

汎化ゴールにおける連続動作型ロボットアームの深層強化学習手法の開発

Generalized goal oriented deep reinforcement learning for robot arm training with continuous action space

木村 友彰¹, 渡部 雅也¹, 坂本 克好², 山口 浩一², Dinesh Bahadur Malla³, 曾我部 東馬^{*2,3,4}

¹ 電気通信大学 機械システムプログラム

² 電気通信大学 先進理工学科

³ 株式会社 グリッド

⁴ 電気通信大学 i-パワードエネルギーシステム研究センター

In multigoal reinforcement learning, Universal Value Function Approximators(UVFA) that takes not only a state but also a goal for inputs is used. We designed a task by bringing the end effector of the 7DOF robot arm to the goals using UVFA based multigoal reinforcement learning. Meanwhile, we performed the equivalent task by changing the number of goals. We confirmed a superb prediction ability by mapping the goal reachability degree using UVFA.

1. はじめに

強化学習において状態だけではなくゴールも汎化するための価値関数近似として Universal Value Function Approximators(UVFA)[1] が提案され、UVFA により目標が一つではなく複数ある場合のマルチゴール強化学習が可能になった。学習時にマルチゴールをタスクとして学習したエージェントは一つのゴールをタスクとして学習するよりも簡単に学習が進むことが示されており [2]、マルチゴール強化学習の有効性が示されている。[2]では7自由度ロボットアームを用いたマルチゴール強化学習を考えていたが、エージェントの行動はエンドエフェクタに対する操作になっていた。エージェントの行動をエンドエフェクタに対する操作ではなくアームの関節角度に対する操作とすると行動が複雑になりゴールの汎化が難しくなるということが考えられる。そこで本研究では連続動作の7自由度ロボットアームに対してエージェントの行動をアームの関節角度に対する指令値としたマルチゴール強化学習タスクを考え、この場合に対してゴールを汎化することができることを示す。

2. 強化学習

本章では本研究において用いた強化学習手法の説明を行う。

● Deep Deterministic Policy Gradients(DDPG)

Deep Deterministic Policy Gradients(DDPG)[3]は Deep Q-Networks(DQN)[4]を行動空間が連続な場合に用いることができるように拡張したアルゴリズムである。行動を決めるための Actor と行動を評価するための Critic からなる Actor-Critic 法に基づいており、Actor とそのターゲットネットワーク、Critic とそのターゲットネットワークの4つのニューラルネットワークがある。それぞれのニューラルネットワークについて簡単に説明する。Actor ネットワークは方策を表し、状態に対しての行動を出力するネットワークである。Critic ネットワークは行動価値関数を表し、

状態、行動に対しての行動価値関数の値を出力するネットワークである。ターゲットネットワークは DQN と同様に学習するネットワークと構造は同じとし学習の間は固定して学習を行った後ソフトウェアにより更新する。学習の方法に関しては Critic ネットワークは TD 誤差を最小化するように学習し、Actor ネットワークは Critic ネットワークの出力を最大化するように学習する。またエピソードを進めていくときには actor の出力に探索ノイズを付与したものを行動とする。

● Universal Value Function Approximators(UVFA)

Universal Value Function Approximators(UVFA)は通常の価値関数を目標が複数ある場合でも用いることができるように拡張したものである。まず目標が複数ある場合の強化学習つまりマルチゴールの強化学習について説明する。目標が複数あるつまりゴールが複数ありそれぞれのゴールに対して報酬関数が定義されている。エピソードの間は目指すゴールは固定されておりエピソードの最初に初期状態を決めるときに合わせて決められる。UVFA は通常の価値関数の入力に加えてゴールを入力に入れるという構造になっておりこれにより状態、行動だけではなくゴールも含めた価値を考えることが可能になる。また方策も状態だけではなくゴールも合わせて決めることになる。

3. ロボットアームへの強化学習の適用

マルチゴールの強化学習タスクとしてロボットアームのエンドエフェクタを決められた位置(ゴール)に持っていくというタスクを考えた。ここでゴールの位置は固定ではなくロボットの作業範囲の中で変えられるとする。図1がロボットアームに強化学習を適用しこのタスクを行う流れを表したものである。青い矢印がエピソードを進める流れであり、赤い矢印が学習時の流れである。強化学習エージェントとして UVFA を用いた DDPG エージェントを使い状態はアームの各関節の角度、行動は各関節への角度指令値、ゴールはゴールの座標とした。また、報酬はエンドエフェクタの位置とゴールの間の距離(cm)にマイナスをかけたものを

