一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

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

大鹿　海都† 　板谷　英典† 　平川　翼† 　山下　隆義† 　藤吉　弘亘†

† 中部大学　〒487–8501 愛知県春日井市松本町 1200

E-mail: {ohshika102, itaya, hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

あらまし　深層強化学習ではエージェントと環境間の相互作用により学習データを収集するため，環境の効率的な探索は網羅的な学習データの獲得に繋がる．この課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の効率化が提案されている．観測情報の新規性を評価し未知の状態空間への探索を促すことで効率的な探索を実現する．しかし，従来の内発的動機付けは現状態のみに着目しているため，環境の時系列情報を考慮していない．そこで，環境の状態遷移に着目した内発的動機付けを提案する．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．

キーワード　強化学習，内発的動機，状態遷移

1. はじめに

エージェントの最適な振る舞いを環境とのインタラクションから獲得する学習手法として，ニューラルネットワークを用いた深層強化学習がある．深層強化学習は経験からエージェントを学習するため教師データの作成が不要であり，ニューラルネットワークによる関数近似を導入することで画像のような状態空間の広大な環境においてもエージェントの高性能な振る舞いを獲得することが可能である．この性質からロボット制御やビデオゲーム攻略など多種多様な分野で応用されている [1]〜[3].

前述の通り，深層強化学習における学習データは環境とのインタラクションにより収集し，収集した経験に紐づく報酬値をもとにエージェントを学習する．しかし，報酬が疎な環境では最適な振る舞いに繋がる経験を獲得することができず，学習が停滞するという問題がある．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による探索の効率化が注目されている．この手法では，学習時に用いる報酬値として解くべきタスクに関する外部報酬と，エージェントが到達していない状態に訪れるようにエージェント内部で算出する内部報酬を用いる．報酬が疎な環境においても内部報酬により，エージェントによる環境の探索が促進され外部報酬に繋がる経験を獲得する．このように内発的動機付けの導入により，エージェントが環境を網羅的に探索することができ多種多様な学習データの獲得に繋がるため，未知環境におけるエージェントの最適な振る舞いを効率的に学習することが可能である．

内発的動機付けによる従来の深層強化学習手法は，エージェントの現状態に着目した内部報酬を用いる．環境から収集している経験とは，時系列情報であり現時刻の状態だけでなく複数時刻の状態，つまりエージェントの行動による環境の状態遷移である．そのため，状態遷移に着目した内発的動機付けは経験の過程による幅広い探索表現によりエージェントの環境探索の効率化に繋がると考えられる．そこで，我々は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を提案する．提案手法では入力次元を任意の時刻 n まで拡張し，現状態から n 時刻前までの状態配列を内部報酬生成機構の入力として利用することで状態遷移に対する新規性の表現を実現する．Atari2600 を用いた評価実験により，提案手法がエージェントの探索効率化に寄与することを確認し，従来手法と比較しエージェントが高い性能を獲得できることを示す．

2. 関連研究

本節では深層強化学習の手法についてと，深層強化学習における内発的動機付けによる探索の効率化を図った手法を包括的にまとめる．

2.1 深層強化学習

深層強化学習手法はエージェントの方策の学習方法による価値ベースと方策ベースに分類できる．価値ベースとは行動価値関数を TD 学習などにより学習し，この関数をもとに最適な行動を決定する方法である．行動価値関数は一意の値であるため，この関数に従った行動のみでは環境の探索ができない．そのため，価値ベースには ϵ-greedy 法など環境を探索するための仕組みが導入されている．行動価値を記録する Q テーブルを構築し Q テーブルを更新することで学習する Q 学習 [4] や SARSA，ニューラルネットワークを用いて行動価値関数を出力する Deep Q-Network [5] が代表的な手法として挙げられる．方策ベースとは方策勾配法などにより方策を学習し，この方策から直接的に行動を決定する方法である．この手法は価値ベースと異なり，方策は確率的に表現しているため，確率的な方策にもとづき行動選択をすることで環境の探索を実現している．方策の更新に現方策から大きく逸脱しないように制約を設けるクリッピングを導入することで安定した学習を実現した Proximal Policy

