

特集 組立作業の自動化に向けた ロボットマニピュレーション技術の最前線

解説4

サンプル効率の良い強化学習・模倣学習と 双腕ロボットによる布操作タスクへの応用

奈良先端科学技術大学院大学 松原 崇充*

*まつばら たかみつ：研究推進機構 研究推進部門 ロボットラーニング研究室 特任准教授

緒言

少子高齢化が進むわが国では、深刻な労働力不足の到来に備えて、ロボットによる人作業代行を拡大させていくことが重要である。しかし、ロボットに作業を獲得させるために人手による動作指示や複雑なプログラミング作業が必要であれば、結局のところ労働力不足解消につながらない。そこで簡便にロボットに作業を獲得させるために、人工知能や機械学習を利用する研究が活発化している。とりわけ、ロボットが環境との相互作用を通じて、環境から与えられる報酬値を最大化する方策(行動ルール)を自律的に学習する強化学習が注目を集めている。

特に、深層学習と強化学習を組み合わせた深層強化学習は、画像などの高次元センサ入力からの特徴抽出器も合わせて学習可能であるため応用範囲が広い。深層強化学習の草分けと言える Deep Q-Network (DQN) は、さまざまな Atari ゲームタスクにおいて、未加工の画像を入力として人間と同等以上の方策を学習することに成功した。しかしながら、実ロボットを用いた作業のような収集可能なサンプル数が制限されるアプリケーションへの成功事例はあまり多くないのが実状である。

ここでは深層強化学習によりロボットにさまざまな作業を獲得させるうえで、以下の2つの課題に着目する。1つ目はサンプル効率の問題である。深層強化学習は特徴抽出器を含む方策を学習する

ために膨大なサンプルを必要とする。シミュレーションやゲームではそれほど深刻な問題とならないが、実ロボットでは実時間での動作を通してサンプルを収集するため、膨大な量のサンプル収集は現実的に不可能である。

2つ目は報酬関数設計の問題である。シミュレーションとは異なり、実環境で取得可能な情報は限られている。例えば、布操作などの複雑な作業では、衣類の状態を直接計測することが困難であることに起因して、現在の布の状態がどの程度望ましいのかを評価して正しい報酬値を与えることが難しい。これらの問題は、深層強化学習のロボットへの適用を大きく制限している。

本稿では、近年われわれが提案した、実ロボットへの応用に適した深層強化学習手法として Deep P-Network (DPN) を紹介する。この手法は、動的方策計画 (Dynamic Policy Programming (DPP)) の理論に基づいて、深層強化学習の方策を滑らかに更新することによりサンプル効率を高め、比較的少ないサンプル数でも安定した学習が期待できる。滑らかな方策更新の実現には、更新前後の方策の変化量をカルバックライブラーダイバージェンスで定量化し、それを最小化しつつ報酬和を最大化する方策を学習する方法が用いられる。

また、報酬関数設計の問題を緩和するための DPN の拡張として、近年われわれが提案した敵対的生成模倣学習手法である P-GAIL についても紹介する。P-GAIL は敵対的生成ネットワーク