深層強化学習を用いた動作制御に関する考察

理学専攻・情報科学コース　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．深層強化学習とは，強化学習と深層学習を融合し，高次元データに対応することにより高精度なモデルを生成可能にするものである．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なう．具体的には振子やロボット，迷路を対象に深層強化学習の動作制御に対する適応可能性について検証を行なった．

2　振子の倒立制御

振子の倒立制御には，Deep Q-Network[1] と呼ばれる深層強化学習アルゴリズムを用いる．このアルゴリズムでは，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伝搬していくことで学習を行なう．また，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．

ネットワーク構成は図 2 に示すように入力層，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．

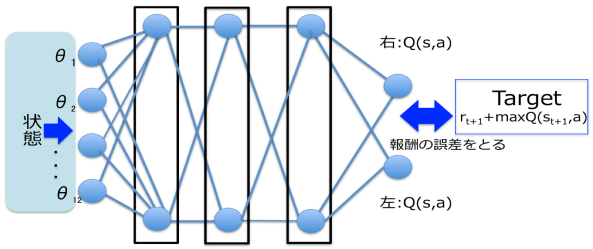

図 1: 本研究で用いるニューラルネットワーク構成

2.1　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．

2.2　実験結果

25,590 エピソード学習時（図 2 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．

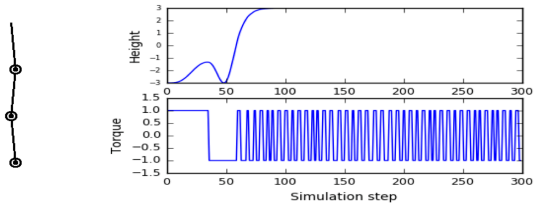

図 2: 25,590 エピソード学習した結果

3　ロボットの自然言語による制御

ロボットに「はめる」という動作を学習させる．自然言語との結びつきのために，表 1 に示すような動作の基本単位とそれを呼称する言葉の辞書を用意する．辞書にはそれぞれの関節ごとの基本動作の名称，方向, 動作の基本単位に対して後に記号列として表現可能な ID を振っている．本研究では動作の基本単位をどのように組み合わせたら「はめる」という動作ができるのかを学習し，得られた ID の系列によって「はめる」動作を表現する．

表 1: 動作の基本単位と言葉の対応関係辞書

名称	方向	ID	方向	ID	動きの単位
上腕を回転	左	A	右	B	0.05×60
上腕を上下	下	C	上	D	0.05×60
肘を回転	右	E	左	F	0.05×60
下腕を上下	下	G	上	H	0.005×60
手首を捻る	右	I	左	J	0.05×60
手首を上下	下	K	上	L	0.05×60
右指を開閉	左	M	右	N	0.05×60
左指を開閉	左	O	右	P	0.05×60

動作系列の学習には Asynchronous Advantage Actor-Critic (A3C)[2] というアルゴリズムを用いた．A3C では強化学習の手法の一つである Actor-Critic というアルゴリズムを用いている．Actor-Critic では，エージェントが行動を選択する Actor と，その選択した行動を評価する Critic という構造を持っている．A3C は Actor と Critic をそれぞれニューラルネットで表現している．また，各 CPU でそれぞれエージェントが学習を行ない，それらを非同期的に走らせ，global network でパラメータ共有するアルゴリズムになっている．このようにすることで学習を効率的にすることができる．

本研究での A3C の用い方を図 3 に示す．環境 1〜環境 4 で「はめる」というタスクをそれぞれの CPU で並列に学習させる．

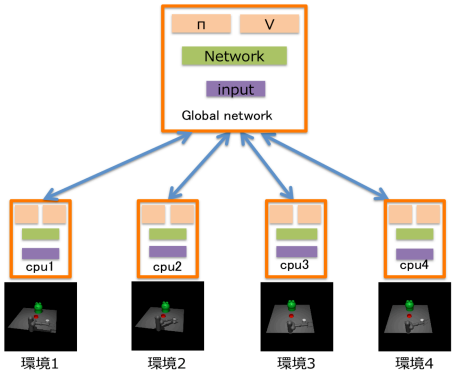

図 3: Asynchronous Advantage Actor-Critic （A3C）

3.1 実験設定

OpenAIGym[1] で提供される強化学習環境にロボットを組込み，chainerrl 上で学習を行なった．学習する環境では，8 次元の関節を持つロボットと，はめる対象となる白い円柱，及びはめる枠組みとなる赤いタブが存在する．1 エピソードを 100 回の試行とし，全 80,000,000 エピソード学習させる．報酬については，掴む際に報酬が高くなるように，ロボットの指 2 本，腕の軸と白い円柱の距離と，白い円柱と赤いタブの距離をそれぞれ計算し，係数を掛けて足し合わせたものとしている．

3.2 実験結果

図 4 に 1,100,031 学習時のモデルを示す．初期の状態からすぐに腕を平行に移動させ, 白い円柱の上まで持っていく様子が確認された. その状態から平行に下に腕を下ろし, 掴んでいる様子が確認できた. 掴んだ後に白い円柱を赤いタブに近づける動作も見られた.

