6Q-03

オフラインデータを利用した意味的探索による世界モデルのサンプル効率の改善

立松 健輔 † 綱島 秀樹 † 森島 繁生 ‡
† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

1. はじめに

深層強化学習手法はゲーム環境やロボットの制御で目覚ましい成果を示している．しかしながら，人間の学習と比べ，環境での膨大な相互作用を行う必要があるという問題を持つ．この問題に対処するため，近年では環境をモデル化する世界モデルを用いたモデルベース強化学習が研究されている．一方，既存手法では世界モデルを学習するための経験をエージェント自身が環境について無知な状態から探索を行うことで獲得しており，サンプル効率向上のボトルネックとなっている．

そこで本研究では，世界モデルを用いた強化学習において，人間が事前に用意したオフラインデータを活用し，エージェントによる環境の探索を改善する新たな学習法を提案する．具体的には図 1 のように，オフラインデータから模倣学習により基本的な行動方針を獲得し，環境の探索時に行動空間上で主成分分析を行う．主成分分析では意味のある主成分を抽出し，その主成分に沿ったノイズを加えることで長期的に意味のある探索を行う．この探索により，環境との相互作用を少なく世界モデルを学習するための経験を獲得し，サンプル効率の改善を目指す．

実験を通して，提案手法は Atari ゲーム [2] の内，模倣学習が高いスコアを獲得する環境において，環境からの限られたデータ量による学習で高いスコアを誇り，サンプル効率が改善することを確認した．

2. 提案手法

世界モデルとは環境の状態遷移を教師ありで学習するモデルである．従来手法では世界モデルを用いた強化学習を以下の 3 ステップを繰り返し行うことで進める．(1) エージェントが環境で探索を行う．(2) 獲得した経験から世界モデルを学習する．(3) 世界モデルから生成された軌跡を用いて強化学習を行う．これらの 3 ステップでは，環境のデータは世界モデルの学習のみに用いられ，エージェントは世界モデルの予測の中で学習を行う．本稿ではまず，従来手法の前にオフラインデータから模倣学習を行い，基本的な行動方針を獲得する．次にその行動方針を初期値とし，行動空間上で主成分分析によるノイズを加えることで意味のある探索を目指す．我々の研究ではベースラインとして，次章で述べる Atari 100K ベンチマーク [4] において優れた結果を示した TWM [6] を採用する．

2.1 模倣学習の導入

従来手法では，エージェントは事前学習をせずに学習を行い，環境のランダムな探索から開始する．その

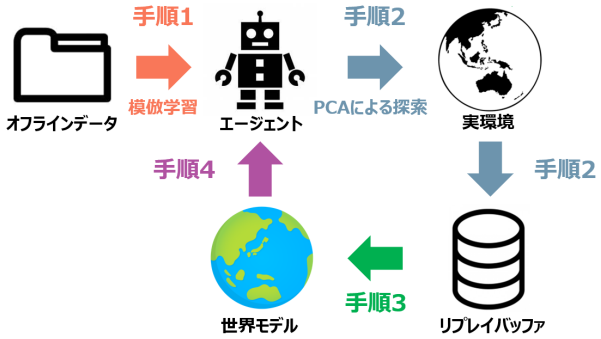

図 1: 提案手法の概要図

ため，世界モデルの学習のための十分な経験を獲得するためには多くの探索回数を必要とする．そこで，本稿では行動を模倣するように学習する Behavior Cloning (BC) を導入する．BC とは事前に人間が用意したオフラインデータを教師データとしてエージェントを学習する模倣学習の一種である．オフラインデータは観測画像とそれに対応する行動のタプルから構成される．学習では BC を入力が観測画像，出力が離散的な行動の分類問題として扱い，クロスエントロピー損失を用いる．本稿では TWM の設定に合わせ，連続した 4 フレームを MLP に入力して行動を出力するモデルを用いた．オフラインデータとして優れたプレイヤーの経験であるエキスパートデータを用いることで，エージェントは基本的な行動方針を獲得することができる．

2.2 主成分分析の導入

従来手法では確率 ε でランダムな行動をとる ε-greedy 方策と，エージェントが偏った行動をとる際に大きくなるエントロピー項によって探索を促している．今回は事前に BC により基本的な行動方針を獲得できているため，その行動空間上で意味のある潜在空間が学習されていると仮定する．その潜在空間上でデータの特性を示す主成分を抽出する主成分分析 (PCA) を行うことで，敵から逃げる・スコアを得るといった行動を分類した意味のある主成分を得られると考えられる．既存研究の GANspace [3] では生成モデルの 1 つである GAN の潜在空間上で主成分分析を行うことで，人間の性別や表情といった意味のある主成分方向を獲得し，その主成分に沿ったノイズを付加することで出力の画像の人間の性別や表情などを制御することに成功している．GANSpace のアナロジーで，環境の探索時に行動空間（エージェントの MLP）に PCA によるノイズを一定ステップ加えることによって意味のある長期的な探索が可能になると考えられる．提案手法ではオフラインデータからランダムに選択したデータをもとに PCA を行う．

Improving Sample Efficiency in World Models through Semantic Exploration Using Offline Data:
Kensuke Tatematsu†, Hideki Tsunashima†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

