エキスパートデータを活用した世界モデルによるモデルベース強化学習

EP21100　町野駿人　　指導教授：山下 隆義

1. はじめに

モデルベース強化学習は，世界モデルで実環境のダイナミクスをモデル化し，エージェントモデルの最適な方策を効率的に学習する手法である．本手法は，世界モデルとエージェントモデルを同時に学習するため，世界モデルの性能が低いと，エージェントモデルの性能低下を誘発する．本研究では，エージェントモデルの学習促進に向け，学習過程により世界モデルで表現可能な状態遷移が異なるため，複数の学習済み世界モデルを用いたモデルベース強化学習を提案する．本手法では，人の振る舞いであるエキスパートデータをもとに模倣エージェントモデルを構築し，それらを用いて模倣エージェント特化型の学習済み世界モデルを獲得する．

2. モデルベース強化学習

モデルベース強化学習は，実環境とエージェントモデル間の相互作用による実データ収集，実データによる世界モデルの学習，世界モデルを用いた強化学習によるエージェントモデルの学習を繰り返し，世界モデルとエージェントモデルを学習する．Δ-IRIS [1] は，画像を状態とする代表的なモデルベース強化学習手法である．Δ-IRIS は，時系列モデルによる次状態の予測と，生成モデルによる確率的な次状態の生成を組み合わせた世界モデルを持つ．時刻 $t-1$ と t の状態，およびエージェントモデルの行動を生成モデルの Encoder に入力し，状態の差分 Δ を算出する．その後，時系列モデルにて時刻 t と $t+1$ の差分 Δ を予測し，生成モデルの Decoder にて 時刻 $t+1$ の状態を再構成する．この方法により，Atari2600 のビデオゲームタスクで，高精度な次状態予測と，高性能なエージェントモデルを獲得している．

3. 提案手法

モデルベース強化学習は，世界モデルとエージェントモデルを同時に学習するため，世界モデルの性能はエージェントモデルの性能に影響する．そこで，人のエキスパートデータにもとづく世界モデルを用いたモデルベース強化学習手法を提案する．本手法は，Δ -IRIS と同様に Atari2600 のビデオゲームタスクを対象とし，その概略を図 1 に示す．

Step1：学習済み世界モデルの獲得　本手法では，人のプレイデータをもとに模倣エージェントモデルを獲得し，このエージェントを用いて世界モデルを学習する．模倣エージェントモデルは，人によるプレイデータのゲームスコアにもとづき，初級者/中級者/上級者の 3 レベルに分類し，個別に模倣エージェントモデルを作成する．学習には，Behavior Cloning (BC) を用い，人のプレイデータを正解とする教師あり学習を行う．そして，模倣エージェントモデルの重みを固定したまま，模倣エージェントごとに適した世界モデルを学習する．この世界モデルは，Δ -IRIS の世界モデルと同様の構造とする．初級者，中級者，上級者の模倣エージェントによる世界モデルをそれぞれ $\mathrm{WM_l}$，$\mathrm{WM_m}$，$\mathrm{WM_h}$ とする．

Step2：エージェントモデルの強化学習　Step1 で作成した 3 つの学習済み世界モデルを学習過程に応じ 200 エポック毎に順に $\mathrm{WM_l}$，$\mathrm{WM_m}$，$\mathrm{WM_h}$ と切り替えながら用いて，モデルベース強化学習を行う．エージェントモデルの学習には，学習済み世界モデルと，実データで学習している世界モデルによって生成されたデータを交互に用いる．これにより，モデルベース強化学習時のエージェントモデルの性能向上を図る．

4. 評価実験

提案手法の有効性を検証するため，Atari2600 のビデオゲームタスクを用いて評価する．本実験では，Ms.Pac-Man に焦点を当て，モデルベース強化学習後のエージェントモデルによるゲームスコアと，世界モデルによる次状態の予測結果から評価する．模倣エージェントの学習には，人のゲームプレイデータセットである Atari Grand Challenge Dataset を用いる．このデータセットは，ゲームスコアにもとづいて，下位 10% の初級者，45 ～ 55% の中級者，上位 5% の上級者に分割する．

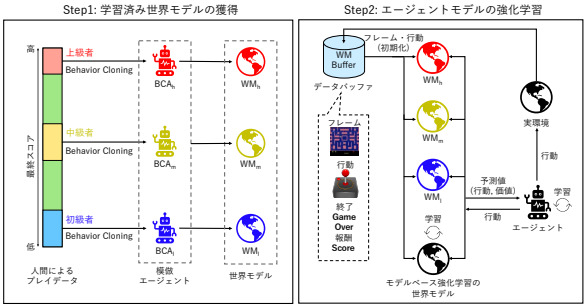

図 1：学習済み世界モデルを用いたモデルベース強化学習

4.1. 定量的評価

モデルベース強化学習後のエージェントモデルのゲームスコアから評価する．学習済み世界モデルを用いたモデルベース強化学習におけるエージェントモデルのゲームスコアを表 1 に示す．表 1 より，提案手法を用いることで単一の世界モデルを用いた場合より，高いゲームスコアを獲得可能であることからエージェントモデルの学習が促進されたと考えられる．

表 1：エージェントモデルの獲得ゲームスコア

$\mathrm{WM_h}$	$\mathrm{WM_m}$	$\mathrm{WM_l}$	ゲームスコア
			1,568.4
		✓	355.6
	✓		746.9
✓			864.3
✓	✓	✓	**1,704.1**

4.2. 定性的評価

モデルベース強化学習による世界モデルを用い，生成した次状態を可視化する．世界モデルが生成した次状態を図 2 に示す．図 2 から，エージェントが赤丸内の操作キャラクタに対して左へ曲がる Left を選択した場合，世界モデルでも操作キャラクタが左へ移動した状態を予測した．このことからエージェントの行動に適した環境の状態遷移が表現できることが確認できる．

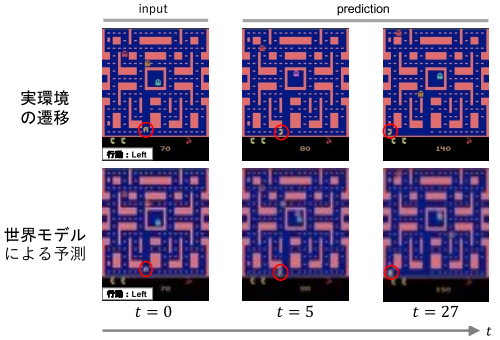

図 2：世界モデルによる次状態の予測結果

5. おわりに

本研究では，人のエキスパートデータから学習した世界モデルによるモデルベース強化学習手法を提案した．評価実験より，提案手法の有効性が確認できた．今後の展望として，Ms.Pac-Man 以外での評価実験が挙げられる．

参考文献

[1] V. Micheli *et al.*, “Efficient World Models with Context-Aware Tokenization”, ICML, 2024.