深層強化学習における状態遷移を考慮した内発的動機付けによる学習効率化に関する研究

TR22003 大鹿 海都　　指導教授：藤吉弘亘

1. はじめに

深層強化学習はエージェントと環境との相互作用により経験を収集し，報酬をもとにエージェントの行動を学習する．しかし，Montezuma's Revenge のような報酬が疎な環境では，報酬に繋がる経験を獲得することが困難となり，その結果学習が停滞する．この問題を解決する手法として，エージェントの内発的動機付けによる探索手法が提案されている [1]．本手法は本来のタスクに関する外部報酬に加え，未知の状態への訪問を誘発する内部報酬を用いる．内部報酬により，報酬が疎な環境においてもエージェントによる環境の探索が促進され，それに伴い外部報酬に繋がる経験の獲得が可能となる．内発的動機付けによる深層強化学習手法はエージェントの現状態に着目して内部報酬を生成するため，到達までの経路が異なる場合でも同じ内部報酬が生成される．そこで，本研究では新たな内部報酬として状態遷移に着目する．これにより，内発的動機付けは経験の過程も考慮するため幅広い探索表現が可能となり，エージェントによる環境探索を効率的に行えると考える．評価実験では Atari2600 を用いた評価実験により，エージェント性能を解析することで状態遷移を考慮する有効性を示す．

2. Random Network Distillation

Random Network Distillation (RND) [1] は内発的動機を用いた強化学習手法の 1 つである．RND は状態を 2 つの畳み込みニューラルネットワークに入力し，それらが出力する値の差を内部報酬とする．Random feature network は教師の役割を担っており，学習開始時に初期化したパラメータで固定する．Predictor network は学習可能なネットワークであり，Random feature network の出力を近似するように学習する．Predictor network の学習によって同じ状態を観測し続けるほど 2 つのネットワーク間の出力誤差が小さくなり，内部報酬値が低下する．一方，未知の状態を観測した場合は未学習の状態が入力であるため出力誤差が大きくなり，大きな内部報酬値が出力される．このように RND は各ネットワークの出力誤差をもとに状態の真新しさを評価する．

3. 提案手法

深層強化学習に内発的動機付けを用いた環境の探索効率化において，状態遷移を考慮した新たな内発的動機付けを提案する．

3.1 提案手法の構造

図 1 に提案手法による内部報酬生成機構を示す．本手法は RND の入力次元を任意の時刻 $n+1$ まで拡張し，現状態から n 時刻前までの状態列を入力とし，各ネットワークの出力誤差を内部報酬として出力する．状態列に対する出力誤差を状態遷移の真新しさとして評価することで，同じ現状態に対しても過去に経験した状態の遷移によって異なる評価を行う．

3.2 状態列の作成

状態列は現在を含めた過去 n 時刻前までの queue 型の配列である．状態列の作成手順を以下に説明する．

1. エージェントの行動 a により遷移後の状態を観測する
2. 状態列の最も過去の状態 s_{t-n} を忘却する
3. 状態列の s_{t-i} の状態は s_{t-i-1} の状態として格納する
4. s_t に観測した状態を格納する

このように状態を観測するたびに状態列の 1 番新しい状態として格納する．また，エピソード終了後は状態列を全てリセットし，0 埋めした状態として扱う．

3.3 エージェントの学習に用いる報酬

式 (1) に内部報酬の算出式を示す．このとき，$\boldsymbol{c}$ は状態列 $[s_t, s_{t-1}...s_{t-n}]$，$f_1(\boldsymbol{c})$ は Predictor network の出力，

