論 文 内 容 の 要 旨

博士論文題目

Deep Reinforcement Learning with Smooth Policy Update for Robotic Cloth Manipulation

(方策を滑らかに更新する深層強化学習による実ロボット衣類操作の学習)

氏　　名

鶴峯義久

(論文内容の要旨)

深層強化学習は高次元情報を入力とする複雑な方策が学習可能な手法である．シミュレーションタスクへの適用では膨大なサンプルから人間と同等以上のパフォーマンスを獲得しており，実ロボットタスクへの適用により洗濯や料理，掃除等の日常的な人間の作業を代替することが期待できる．しかし，深層強化学習を実ロボットタスクに適用した研究は少ない．目標は衣類操作への DRL の適用とする．そのために必要な次の 2 つの問題について検討する．(1) ロボットは実際の行動を通してデータを取得するためシミュレーションのように膨大なサンプルを収集できない，(2) 学習には選択した行動を評価する報酬関数が必要だが，形状が柔軟に変化する衣類の報酬設計は困難である．

初めに，サンプル効率の良い深層強化学習を提案し，実ロボットによる衣類操作タスクを学習することを目指す．従来の価値関数ベース深層強化学習は，膨大なサンプルから学習することで深層学習による価値関数の近似を安定化させている．本研究では方策を滑らかに更新することで少数サンプルから安定した学習を実現する．この解決策は，方策の更新量を制約することで方策の過学習を避ける．

本研究はサンプル効率の良い深層強化学習である Deep P-Network (DPN)と Dueling Deep P-Network (Dueling DPN)を提案する．提案手法は方策を滑らかに更新する価値関数ベース深層強化学習であり，方策の更新をカルバックライブラー・ダイバージェンスで定量化することで制約する．Dueling DPN は価値関数の近似に適した深層ネットワーク構造を持ち，行動空間が大きいタスクにおいてサンプル効率を改善する．

次に，報酬関数と方策の両方を学習することで，報酬関数の設計なしに衣類

操作方策を学習することを目指す．本研究ではエキスパートデモンストレーションから報酬関数と方策を同時に学習する敵対的模倣学習フレームワークに注目する．しかし，人間がロボットの状態行動空間を用いて適切なエキスパートサンプルを収集することは困難であり，実ロボットタスクでのパフォーマンスは良くない．本研究では人間が適切に提示できる目標状態ラベルを利用し目標状態の報酬を高く見積もることで，学習方策の高いパフォーマンスは期待できない．本研究は目標状態を考慮する Double Discriminator P-Generative Adversarial Imitation Learning (DDP-GAIL)を提案する．DDP-GAIL の方策はエキスパート判別器と目標状態判別器を含む報酬関数から学習する．DDP-GAIL では二つの判別機から生成される複雑な報酬値から安定した学習を実現するために，方策を滑らかに更新する DPN を用いて方策を更新する．

本研究はシミュレーションの n DOF リーチングタスクに提案手法を適用し，学習方策のパフォーマンスやサンプル効率を従来手法と比較した．実機実験においては提案手法を双腕ロボットによるハンカチ裏返しタスクと衣類折り畳みタスクに適用した．ハンカチ裏返しタスクでは人間が設計した報酬関数で学習方策のパフォーマンスを調べた．衣類折り畳みタスクでは日常的なタスクに近い衣類操作方策を学習できるかを検証した．

氏　名	鶴峯　義久

（論文審査結果の要旨）

本論文では、洗濯や料理、掃除等の人間の日常的作業を代行するロボットの実現に向けて、実ロボットタスクへの応用に適した深層強化学習手法の提案および実ロボットタスクへの適用に関する研究を行っている。深層強化学習は、深層ニューラルネットワークと強化学習を組み合わせた機械学習手法であり、エージェントが試行錯誤によって収集する経験サンプルのみから、報酬最大化を規範として画像などの高次元情報を入力とする複雑な行動方策の学習が可能である。先行研究において、シミュレーションタスクへの適用では膨大な経験サンプルから人間と同等以上のパフォーマンスを獲得した事例も数多く報告されている。一方、実ロボットタスクへの応用事例に関する報告は少なく、また扱われるタスクも物体把持などの比較的単純なものに留まっている。

本論文では、ロボットによる衣類操作タスクに焦点を当てて、深層強化学習を実ロボットタスクに適用する際に障壁となる次の二つの問題を取り上げている。（1）手間・安全性の観点からシミュレーションのように膨大な経験サンプルの収集は困難、（2）形状が多様に変化する衣類の状態を適切に評価する報酬設計は困難、の2点である。

問題（1）の解決策として、サンプル効率を高めた深層強化学習アルゴリズムを提案している。従来の価値関数ベースの深層強化学習では、膨大な経験サンプルの利用により深層学習の学習性能を安定化させているのに対して、提案手法では方策の更新規則において、方策の急激な変化を回避するための正則化を考慮することで少数サンプルでの学習安定化を図っている。この着想に基づいて、Deep P-Network (DPN)と Dueling Deep P-Network (Dueling DPN)の2つのアルゴリズムを提案している。問題（2）の解決策として、方策と同時に報酬関数も深層ニューラルネットワークで学習することにより、事前の報酬関数設計なしに衣類操作を獲得するアプローチを提案した。熟練者によって提供される教示データと目標状態ラベルを利用し、報酬関数と方策を同時に学習する敵対的模倣学習アルゴリズムとして Double Discriminator P-Generative Adversarial Imitation Learning (DDP-GAIL)を提案している。

数値シミュレーションおよび実ロボットを用いた検証実験の結果、提案手法は先述の2つの問題に対して有効であることを確認している。また、実ロボットにより数時間程度で収集された経験サンプルのみから、ハンカチの裏返し作業やシャツ、ズボンの畳作業などの複雑な衣類操作の学習を達成している。