深層強化学習を用いた動作制御に関する一考察

橋本さゆり　(指導教員：小林　一郎)

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深層強化学習の動作制御に対する適用可能性を検証する．

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，$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\ [r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)]$ で行う．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 [1, 2] 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 target=$r_{t+1} + \gamma max Q(s, a)$ 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誤差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 target 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．

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 1 に制御対象となる三重倒立振子を示す．

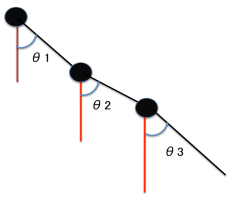


図 1: 三重振子

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を状態とする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．

$$\theta_1 \leftarrow \theta_1 + power1$$
$$\theta_2 \leftarrow \theta_2 + power2$$
$$\theta_3 \leftarrow \theta_3 + power3$$

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．

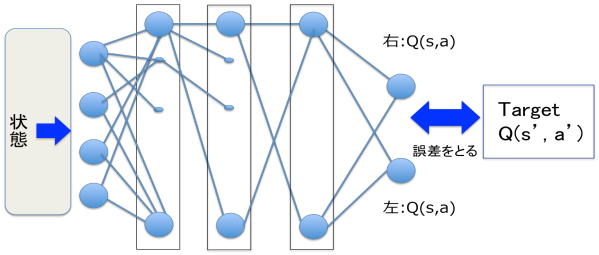


図 2: 深層ニューラルネットワーク

4　実験

4.1　実験 1

本研究では，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．

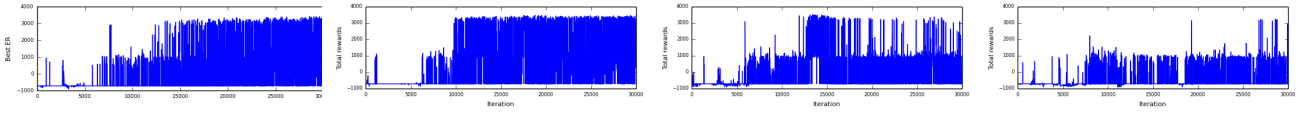
4.1.2　実験結果

- 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．

- 940 エピソード学習時（図 4 参照）

*http://chainer.org



(100,32) (50,32) (25,25) (10,10)

図 7: 実験 2 結果

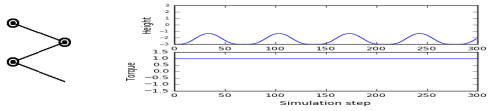


図 3: 1 エピソード学習した結果

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．

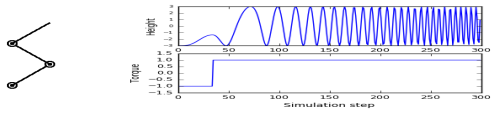


図 4: 940 エピソード学習した結果

- 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．

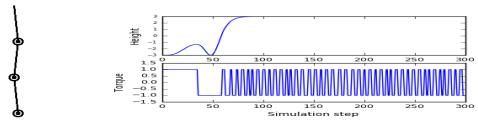


図 5: 25,590 エピソード学習した結果

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．

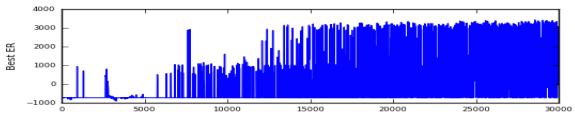


図 6: 30,000 エピソードまでの累積報酬の遷移

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 のサイズを変更することによる動作制御の性能を調査する．

4.2.1 実験設定

Experience Replay のサイズを 100，50，25，10 と変更して実験を行う．また，その際のミニバッチ数はそれぞれ 32，32，25，10 とした．その他の設定は実験 1 と同様にした．

4.2.2 実験結果

実験結果を図 7 に示す．図中の括弧の左に Experience Replay のサイズ，右にミニバッチ数を示す．サイズが 100，50 の場合は 30,000 エピソードで振子を倒立させることができていた．サイズが 25 の場合，30,000 エピソード学習するまでに良いモデルができることは多かったが 30,000 エピソード時のモデルでは三重倒立振子は倒立しなかった．サイズが 10 の場合は，三重振子が倒立した状態を維持できる良いモデルはほとんどできなかった．

4.2.3 考察

Experience Replay のサイズが 100 の場合と 50 の場合を比較すると，50 の方が良いモデルが構築できている．これは 100 の時には学習には不適なデータが混ざっていた一方で 50 の場合は 100 に比べて良質なデータのみを抽出できたためだと考えられる．また，Experience Replay のサイズを 25，10 とし，ミニバッチ数も小さくした場合にはデータ不足により多くの状態空間を網羅的に捉えられることができず，良いモデルが得られなかった．所々総報酬が良いモデルが得られているのは偶然そのエピソードで良いモデルが学習されたことによると考えられる．

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，深層強化学習の学習モデルの構築には，Experience Replay のサイズとミニバッチのサイズが関与することを確認した．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.