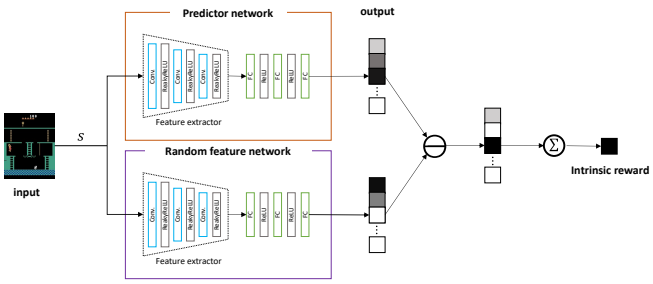


図 1: RND の内部報酬生成機構

Optimization (PPO) [6] が代表的な手法として挙げられる．

これらの手法は環境から観測した報酬値をもとにエージェントの振る舞いを最適化するため，外部報酬をもとに学習が行われる．そのため，報酬が疎な環境下では外部報酬の獲得が困難であり，エージェントの学習が困難になるケースや学習に膨大な時間を要してしまうケースがある．

2.2 内発的動機付けによる探索の効率化

報酬が疎な環境下で効率よく報酬を得るためにエージェントの内発的動機付けによる探索促進が提案されている．内発的動機付けを用いた深層強化学習は，環境から観測する外部報酬と未知領域への遷移によりエージェント内部で生成する内部報酬にもとづき学習することで，状態行動空間が広大な環境においても効率的な学習を実現している．この内部報酬は，エージェントが未知な状態に訪れた際に高い報酬値となり，既知な状態に訪れた際には低い報酬値となる．つまり，外部報酬が疎で学習が進まない環境において内部報酬は，環境の探索を促進し外部報酬の手がかりとなる経験を効率的に収集することが可能となる．Pathak らは，Forward model と Inverse model を利用し予測した次状態と実際の次状態との誤差を内部報酬として生成する Intrinsic Curiosity Module (ICM) [7] を提案している．ここで Forward model は現状態特徴と行動から次状態特徴を予測するモデル，Inverse model は現状態特徴と次状態特徴から行動を予測するモデルとして構築する．ICM はこの時の各モデルにおける次状態特徴の差を内部報酬として出力する．Burda らは，重みパラメータを固定した Random feature network と学習可能なモデルである Predictor network の出力誤差を内部報酬として生成する Random Network Distillation (RND) [8] を提案している．図 1 に RND による内部報酬生成の概略図を示す．Predictor network と Random feature network は CNN 構造となっており，各ネットワークの入力を現状態とし Predictor network の出力を Random feature network の 出力に近似するように Predictor network を学習する．これにより，未知な状態には出力誤差が大きく，つまり内部報酬が高くなり，既知な状態は出力誤差が小さく，つまり内部報酬が低くなる．Badia らは，RND による長期的な探索評価とエピソードメモリ部による局所的な探索評価によりエピソード間の探索を促進する Never Give Up (NGU) [9] を提案している．この手法は後に提案された Atari2600 の全ゲームタスクで人間のスコアを凌駕した Agent57 [10] に導入されている．また Parisi らは様々な行動空間での探索が可能なエージェントを獲得し転移学習する Change-Based Exploration Transfer (C-BET) [11] を提案している．この手法は環境の探索に特化したエージェントを内部報酬のみで獲得することで，多種多様なタスクにおいて効率的に学習可能な汎用性の高いエージェントを実現している．Ecoffet らは過去に経験した良い状態を呼び出し探索のスタート地点とする Go-Explore [12] を提案している．この手法は従来の学習方法とは異なり，ランダムな状態からのランダム探索と経験した状態に到達するような行動の模倣学習によって状態空間の効率的な探索を実現した．

これら従来の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している．しかし，エージェントの学習に用いる経験は時系列情報であり，ある状態に到達する過程が異なれば現状態が同一であっても異なる経験として考慮すべきである．そこで，我々は従来の方法と異なり内部報酬の生成にあたり状態遷移に着目することで，より効率的な環境の探索を実現する．

3. 状態遷移を考慮した内発的動機付け

深層強化学習における内発的動機付けによる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

内部報酬の生成方法は RND をベースとし，入力を状態遷移とする．図 2 に状態遷移を考慮した提案する RND の概略を示す．

本手法は RND の入力次元を任意の時刻 n まで拡張し，現状態から n 時刻前までの状態遷移を入力とし，各ネットワークの出力誤差を内部報酬としてエージェントに付与する．図 3 に本手法における状態遷移情報を示す．状態遷移情報は現在を含めた過去 n 時刻間の状態配列であり，この状態配列は queue 型の配列とする．つまり，新たに状態を観測した際には格納した状態遷移内で最も過去の状態は忘却する．また，エピソード開始時など過去の状態が存在しない場合は 0 埋めされた状態として扱う．