連絡先: 曾我部 東馬, 電気通信大学 i-パワードエネルギー・システム研究センター sogabe@uec.ac.jp

用いエピソードの終了条件はこの距離が決められた範囲内に入っているかどうかとした。

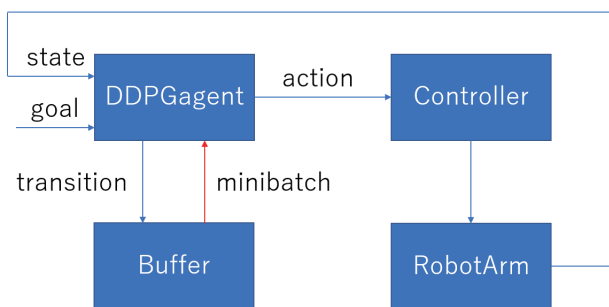


図1 ロボットアームの強化学習

エピソードの開始にエージェントはそのエピソードのゴールを受け取りエピソードが始まる。このゴールの与え方がロボットの作業範囲の中のゴールを汎化するために重要となると考えられるのでゴールの与え方を変えいくつか実験を行った。実験にあたっては Gazebo シミュレータを用いロボットのモデルには株式会社アールティの7自由度のロボットアーム CRANE-X7を用いた。

4. 実験

● ゴールを二点与える

ゴールを二つ用意し学習を行った。用意した二つのゴールの座標は(0.3, 0.2, 0.15), (0.3, -0.2, 0.15)とした。学習後の汎化性能をこれらの二つのゴールに対して x, y 方向にゴールを変更しどのくらいの距離まで近づくことができるのかによりテストした。図2より学習したゴールから離れると汎化性能は悪化していることがわかる。また、同じ距離だけ学習したゴールから離れていてもこれらの間にある方のゴールについては間もないゴールより高い汎化性能を出せていることがわかる。

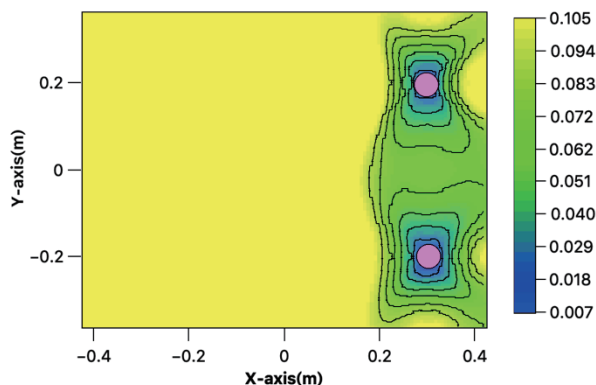


図2 ゴール二つの場合の汎化性能

● ゴールを三点与える

ゴールを三つ用意し学習を行った。用意した三つのゴールの座標は(0.3, 0.2, 0.15), (-0.2, -0.15, 0.35), (0.1, -0.4, 0.1)とした。学習後の汎化性能をこれらのゴールに対して x, y, z 方向にゴールを変更しどのくらいの距離まで近づくことができるのかによりテストした。図3において10cm近づけていることを汎化性能があるといえる限界と考えこれより離れてしまった場合汎化性能はないと考え色はグレーとした。図3よりゴール二つの場合と同じように学習したゴールから離れると汎化性能は悪化していることがわかる。また、学習したゴールの間にあるゴールに対しては離れていてもそれなりの汎化性能を出すことができています。

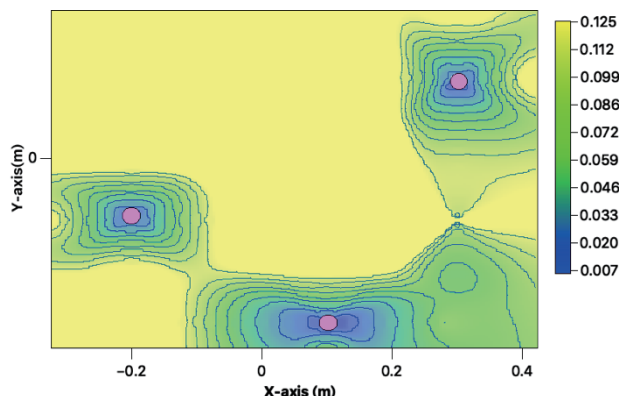


図3 ゴール三つの場合の汎化性能

5. 考察とまとめ

ゴール二つの場合、三つの場合より二つのことがわかる。一つは学習したゴールから離れると汎化性能は悪化するということがわかった。もう一つは学習したゴールの間にあるゴールにはある程度の汎化性能を持たせることができるということである。これより学習するゴールをうまく設定することができればロボットの作業範囲全てに対して汎化性能を持たせることが可能であると考えられるが、このゴールの設定方法に明確な指標はない。よって今後このゴールの設定の仕方をエージェントが学習するという手法を開発しすべてのゴールに対する汎化機能を出せるようにしていきたいと考えている。

参考文献

- [1] Schaul, Tom, Horgan, Daniel, Gregor, Karol, and Silver, David *Universal value function approximators*, In International Conference on Machine Learning, 2015.
- [2] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba *Hindsight Experience Replay*, arXiv preprint arXiv:1707.01495, 2018.
- [3] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver and Daan Wierstra *CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING*, arXiv preprint arXiv:1509.02971, 2016.
- [4] Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. *Humanlevel control through deep reinforcement learning*, Nature, 518(7540):529–533, 2015.