拡散モデルにおける模倣学習の動作保持と環境適応を両立した強化学習

ER22027 児玉将汰　　指導教授：藤吉弘亘

1. はじめに

模倣学習は人の動作を教師として模倣するように学習するため，未知の状況下では適切な動作ができない課題がある．この課題に対して強化学習による方策の更新が有効だが，強化学習することで模倣学習により獲得した動作を忘却するという問題がある．本研究では，模倣動作を保持しつつ強化学習する手法を提案する．提案手法により，動作の忘却を低減しつつ環境適応を図ることを目的とする．

2. 従来手法

模倣学習したモデルの重みを初期値として，強化学習により，方策を更新する DPPO [1] が提案されている．DPPO は，Diffusion Policy [2] の多様な動作候補を生成できる能力と強化学習を組み合わせることで，把持に失敗してもその位置を再認識し，正しい場所に戻す動作の自律的な学習が可能である．しかし，方策を逐次更新するため，方策が学習の進行で徐々に変化する．結果として，模倣学習で獲得した動作を忘却することがある．

3. 提案手法

本研究では，DPPO [1] に最適化手法 GRPO(Group Relative Policy Optimization) [3] を導入することで模倣学習で獲得した方策を参照モデルとして忘却をしないように方策の更新を行う手法を提案する．GRPO は複数の動作候補を生成し，グループ内での相対的な報酬で学習する手法であり，多様な動作候補の中から適切な動作を探索可能である．学習の際には，ResNet18 で抽出したカメラ画像の特徴量と関節の状態を入力し，関節の目標角度を出力する．GRPO による方策更新の手順を図 1 に示す．

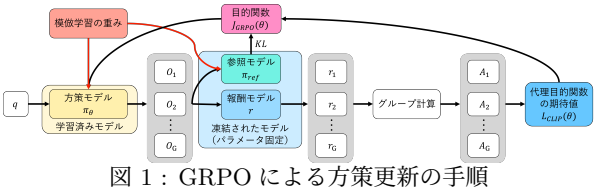

図 1：GRPO による方策更新の手順

黒色の矢印は学習中のデータの流れや演算で，赤色の矢印はパラメータの継承（初期値）を表す．まず，方策モデルが入力 q に対して複数の動作候補 O のグループを生成する．次に，報酬モデルで各応答を評価して報酬 r を算出した後，グループ内での相対的な優劣を示すアドバンテージ A を決定する．学習には，式 (1) に示す目的関数 $J_{GRPO}(\theta)$ を用いる．

$$J_{GRPO}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_{CLIP}(\theta)\right] - \beta\mathcal{D}_{KL}(\pi_\theta||\pi_{ref}) \tag{1}$$

$\mathcal{L}_{CLIP}(\theta)$ はアドバンテージに基づく代理目的関数の期待値を表し，π_{ref} は模倣学習の重みを継承しパラメータを固定した参照モデルである．学習中の方策モデル π_θ と参照モデル π_{ref} の KL ダイバージェンスを算出しており，これにより π_θ が模倣学習時の動作から過度に乖離することを抑制する．これにより，強化学習による未知環境への適応を促進しつつ，模倣学習で獲得した基本動作の維持を図る．

4. 評価実験

本実験では模倣学習と 2 種類の強化学習（PPO/GRPO）によるロボットの動作制御を行い，環境の差異が与える影響や，軌道生成の評価を行う．

4.1. 実験概要

Genesis シミュレータ内でヒューマノイドロボットの Unitree G1 を用いた実験を行う．両腕 28 関節を制御対象とし，物体把持タスクの模倣学習と強化学習を行う．使用するデータセットは実機の Unitree G1 で収集されたブロック積み上げタスクのデータであり，頭部 2 視点と両手カメラの計 4 視点の RGB 画像から構成される．強化学習の報酬は，把持対象への接近・接触および，把持の成功だけを条件にした簡易的なものとし，模倣学習による事前学習動作を活用しやすくする．

4.2. 実験条件

学習では共通して，方策モデルに Diffusion Policy，最適化手法に AdamW，学習回数を 1000 として学習する．模倣学習では，学習手法を Behavior Cloning，バッチサイズを 8，学習率を 1e-4 として学習する．強化学習では，学習手法を PPO，バッチサイズを 256，学習率を 1e-4 として学習する．提案手法は，学習手法を GRPO，バッチサイズを 32，学習率を 1e-6 として学習をする．模倣学習ではオープンソースの G1_Dex3_BlockStacking_Dataset をデータセットとして使用する．強化学習では容量削減や学習速度向上のため，カメラの画質を最低限にして実行する．

4.3. 実験結果

従来手法を青線，提案手法を赤線として学習中の報酬の推移を図 2 に示す．

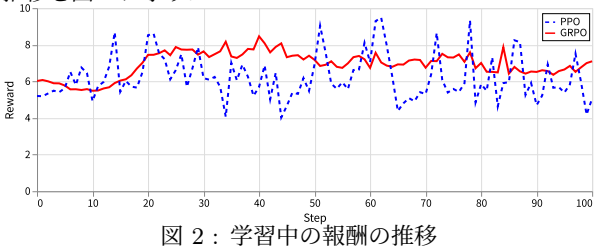

図 2：学習中の報酬の推移

従来手法は報酬の変動が激しく，提案手法は安定している．従来手法および提案手法におけるロボット頭部のカメラ視点の描画結果を図 3，図 4 に示す．模倣学習のみでは，把持対象を認識して，腕を伸ばす動作や，把持対象付近で指を曲げて，腕を上げる動作はできたが，把持には至らなかった．モデルを対象とする従来手法は，把持動作を確認（青丸）できたが，学習を進めるほど事前学習で獲得した動作から逸脱する傾向が見られた．動作が不自然で，別のブロックへの接触（黄丸）や，右腕の画面外への移動（赤丸）等の動作が見られた．一方，提案手法は，従来手法よりも事前学習に沿った動作が維持され，把持対象に充分接近してから指を曲げる動作（紫丸）が確認できた．従来手法よりも安定した動作でありながら，模倣学習よりも確実に把持対象を認識して近づく傾向（緑丸）が見られた．

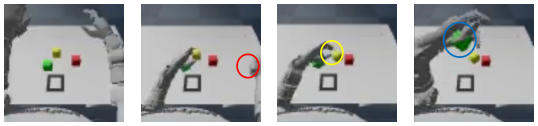

図 3：従来手法の把持動作

図 4：提案手法の把持動作

5. おわりに

本研究では，模倣学習後の強化学習によるロボットの動作精度の向上の手法の提案した．提案手法によって，模倣学習で獲得した動作を保持しつつ，強化学習が可能だと確認できた．今後は，提案手法による物体把持後の積み上げタスクを行う予定である．

参考文献

[1] AZ. Ren, *et al.*, “Diffusion Policy Policy Optimization”, ICLR, 2025.

[2] C. chi, *et al.*, “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion”, RSS, 2023.

[3] Z. Shao, *et al.*, “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models”, arXiv, 2024.