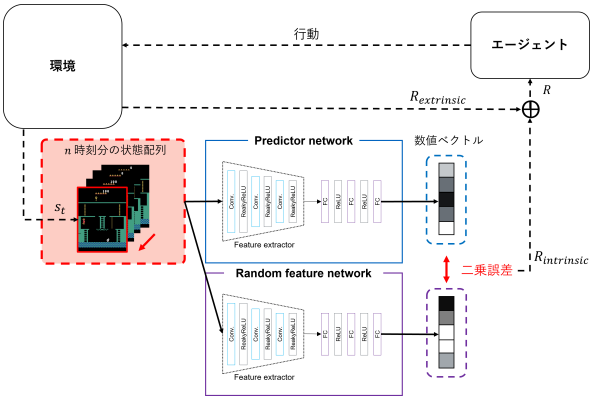


図 1: 提案手法による内部報酬生成機構

$f_2(\boldsymbol{c})$ は Random feature network の出力を示す．

$$R_{intrinsic} = ||f_1(\boldsymbol{c}) - f_2(\boldsymbol{c})||^2 \tag{1}$$

エージェントが学習に用いる報酬値は式 (2) で求める．

$$R = R_{extrinsic} + R_{intrinsic} \tag{2}$$

環境から得られた外部報酬と式 (1) によって求めた内部報酬の総和を報酬値とすることでエージェントは状態遷移を考慮した探索的な行動選択を実現する．

4. 評価実験

Atari2600 のビデオゲームタスクを用いて，状態遷移を考慮した内発的動機付けの有効性を確認する．

4.1 実験概要

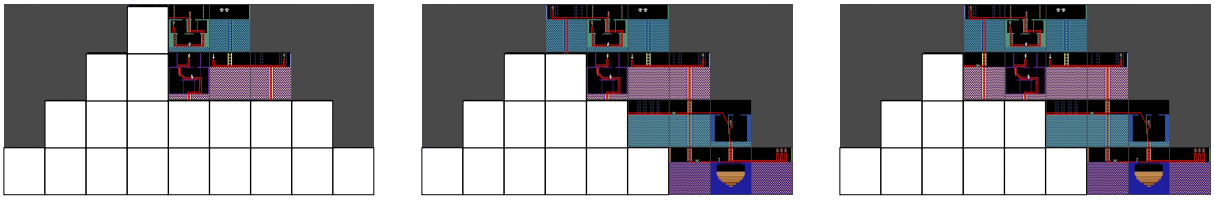
本実験ではまず，内発的動機付けなしのモデル (PPO)，RND による内発的動機付けを用いたモデル (PPO+RND)，提案手法を用いたモデル (PPO+Ours) を用いて学習済みモデルによる獲得スコア平均及び学習時の外部報酬推移を比較する．評価環境は Atari2600 の 報酬が疎な環境である Montezuma's Revenge(MR)， Venture(VT)， 報酬が密な環境である Breakout(BO) の 3 つである．比較の際は，Montezuma's Revenge は 1.5×10^3 エピソード，Venture は 1.0×10^3 エピソード，Breakout は 4.5×10^3 エピソード学習したモデルを使用する．次に Montezuma's Revenge 環境を用いて学習中に訪れた部屋の可視化を行うことで各手法に対するエージェントの探索範囲を確認する．最後に，学習した提案手法による内部報酬生成モデルを用いて観測状態に対する内部報酬を確認することで出力される内部報酬の特徴を明確に解析する．

4.2 ゲームスコアによる比較

図 2 に学習時の外部報酬推移，表 1 に各学習モデルを 100 エピソード実行したときの平均スコアを示す．このとき，表 1 の random はエージェントがランダムで行動選択を行った際の平均スコアである．Montezuma's Revenge では PPO+Ours が最も高いスコアを獲得している．さらに図 2(a) においても Ours が最も高い報酬を獲得していることが分かる．この結果から状態遷移を考慮した網羅的な探索が高いスコアの獲得に貢献できていると考えられる．Venture でも PPO+ours が最も高いスコアを獲得している．しかし，図 2(b) より報酬推移に大きな変化は見られなかった．Venture の環境は部屋数が Montezuma's Revenge と比較してかなり少なく，部屋の切り替わりによる経験の種類が限られている．提案手法では部屋の切り替わり等による大きな状態変化に着目しているため，部屋数が少ない Venture では RND と類似した推移が見られたと考えられ

