7N-03

Deep Q Network による人型ロボットの把持動作選択

中村　郁仁†　　谷津　元樹‡　　原田　実‡

青山学院大学大学院理工学研究科理工学専攻知能情報コース†

青山学院大学理工学部情報テクノロジー学科‡

1. 概要

近年、Pepper など対話型ヒューマノイドロボットが広く社会に広まりつつある。そうした中で人間とロボットのより親和的な活動に関して、原田研究室では Nao や Pepper を用いたロボット対話的動作指示システム Athena[1]の開発を行っている。一方、機械学習の分野では Deep Learning の技術に対して大きな注目が集められており、特に画像処理や音声認識、ロボット制御にこの技術を応用し、従来の他の手法と比較して大幅な性能向上が見られている[2]。本研究では、汎用ロボットとして Pepper を採用し、把持動作における行動選択を Deep Q Network[3]によって行うシステムの構築を行う。

2. Deep Q Network

Deep Learning によるロボット行動決定では Deep Q Network を使用する。Deep Q Network で用いられる Q 学習とは周囲の状態sにおいて行動aをとったときの評価値を与える Q 値$Q(s,a)$をもとに次の行動a'の価値を最大化させる行動を選択する強化学習[4]の一種である。式(1)は Q 値の更新式で、αは学習率、γは割引率、rは状態sにおいて行動aを取ったときの即時報酬である。

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma max_{a'} Q(s',a')\right) \quad (1)$$

この Q 学習に対して深層学習を適応したものが Deep Q Network である。Q 学習では理論的には無限回の学習によって全ての(s,a)の組に対する真の Q 値を得ることができる。しかし、高次元な環境下では(s,a)の組が膨大となり、現実的に真の Q 値の取得は不可能となる問題がある。そのためいくつかの (s,a,r,s',a')の組を観測し、それらを学習データとし、与えられた状態sに対して各行動aに対応する Q 値を出力するニューラルネットワークの学習を行う。

A method using Deep Q Network to select gripping action of humanoid robots

Fumihito Nakamura†, Motoki Yatsu‡, Minoru Harada‡

†Graduate School of Science and Engineering, Aoyama Gakuin University

‡Faculty of Science and Engineering, Department of Integrated Information Technology, Aoyama Gakuin University

3. 提案手法

3.1. システムの概要

システム構成を図 1 に示す。図中のロボットは、把持動作のための機構を有するロボット実機を示し、本研究では Pepper を使用する。物体検出部は、三次元カメラによるキャプチャ画像から把持対象物とロボットの指先の座標値を算出すると共に開発者への GUI の提供や上記 2 つのシステムのインタフェースとして使用される。また、三次元カメラには Intel RealSenseCamera F200 を使用する。Q-net 処理部では取得したデータを基に Deep Q Network の学習を行っている。

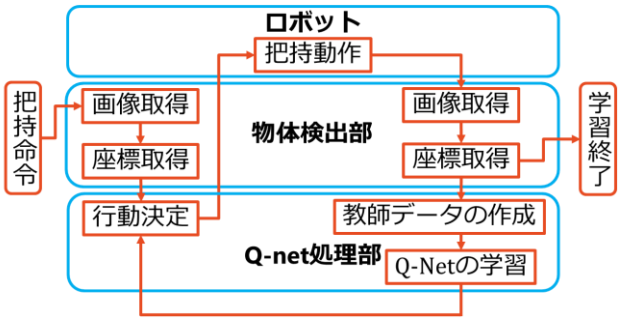

図 1. システム全体における学習フロー

Deep Q Network において、状態sは図 2 に示す通りに取得したカラー画像から把持対象物とロボットの指先を色検出し、深度画像から算出した(x,y)座標に対応するz座標の値を記した大きさ 84×84 の行列とする。行動aはロボット(Pepper)の右腕にある 6 個の関節の角度の組み合わせで 14 通り作成する。即時報酬rは状態sの行列において把持対象物とロボットの指先座標により把持動作が成功したと判定された場合は+1、動作中において把持に至っていない場合は 0、ロボットの指先が検出できない場合は-1 とした。また、+1 又は-1 の報酬を受け取った時点で一連の把持動作を終了する。