2.3 学習の流れ

提案手法では以下の手順に沿った学習（図 1）を行う．世界モデルの学習では過去の経験を保存し，学習に利用するリプレイバッファを用いている．

手順 1. オフラインデータを用いた模倣学習を行う．
手順 2. エージェントに PCA によるノイズを加えつつ環境の探索を行い，経験をリプレイバッファに保存する．
手順 3. リプレイバッファからデータをサンプリングし，世界モデルを学習する．
手順 4. 世界モデルから生成された軌跡からエージェントを強化学習する．
手順 5. 手順 2-手順 4 を繰り返す．

3. 実験結果

データセット 実験にはオフラインデータとして DQN Replay Dataset [1] を用いた．DQN Replay Dataset は代表的な深層強化学習手法である DQN [5] エージェントに様々な能力を評価するためのベンチマークである Atari 2600 の全 60 ゲームを学習させた際の経験を保存したものである．各ゲームで 2 億フレーム分の学習が行われ，学習中の各ステップの観測画像・行動・報酬のタプルが保存される．実験では学習終盤の優れたエージェントによる 70 万フレーム分の経験をエキスパートデータとして用いた．

ベンチマーク 実験には Atari 100K ベンチマークを用いた．Atari 100K は 26 個の Atari ゲームから構成されており，各ゲーム環境において 10 万回のアクションのみを許可している．これは人間の約 2 時間分のゲームプレイの経験に相当しており，従来の 500 分の 1 の制限となっている．今回はその中でも提案手法における性能の変化を確認するために TWM において性能が低い Alien・DemonAttack と性能が高い MsPacman の 3 つを用いて実験を行った．

評価指標 今回の実験ではサンプル効率を評価指標として用いた．サンプル効率の向上はデータ量を固定した上で性能を向上させる，もしくはより少ないデータ量で同じ性能に達することの 2 点から確認できる．Atari 100K では前者によりサンプル効率を調査する．また，Atari 100K の性能は各ゲームにおいて獲得されるスコアの累積で表される．

3.1 実験概要

提案手法の有効性を確かめるため，従来手法として TWM を用い，従来手法・模倣学習・模倣学習＋従来手法・提案手法の 4 つの比較実験を行った．模倣学習＋従来手法は図 1 の手順 2 で PCA によるノイズを加えずに探索を行う学習と対応している．学習や環境のシードの影響を考え，2 つのシードで学習を行い，各エージェントの学習終了時の 100 回分のエピソードにおいて獲得した累積スコアの平均を算出した．エピソードはゲーム開始時からゲームオーバーまでのプレイを意味する．

図 1 の手順 1 では DQN Replay Dataset から選択した 70 万フレーム分のエキスパートデータの内，50 万フレーム分を用い，手順 2 では同じ 70 万フレームの内，ランダムにサンプリングした 1 万フレーム分のデータを PCA に用いた．そして，手順 3 ではエージェントの MLP の第一層の出力に対して PCA を行い，ノイズを加える主成分は 512 成分中の第 1〜50 主成分からランダムに選択した．

表 1: 各手法における Atari100K の平均スコア

ゲーム名	従来手法	模倣学習	模倣学習＋従来手法	提案手法
Alien	576.2	794.0	911.7	1174.5
DemonAttack	381.0	107.1	299.4	307.5
MsPacman	1221.6	1078.4	1215.0	1986.3

3.2 結果

実験の結果を表 1 に示す．Alien・MsPacman では提案手法が最も高いスコアを獲得できていることが分かる．よって, オフラインデータを用いた提案手法が Atari 100K の設定でより高い性能のエージェントを学習でき，サンプル効率の改善が確認できた．一方，DemonAttck では従来手法からの性能の向上が見られなかった．また，模倣学習時点で従来手法及びランダムに行動した場合よりも平均スコアが低くなる結果になった．この原因としてはゲームの性質の違いが考えられる．Alien・MsPacman では常に同じ構造のステージ上をエージェントが探索するのに対して，DemonAttack では上下左右に敵が配置され，不規則に動く敵を正確に狙い攻撃をする必要がある．そのため，本稿で用いているような現在の状態から一意に行動を選択する模倣学習手法では学習が困難である．模倣学習の段階で基本的な行動方針を獲得できなかったため，模倣学習を初期値とした 2 種類の学習の性能が悪化したと考えられる．今後の展望として提案手法のより詳しい調査のために Atari 100K の他のゲームについての実験を考えている．

4. おわりに

本稿では世界モデルを用いた強化学習のサンプル効率を改善するために，オフラインデータを用いた模倣学習・PCA を適用する手法を提案した．実験から，Alien・MsPacman では提案手法が有効であることを確認したが，DemonAttack については模倣学習の改善が必要である．今後はオフラインデータのより良い活用のために，模倣学習に代わるオフライン強化学習や PCA を用いたノイズのかけ方について調査を行う予定である．

謝辞 本研究は，JSPS 科研費 (21H05054) の補助を受けています．

参考文献

[1] Agarwal R. *et al.* “An optimistic perspective on offline reinforcement learning”. *ICML*, 2020.

[2] Bellemare Marc G. *et al.* “The arcade learning environment:An evaluation platform for general agents”. *Journal of Artificial Intelligence Research47*, pp. 253–279, 2013.

[3] Härkönen E. *et al.* “GANSpace: Discovering Interpretable GAN Controls”. *NeurIPS*, 2020.

[4] Kaiser L. *et al.* “Model Based Reinforcement Learning for Atari”. *ICLR*, 2020.

[5] Mnih V. *et al.* “Human-level control through deep reinforcement learning”. *Nature 518*, pp. 529–533, 2015.

[6] Robine J. *et al.* “Transformer-based World Models Are Happy With 100k Interactions”. *ICLR*, 2023.