(a) Montezuma's Revenge (b) Venture (c) Breakout

図 2: 学習時の報酬推移



(a) Montezuma's Revenge (b) Venture (c) PPO+Ours

図 3: 各手法毎の探索範囲

表 1: 各環境の平均スコア

env	methods			
	Random	PPO	PPO+RND	PPO+Ours
MR	0	2500	4600	**6600**
VT	0	0	1660	**1810**
BO	2	**422**	417	402

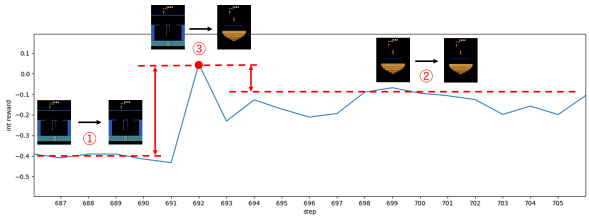


図 4: 各ステップの内部報酬の推移

る．Breakout では 3 つの手法間でスコアと図 2(c) による報酬推移ともにに大きな変化が生まれなかった．これは Breakout が報酬が密な環境であり，表 1 の Random のスコアからも，1 以上獲得できていることが分かる．そのため，外部報酬のみで十分学習可能であり，内部報酬導入による効果は低いことが考えられる．

4.3 Montezuma's Revenge の探索範囲の可視化

図 3 に各手法における学習時の探索範囲を示す．学習を通して PPO は 5 種類，PPO+RND は 12 種類，PPO+Ours は 13 種類の部屋を学習全体を通して訪れた．この結果から，提案手法である Ours が最も多くの状態を訪れていることが分かる．また，提案手法のみが訪れている部屋に着目すると，学習によって進んでいる右下の部屋ではなく左方向の部屋であることから探索的な行動の促進が行われていると考えられる．

4.4 内部報酬の可視化による比較

図 4 に Montezuma's Revenge にてエージェントが部屋を移動する瞬間における内部報酬推移を示す．図 4 の①と②に着目すると，②で高い内部報酬が出力されていることが分かる．このことから②の部屋が①の部屋より珍しいことが分かる．さらに③に着目すると，②より高い内部報酬が出力されていることが分かる．これは③は部屋が切り替わる際の内部報酬であり，部屋が切り替わる際の経験は切り替わった後の経験より少ないからだと考えられる．この結果から部屋の変化に対して変化した際とその後に対して異なる内部報酬が出力されていることが分かる．これにより長期間の探索により状態への探索促進が低下しても，状態遷移を考慮した探索促進により探索の継続が期待できる．

5. おわりに

本研究では，深層強化学習において過去の状態にも着目した幅広い探索表現を実現するため，状態遷移を考慮した内発的動機付けを提案した．提案手法では，現状態を始めとした複数時刻分の状態を RND による内部報酬生成機構に入力することで状態遷移を考慮した内部報酬の出力を実現した．評価実験では 5 時刻分の時系列情報を用いて状態遷移を表現したモデルを Atari2600 の環境を用いて評価した．その結果，状態変化に対して変化の際と変化後で異なる評価ができており，変化の際に対する探索促進の低下が抑制されていることを確認した．また，探索部屋の比較にて提案手法が一番多くの部屋の探索が出来ていることを確認するとともに，提案手法による網羅的な探索が外部報酬の獲得に繋がることを確認した．今後は，状態遷移を用いた内部報酬生成機構に Transformer encoder 構造を導入することで状態遷移の表現力向上を図る予定である．

参考文献

[1] Y. Burda, *et al.*, "Exploration by Random Network Distillation", 2019.

研究業績

[1] 大鹿海都 等, "深層強化学習における状態遷移を考慮した内発的動機付けによる探索の効率化", 電子情報通信学会技術研究報告, パターン認識・メディア理解研究会, 2024