Vol.2020-MPS-127 No.6
2020/3/2

深層強化学習における時系列的内部報酬生成器による探索の改善

村上 知優[1,a)]　森山 甲一[1]　松井 藤五郎[2]　武藤 敦子[1]　犬塚 信博[1]

概要：近年，高次元状態における強化学習手法として深層強化学習という手法が注目されている．しかし，深層強化学習を含む強化学習全般において，報酬が疎な環境における学習が困難であることが知られている．この問題を解決する手段として，目新しい状態の訪問に対して内的な報酬を発生させ，エージェントに多様な状態への訪問を促進させる手法が存在する．本研究ではそれを時系列的なものへ拡張し，目新しい状態遷移に対して内部報酬を生成するようにした．これにより部分観測マルコフ決定過程における探索にも対応できるようにし，実験を行った結果，その有効性を確認した．

キーワード：強化学習，深層学習，深層強化学習，探索，内部報酬

Exploration Improvement by Sequential Intrinsic Reward Generator in Deep Reinforcement Learning

1. はじめに

近年，様々な場面で物事の自動化が進められている．従来は人間が動作ロジックを全て定義することで行われてきたが，現実世界のより複雑なものを自動化するにあたって，動作ロジックを全て定義することは非常に難しく，現実的でない．こうした背景のもと，動作ロジック自体を機械が獲得する機械学習の分野が注目されており，中でも与えられた環境下で最適な行動を獲得する強化学習が大きな成果を挙げている [1].

また，近年画像認識や自然言語処理，音声認識などの分野で深層学習が大きな成果を挙げている．深層学習は入力データから特徴を抽出することに長けており，人間には認識できないような特徴を抽出することができる可能性を秘めている．実際，画像認識のコンペティションとして有名な ImageNet Large Scale Visual Recognition Challenge(ILSVRC)[2] において，ResNet[3] が人間の誤認率 5.1% を下回る誤認率 3.6% を記録し，話題となった．

さらに，深層学習と強化学習を組み合わせた深層強化学習の手法として，Deep Q Network(DQN)[4][5] が提案された．DQN は Atari2600[6] のゲーム画面を入力とし，得られたスコアを報酬とすることで，上級者を上回るスコアを出せる行動を獲得するに至っている．

一方で深層強化学習を含む強化学習全般における問題点として，報酬が得られない限り学習が進まないというものがある．強化学習では人間が設計した報酬を可能な限り多く集めるためにはどのように行動すればよいかを学習するが，どの状態で報酬が獲得できるかという知識があらかじめ存在しないため，まずは様々な状態を訪問して報酬を獲得しなければならない．しかし，あらゆる状態の中で報酬がごく一部の状態でしか得られない場合は報酬を見つけることが極めて困難である．この問題の解決策として，人間が設計した報酬を可能な限り多く集めるための行動を学習しつつも，報酬を探すための行動を学習するという手法が存在し，その中でも未知の状態に遭遇した際に報酬を発生

[1] 名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology, Gokiso, Showa, Aichi, Japan
[2] 中部大学 生命健康科学部 臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University, Mastumoto, Kasugai, Aichi, 1200, Japan
[a)] k.murakami.638@nitech.jp

させる手法を好奇心探索 [8] という．好奇心探索では人間が設計した報酬とは別に，自発的に報酬を生成する機構をエージェントに実装する．この機構はエージェントがまだ訪問したことのない状態を訪問した時に報酬を発生するものになっているため，エージェントは未知の状態を求めて行動を決定するようになる．一方で従来手法ではある状態の目新しさを評価することで報酬を生成しているため，マルコフ決定過程 (MDP) における探索にのみ対応している．また，行動の目新しさについては考慮されていない．そこで本研究ではある状態のみの目新しさを測るのではなく，状態の系列に対する目新しさも測るよう拡張し，同時に目新しい行動についても評価できるよう改良した．これにより部分観測マルコフ決定過程 (POMDP) における探索にも対応することができる．このような状態の評価のみでは十分に報酬を探すことができない場合にも対応するような新たな機構を提案することで，好奇心探索の性能の向上を図る．