3.2 内部報酬生成機構

内部報酬生成機構について説明する．式 (1) に内部報酬の生成式を示す．

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

ここで，c は状態遷移配列 $[s_t, s_{t-1}...s_{t-n}]$，$f_1(c)$ は Predictor network の出力，$f_2(c)$ は Random feature network の出力を示している．内部報酬生成機構では，各時刻の情報を格納した状態遷移 c を Predictor network，Random feature network に入力する．その後，各ネットワークの出力 $f_1(s)$，$f_2(s)$ 間の二乗誤差を内部報酬値として生成する．式 (2) にエージェントが学習に用いられる報酬値の計算式を示す．

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

環境から得られた外部報酬と式 (1) によって生成された内部報酬の総和を報酬値とすることでエージェントの振る舞いを学習し，同時に内部報酬値を基づき Predictor network を学習する．

つまり，既知な状態遷移に対する出力誤差は低くなるように，

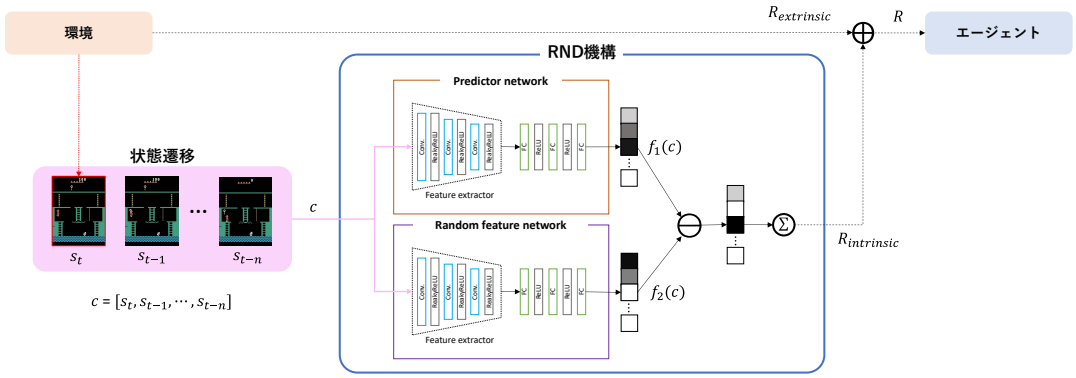


図 2: 状態遷移を考慮した RND の概略

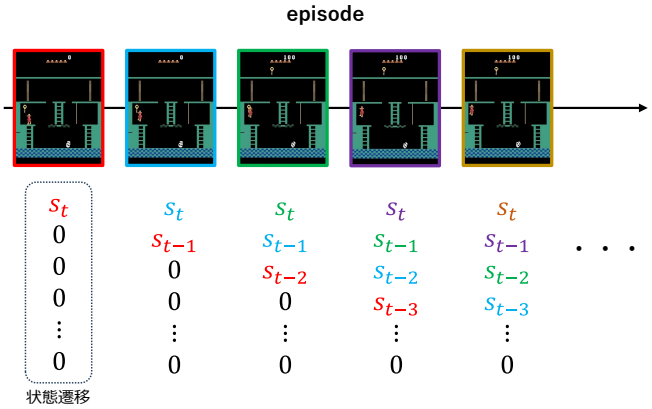


図 3: 状態遷移情報

未知な状態遷移に対する出力誤差は高くなるように学習する．したがって，状態遷移に着目し内部報酬を生成することで，現状態が既知であっても現状態までの遷移が異なる場合は未知な状態遷移として高い内部報酬値の生成を実現している．このように，本手法はエージェント内部で未知な状態遷移に着目し内部報酬を生成することで未知な経験に対する表現力を向上させる．

4. 評価実験

Atari2600 [13] 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．本実験では，深層強化学習アルゴリズムとして PPO を用い， 内発的動機付けなし (PPO), RND による内発的動機付け (PPO+RND), 状態遷移を考慮した内発的動機付け (PPO+Ours) で性能を比較する．評価環境として Atari2600 の Montezuma's Revenge， Venture, Breakout の 3 ゲームで評価する．各比較モデルの学習ステップ数は， Montezuma's Revenge では 1.5×10^7 エピソード, Venture では 1.0×10^7 エピソード, Breakout では 5.0×10^7 エピソードとした．評価指標は以下のとおりである．

- ゲームスコアによる比較.
- Montezuma's Revenge における探索範囲の可視化.
- 内部報酬の可視化による比較.