3.2. Q-net

本研究で使用したモデルQ-netを図2に示す。

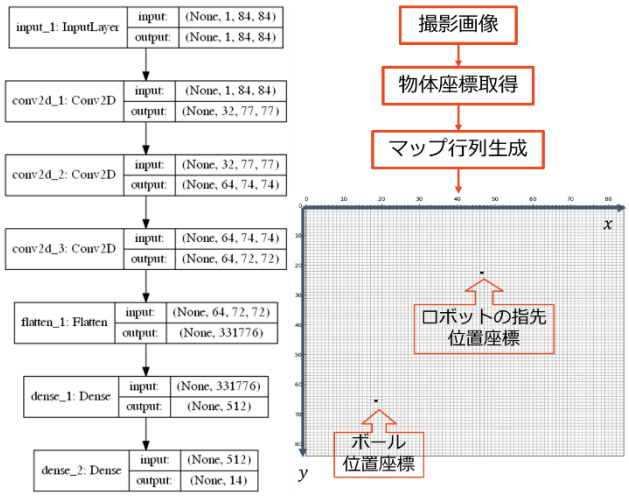

図2. Q-netと入力データ

Q-netの活性化関数には、3層のConvolution層とその後の全結合層ではReLUを、出力層につながる全結合層ではLinearを用いた。さらに、勾配降下法によるネットワークの最適化にはRMSpropを使用した。Q-netへの入力は前述した状態sの行列であり、出力は状態sに対する各行動aをとったときのQ値をもつ14個のノードである。

3.3. Q-netの学習

本研究ではこのQ-netの更新において幾つかの手法を用いた。まず、行動の選択方法においてQ-netによる出力から最大のQ値を持つ行動以外に、ある確率でランダムに行動を選択するε-greedy法を採用した。また、時系列データの相関を減らし、損失関数の値を収束しやすくするために観測した経験データからランダムにサンプリングしてQ-netの学習に用いるexperience replayを実装した。さらに行動を決定するためのモデルと学習を行うモデルを分け、一定回数学習を行った後に行動選択モデルに適時コピーし学習の安定を図っている。

また、Q-netの学習過程において、ステップ学習を採用した。これは学習初期ではランダムに設置する把持対象物に対する把持成功判定の範囲を広範囲とし、学習を続けて+1の報酬を得る確率が増えた段階で徐々に報酬付与の条件を厳しくしていくことにより、最終的に把持動作を学習することを意図した方策である。

また、本研究ではPepperの実機を用いて学習を行うなかで、稼動部の熱による故障を防ぐ安全機能などが働き処理が途中で止まってしまうことがある。そのため、これらの学習を復旧するための対策をとったシステムとした。

4. 評価実験

把持動作の学習過程を観測する。学習のための実験環境は使用するカメラからロボットの指先が検知可能な距離としてロボット(Pepper)の正面60cm程度にカメラを設置し、ロボットの右手人差し指に認識用の緑色のテープを貼り付ける。また周囲に障害物は設置しない。学習過程において+1の報酬を得る推移を以下の図3に示す。

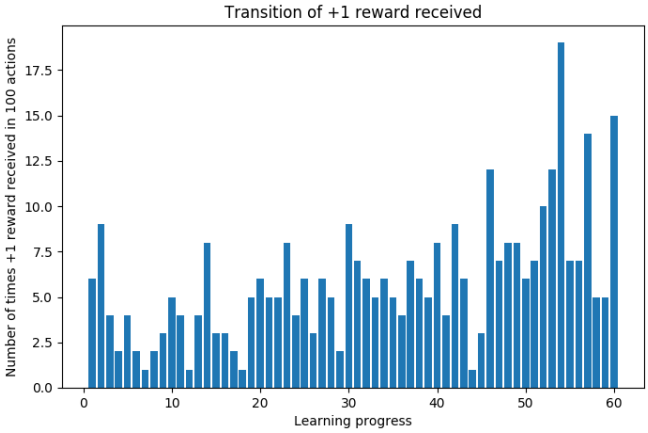

図3. 評価実験の結果

5. 結論

Deep Q Networkの適応には状態s、行動a、報酬rの適切な設定と、それぞれの適応事例に応じて適切な学習手法を定めることが重要である。また、Pepperなど汎用ロボットでは産業ロボットなどと比較して高精度な可動機構部を備えていないため動作誤差は大きい。本研究では実環境下において、そのような汎用ロボットを用いてDeep Q Networkにより把持動作をするための実装例を示し、また、評価実験結果から本研究の学習方針の有用性を示すことができた。

参考文献

[1] 田村 優樹, 長崎 達也, 中野 雅広, 原田 実:"自然言語によるロボット指示システムAthena2011",情報処理学会研究報告,2013-NL-211(9),No.5,pp.1-6,(2013.5).

[2] 分散深層強化学習でロボット制御: https://research.preferred.jp/2015/06/distributed-deep-reinforcement-learning

[3] Volodymyr Mnih,Koray Kavukcuoglu, et al. "Playing Atari with Deep Reinforcement Learning" NIPS Deep Learning Workshop 2013.

[4] Richard S.Sutton, Andrew G.Barto,"Reinforcement Learning: An Introduction" A Bradford Book, (1998.2).