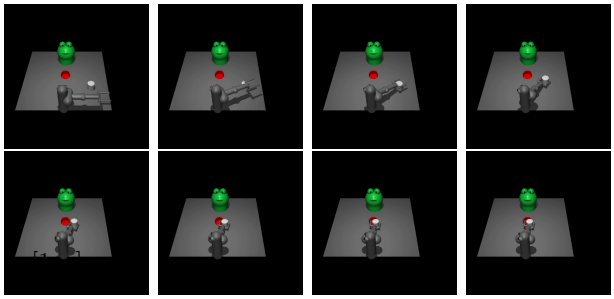

図 4: 1,100,031 エピソード学習時

この動作を行なった際に得られた動作列は以下のようになった．動作列の書き方については，ID の後に数字がくるように書いており，数字はその ID が連続で行なわれた回数を示している．

```
A8C1A2C2A1C2A1C1A1C1O1A3C1A6O1N1A2G1A2G1A1O3
・・・・・・・・・・・・・・・・・
A1G4N1G4N2G2O3N2O1G1O1G2O1G2N1G1N1O1N1G3N1G1
```

得られた記号列については他のエピソードで得られた記号列と比較することで同じ動作であっても得られる記号列は少しずつ違いがあることを確認した．

[1]https://github.com/openai/gym

4 迷路における動作の転移学習

動作の転移学習には，Distral: Robust Multitask Reinforcement Learning[3] と呼ばれる深層強化学習アルゴリズムを用いる．図 5 に概要を示すように，並列に学習を行なう環境が n 個存在し，通常それぞれに最も最適な方策 π_1〜π_n が存在するが，ここでそれぞれの方策を求めつつ，タスクに共通する部分を蒸留し，学習している全ての環境に対応できるような理想的な方策 π_0 を作り出すことを目標に学習を進めていく．この方策 π_0 を作り出すことでより効率の良い転移学習を目指す．

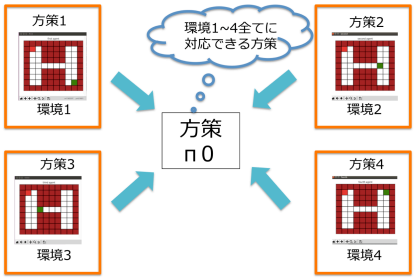

図 5: 概要図

それぞれの環境 π_i に適応する方策は π_0 を用いると式（1）のように表せる．

$$\pi_i(a_t|s_t) = \pi_0^{\alpha}(a_t|s_t)e^{\beta A_i(a_t|s_t)} \quad (1)$$

式（1）内の α,β は係数であり，A_i はそれぞれの行動価値関数と状態価値関数を用いて $A_i = Q_i(a,s) - V_i(s)$ で表される．この式（1）を用いることで，それぞれの環境に最適となる方策を学習しつつ全ての環境に適応可能な方策 π_0 を作成することを目指す．

5 おわりに

本研究では，DQN[1] を用いて振子の倒立制御に取り組んだ．DQN の適応可能性について検証し，振子以外の制御も可能であることを確認した．次にロボットに対して自然言語による動作制御を目指して A3C[2] を用いて動作制御に取り組んだ．報酬関数について動作を細分化した報酬関数を作る必要があり，複雑な動作への適応が難しいことを確認した．またマルチタスク強化学習に着目し，Distral[3] を用いての転移学習についても検証した．今後の課題として，Distral によって蒸留されたタスクに共通する方策を用いた状態価値と DQN によって得られた方策を用いた状態価値を比較したいと考えている．

参考文献

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

[2] Volodymyr Mnih, Adri Puigdom, Nech Badia, Mehdi Mirza, Alex Graves,Tim Harley, Timothy P. Lillicrap, David Silver, Koray Kavukcuoglu,“Asynchronous methods for deep reinforcement learning”

[3] Yee Whye Teh, Victor Bapst, Wojciech Marian Czarnecki, JohnQuan, JAmes Kirkpatrick, Raia Hadsell, Nicolas Heess, Razvan Pascanu, “Distral: Robust Multitask Reinforcement Learning”