また，Montezuma's Revenge において学習の際に訪れた部屋を記録し，学習全体を通して訪れた部屋を可視化し比較する．

4.1 Atari2600 における学習環境

以下に本実験で用いる環境で 3 つのビデオゲームタスクについて述べる．

Montezuma's Revenge (MR): プレイヤーを操作して様々な部屋を訪れるゲームである．敵と接触したり，高所から落下するとエージェントの残機が 1 つ減り，残機が 0 になるとエピソード終了である．鍵やコインの取得，敵の討伐を行うとそれに応じた報酬を獲得する．本環境は外部報酬が疎な環境である．

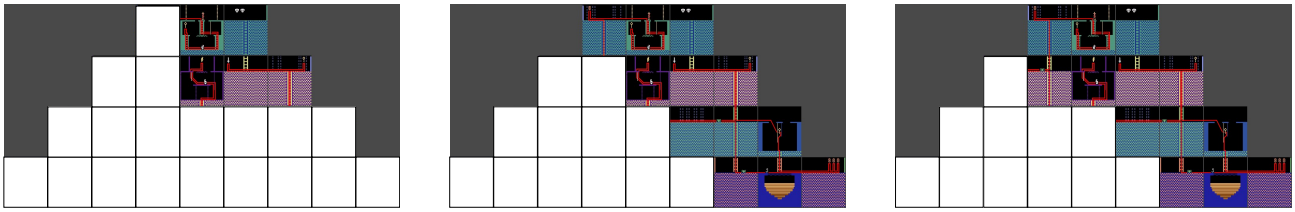
Venture (VT): プレイヤーを操作して各部屋に配置されている宝石を獲得するゲームである．部屋内には敵が配置されており接触すると残機が 1 つ減り，残機が 0 になるか全ての宝石を獲得するとエピソード終了である．宝石の獲得や敵の討伐することで報酬が与えられる．本環境は外部報酬が疎な環境である．

Breakout (BO): ポールを操作して球を打ち返すゲームである．ボールがブロックに衝突するとブロックが消えスコアを獲得する．球の打ち返しに失敗すると残機が 1 つ減り，残機が 0 になるか全てのブロックを壊すとエピソード終了である．ブロックを壊すと報酬を獲得し，最終的に壊したブロックの数がスコアとして反映される．本環境は外部報酬が密な環境である．

4.2 ゲームスコアによる比較

各ゲームタスクにおいて学習済みモデルを用いた 100 エピソード間の平均スコアにより比較する．

表 1 に各ゲームにおける平均スコアを示す．ここで Random はエージェントの行動をランダムに選択した場合のスコア，Human はプロの人間により獲得したスコア [14] を示す．MR では PPO+Ours が最もスコアが高く，PPO+RND より 2000 高いスコアを獲得している．また同様に VT においても，PPO+Ours は PPO+RND より 150 高いスコアを獲得している．一方で，BO では Random と Human を除く手法間でスコアに大きな差はないことが分かる．これらの結果から，MR や VT のような報酬が疎な環境では PPO+Ours が最もスコアが高いことから，状態遷移を考慮することで外部報酬に繋がるより良い環境の探索を実現できていると考えられる．一方で，BO のようなランダムな行動で報酬が獲得できる密な環境では，スコアに大きな差がないため外部報酬のみで十分に学習が可能であり，内部報酬導入による効果は低いと考えられる．



(a) PPO (b) PPO+RND (c) PPO+Ours

図 4: Montezuma’s Reveng におけるエージェントの探索範囲

表 1: Atari2600 における 100 エピソード間の平均スコア

env	methods				
	Random	Human	PPO	PPO+RND	PPO+Ours
MR	0	4367	2500	4600	**6600**
VT	0	1188	0	1660	**1810**
BO	2	32	422	417	**402**

