおけるミニバッチ数を変更した際の総報酬の変化

4.2.1 実験設定

Experience Replay のサイズを 100，80，60，40 と変更して実験を行なう．また，それぞれのサイズに対するミニバッチ数を 100，80，60，40，20 と変更した．その他の設定は実験 1 と同様にした．

4.2.2 実験結果

Experience Replay のサイズ及びミニバッチ数を変更した際の総報酬の変化を図 7 に示す．また，それぞれの図については見易さのため 1000 エピソードごとの総報酬の平均をプロットした．ミニバッチ数が 100 の場合はピンク，80 の場合は青，60 の場合はオレンジ，40 の場合は黒，20 の場合は紫でプロットした．
また，各 ExperienceReplay とバッチ数ごとの三重倒立振子が初めて倒立したエピソード数を以下の表 1 にまとめた．

表 1: 倒立に必要であったエピソード数の比較

ER 数 / ミニバッチ数	100	80	60	40	20
100	1880	4150	8780	6840	8130
80	4450	5460	4020	3790	7020
60	3460	5190	5040	6200	8480
40	3430	1370	7740	3360	9320

Experience Replay がどのサイズの場合も，バッチ数が 100 の場合が一番早い段階で総報酬が高くなっていた．また、全エピソードを通しても高い総報酬を得られていた．どの Experience Replay の大きさでもバッチ数が 100，80 では他のバッチ数と比較して高い総報酬を得られている場面が多く見られた．一方，バッチ数が 60 の場合は早い段階で総報酬が高くならず，後半にバッチ数が大きいものと同様に総報酬が高くなっているものと高くならなかったもの両方が得られた．バッチ数が 40 の場合は 100，80，60 の場合と比べると総報酬は全体的に低いものが多かった．特に 30,000 エピソード付近ではかなり総報酬が低くなっていた．バッチ数が 20 の場合は全エピソードを通して他のバッチ数の場合より総報酬が低い状態がほとんどであった．

4.2.3 考察

三重倒立振子が初めて倒立したエピソード数を Experience Replay のサイズとミニバッチ数ごとに比較した結果から，Experience Replay のサイズと学習の速さには明確な相関関係が見られなかった．これは，Experience Replay にはランキング上位の各エピソードごとのデータとランダムにエピソードを取ってくるという動作が含まれていることと，ランキングをつけたエピソード数に対して，とってくるミニバッチ数が小さいため学習が十分でなかったことがその要因と考えられる．Experience Replay のサイズを小さくすると様々な状態における行動の経験が含まれなくなり，学習が難しくなることが考えられる．また，ミニバッチを作成する際にデータに偏りが生じやすくなるためサイズが小さいのはよくないと考えられるが，サイズが大きすぎると不要な経験も含まれることから，学習させるデータは，不要な経験が含まれない様なサイズのデータを使うことが望ましいと考えられる．一方で，ミニバッチ数ごとにエピソード数を比較するとミニバッチ数が多いほうが倒立するまでのエピソード数が比較的小さくなっており，学習の速度が速いと言える．これらの結果は，ミニバッチ数が多いほど一度に学習するデータの量が増えるため，学習が早く，また安定したと考えられる．ミニバッチ数が 60 程度のものは学習が安定しているものと不安定なものが得られたため今回の実験課題についてはミニバッチ数は 60 以上に設定したほうが良いことがわかる．ミニバッチ数が 20 や 40 のものの結果はミニバッチ数が小さいことから起こるデータ不足のためだと考えられる．

5. まとめ

本研究では，三重倒立振子を用いて深層強化学習の可能性を検証した．この検証結果から，深層強化学習が単純な動作を行なう物体に対する動作制御において有用であることを確認した．また，深層強化学習の性能の検証としては，学習に対する適切な Experience Replay のサイズとミニバッチ数について調査を行なった．その結果からミニバッチ数は 80 よりも大きいほうが学習が安定し，また学習速度も速いことがわかった．今後の課題として，深層強化学習を用いてより複雑な対象の動作制御を行なう．

謝辞 ：本研究において深層強化学習のプログラム作成において Qiita における ashitani 氏のプログラムを参考にさせていただきました．ここに深謝いたします．

参考文献

[1] Volodymyr Mnih,Koray Kavukcuoglu,David Silver,Alex Graves,Ioannis Antonoglou,Daan Wierstra,Martin Riedmiller,”Playing Atari with Deep Reinforcement Learning”,NIPS Deep Learning Workshop 2013

[2] Volodymyr Mnih,Koray Kavykcuoglu,David Silver, Andreu A.Rusu,“Human-level control through deep reinforcement learning ”, Nature 14326, 2015.