4.3 Montezuma’s Revenge における探索範囲の可視化

Montezuma’s Revenge におけるエージェントの学習時に到達した範囲を可視化することで，状態遷移を考慮した内部報酬による探索の範囲向上を確認する．本実験では外部報酬のみで学習した際の探索範囲 (PPO)， 現状態に着目した内部報酬を導入した探索範囲 (PPO+RND)， 状態遷移に着目した内部報酬を導入した探索範囲 (PPO+Ours) で比較する．図 4 に各手法における学習時の探索範囲を示す．PPO と比較して PPO+RND と PPO+Ours は，多くの状態に到達できていることが分かる．また，PPO+Ours は PPO+RND と比較し到達した状態数が多い．PPO+RND と PPO+Ours はどちらも右下のエリアに進むよう学習されているが，PPO+Ours は左のエリアも探索している．この結果から，現状態のみではなく状態遷移を考慮することで，一定の探索方向だけでなくより広範囲なエリアの探索を実現している．

4.4 内部報酬の可視化による比較

生成する内部報酬を可視化することで，どのような場面で高い内部報酬を生成しているか解析し，状態遷移を考慮した内部報酬の有効性を確認する．本実験では現状態に着目した内部報酬による探索範囲 (PPO+RND)， 状態遷移に着目した内部報酬による探索範囲 (PPO+Ours) で比較する．解析に用いるエピソードは比較手法間で同一エピソードとすることで，PPO+RND と PPO+Ours 間での内部報酬値の相違を明確に解析する．

図 5 に各ゲームタスクにおける内部報酬推移を示す．図 5a から ①の場面でエージェントが落下により残機を失っている．このとき PPO+RND は落下したタイミングで報酬値は低いが，PPO+Ours では落下した瞬間は少し上昇し落下直後は低い報酬値であることが分かる．この結果から，PPO+Ours は現状態だけでなく，その後の状態に対して着目し異なる経験として捉えられていると考えられる．図 5b から 300 ステップ付近の ②では，PPO+Ours は高い報酬値であるが PPO+RND は低い報酬値である．ここで，図 5b の 30 ステップと 300 ステップ付近の ② はどちらも部屋から他の部屋への移動する状態であり，これは状態が大きく変化する経験である．しかし，PPO+RND は 30 ステップと 300 ステップ付近の状態間で，生成する内部報酬が異なり，状態を正確に表現できていない．一方で，PPO+Ours は 30 ステップと 300 ステップ付近どちらの状態も高い内部報酬を生成することができている．このことから異なる部屋間へ移動のような時系列情報が重要な状態において，状態遷移を考慮する本手法は正確に内部報酬の生成することが可能である．

また図 5c から BO における内部報酬推移を比較すると PPO+RND と PPO+Ours はどちらも同一の報酬推移である．これは BO が外部報酬が密な環境であり，内部報酬なしでエージェントの学習が十分にできるためと考えられる．

これらの結果を通して本手法は状態遷移に対して正しく報酬生成を行うことができると同時に，経験の過程による幅広い探索表現が可能となっていることが考えられる．

5. ま と め

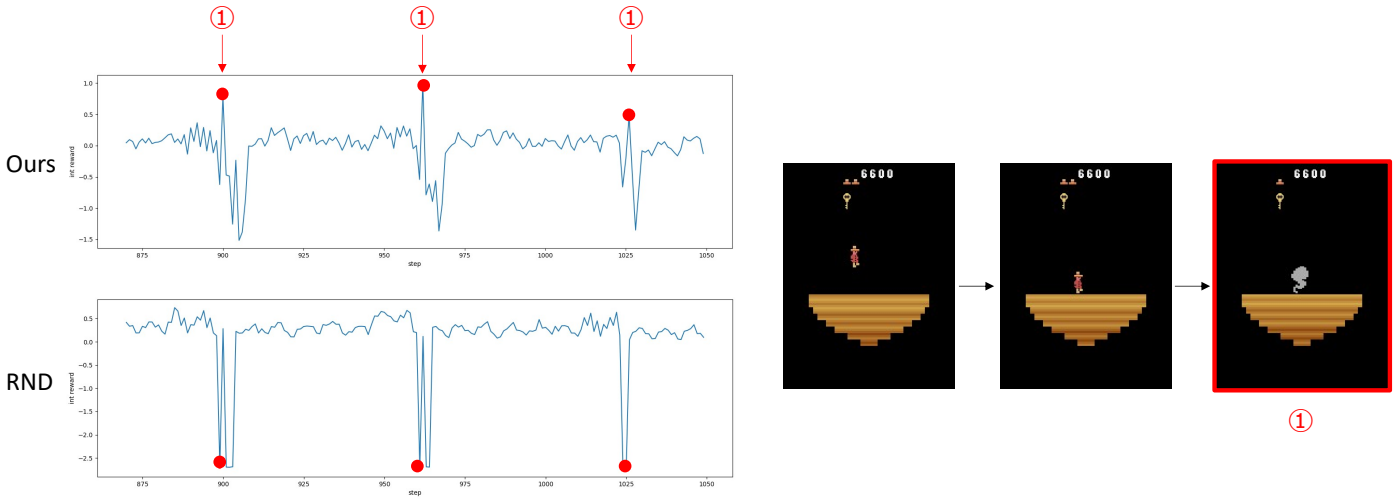
深層強化学習ではエージェントの内発的動機付けによる学習の効率化が注目されている．しかし，現状態のみ着目した発的動機付けが主流であり，エージェントが獲得した経験，つまり時系列情報には着目していない．そこで本研究では，この問題に対し深層強化学習における状態遷移を考慮した内発的動機付けを提案した．RND による内部報酬の生成する際に，現状態を始めとした複数時刻分の状態を入力とすることで状態遷移を考慮した内部報酬の生成を実現した．評価実験では 5 時刻分の時系列情報を用いて状態遷移を表現したモデルを Atari2600 の 3 つの環境で評価を行った．その結果，Montezuma’s Revenge や Venture のような外部報酬の疎な環境において RND より高いスコアを獲得した．さらに Montezuma’s Revenge では本手法がより多くの状態に到達していることを確認した．一方で，Breakout のような内発的動機を用いなくても学習可能な環境では内部報酬の導入は効果がなかったと考えられる．

今後は，状態遷移情報の特徴抽出器として Transformer 構造を導入することで内部報酬生成の表現力向上を図る予定である．

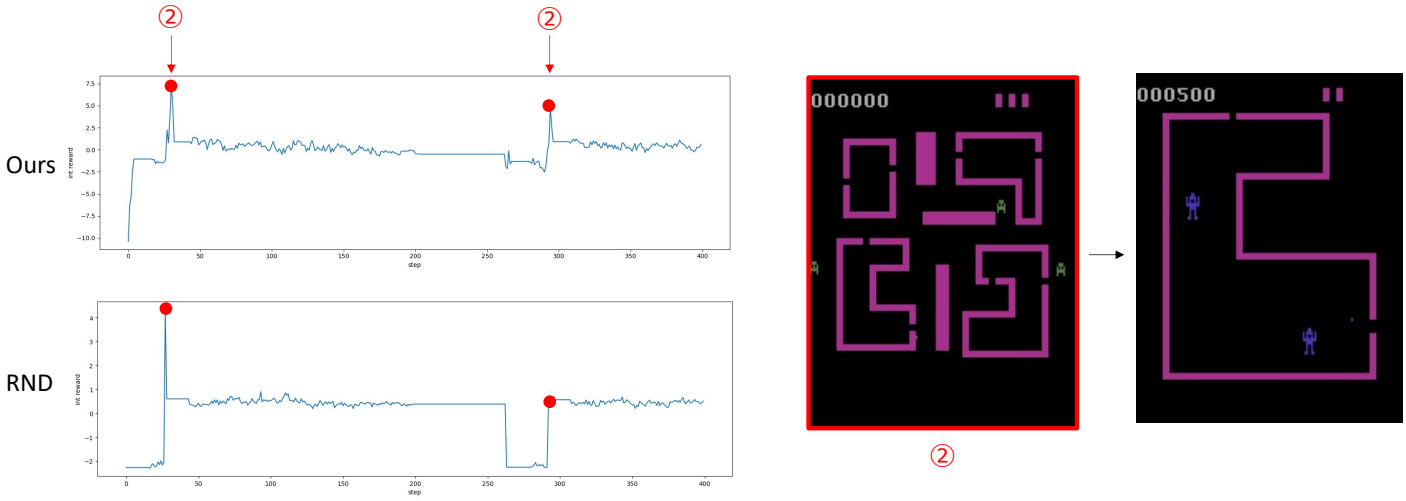
文 献

[1] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. “Mastering the game of Go with deep neural networks and tree search”, Nature, 2016.

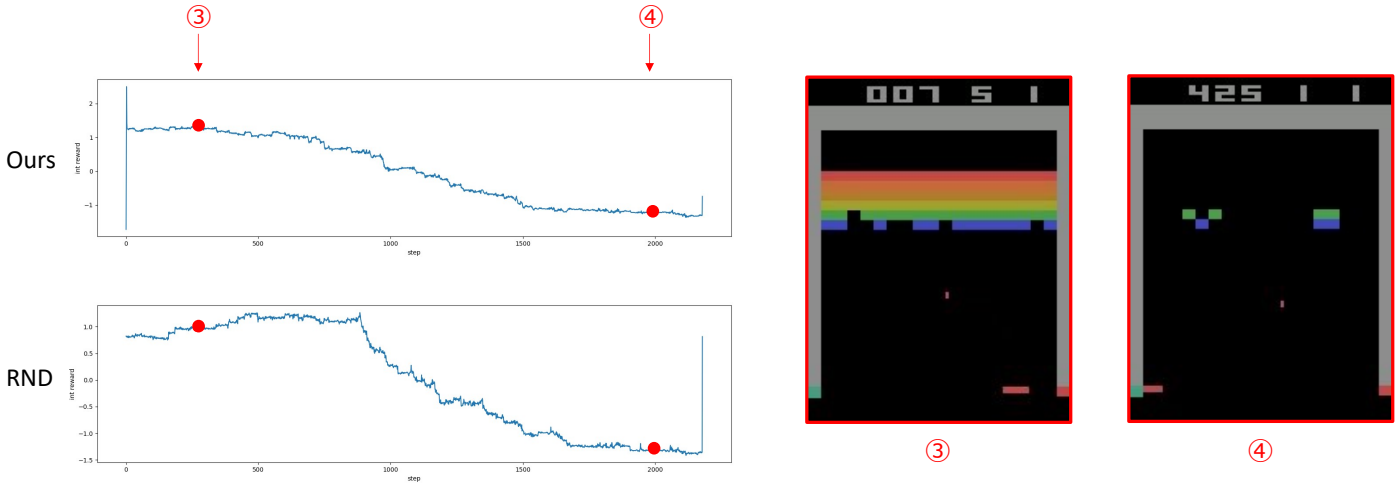
[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller, “Playing



(a) Montezuma's Revenge



(b) Venture



(c) Breakout

図 5: 各ゲームタスクにおける内部報酬推移

Atari with Deep Reinforcement Learning", NeurIPS, 2013.

[3] Shixiang Gu, Ethan Holly, Timothy Lillicrap, Sergey Levine, "Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates", *arXiv preprint arXiv:1610.00633*, 2016.

[4] Christopher Watkins, Peter Dayan, "Q-learning. Machine Learning", May 1992.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller, "Playing Atari with Deep Reinforcement Learning", NeurIPS, 2013.

[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, "Proximal Policy Optimization Algorithms", *arXiv preprint arXiv:1707.06347*, CoRR, 2017.

[7] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell, "Curiosity-driven Exploration by Self-supervised Prediction", ICML, 2017.

[8] Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov, "Exploration by Random Network Distillation", CoRR, 2012.

[9] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, and Charles Blundell. "Never Give Up: Learning Directed Exploration Strategies", ICLR, 2020.

[10] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell, "Agent57: Outperforming the Atari Human Benchmark", ICML, 2020.

[11] Simone Parisi, Victoria Dean, Deepak Pathak, Abhinav Gupta, "Interesting Object, Curious Agent: Learning Task-Agnostic Exploration", NeurIPS, 2021.

[12] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, Jeff Clune, "Go-Explore: a New Approach for Hard-Exploration Problems", *arXiv preprint arXiv:1901.10995*, 2019.

[13] Marc G. Bellemare, Yavar Naddaf, Joel Veness, Michael Bowling, "The Arcade Learning Environment: An Evaluation Platform for General Agents", *arXiv preprint arXiv:1207.4708*, 2012.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, "Human-level control through deep reinforcement learning", Nature, 2015.