2. 好奇心探索

好奇心探索はエージェントに人間でいう好奇心に当たるものを搭載することで未知の状態への訪問を促進させ，報酬を探させるというものである．未知の状態への訪問に対する報酬 (内部報酬) を与えることで，未知の状態を探すような方策を学習することになる．それと同時に人間の設計した本来の報酬 (外部報酬) を獲得することで，本来エージェントに学習してほしい方策も学習することができる．従って，探索するための方策と本来の方策の混合方策を学習することになる．しかし，最終的に獲得してほしいのは本来の方策であるため，好奇心探索では既訪問状態に対する内部報酬がしっかりと減少し，最終的には 0 に収束するような内部報酬生成器をエージェントに組み込むことが重要である．

2.1 深層強化学習における内部報酬生成器

好奇心探索を実現するためには訪問済みの状態を記録する必要がある．しかし，深層強化学習は状態の次元数が大きいことを仮定しているので，これらを全て記録することは難しい．よって一般的に次のような枠組みが用いられる．

$$\boldsymbol{y} = f(\boldsymbol{s})$$
$$i = ||\boldsymbol{t} - \boldsymbol{y}||_2^2$$

ここで，$\boldsymbol{s}$ は状態，f は Deep Neural Network(DNN)，$\boldsymbol{y}$ はその出力，$\boldsymbol{t}$ は $\boldsymbol{y}$ に対する教師データ，i は内部報酬を表す．このように DNN によって状態から何らかの推定を行い，その誤差を用いて内部報酬を計算する．この枠組みの挙動を考えてみると，頻度の大きい入力に対する出力は妥当なものとなり，誤差が小さくなって内部報酬も小さくなる．逆に頻度の小さい入力に対する出力は妥当性に欠けるため，誤差が大きくなって内部報酬も大きくなる．従って，見慣れた状態に対する内部報酬は小さくなり，見慣れない状態に対する内部報酬は大きくなる．ではこの DNN に何の推定をさせるかが問題となるが，この時に好奇心探索の項で挙げた，「既訪問状態に対する内部報酬がしっかりと減少する」という条件が重要となってくる．この条件を満たすにあたって，内部報酬がどのようなときに大きくなるのかを考えると，以下の 4 つの要因に分けられる [9]．

(1) 入力される頻度の少ない状態が入力された
(2) 入力に対する教師データが一意でない
(3) 入力された情報からでは推定ができない
(4) DNN の重みの最適化に失敗する

要因 1 は好奇心探索を実現するうえで不可欠な要素であると同時に，これ以外の要因で内部報酬が大きくなることは好ましくない．要因 2 はこれによって誤差が一向に減らなくなってしまい，結果的にどの状態を訪問しても内部報酬を獲得できてしまう．要因 3 はそもそもその推定タスクを解くことができず，誤差が一向に減らないことで，どの状態においても内部報酬が大きいまま維持されてしまう．要因 4 は DNN の重みが局所最適解に陥り，それ以上誤差が減少しなくなることで内部報酬が減少しないという状況である．要因 4 は DNN の構造や学習則に依存することであるが，要因 2 と要因 3 は何の推定を行うかに依存することであるため，これらを回避するような推定タスクにする必要がある．

2.2 Random Network Distillation

Random Network Distillation (RND)[9] は，上記要因 2,3 の回避を実現した内部報酬生成器であり，本研究における提案手法のもとになっている手法である．RND では内部報酬生成器に 2 つの DNN を使い，次のように内部報酬を計算する．

$$\boldsymbol{y} = \hat{f}(\boldsymbol{s})$$
$$\boldsymbol{t} = f(\boldsymbol{s})$$
$$i = \frac{1}{N}||\boldsymbol{t} - \boldsymbol{y}||_2^2$$

ここで，N は t 及び y の次元数である．また，$\hat{f}$ は学習を行う DNN であるが，f は学習を行わない DNN である．$\hat{f}$ を Predictor Network，f を Target Network と言う．つまり，Target Network の出力を教師データとして Predictor Network の学習を行うということである．これにより，入力に対する教師データの一貫性が保たれるため，要因 2 の回避が可能となる．また，要因 3 は Target Network と Predictor Network の入力を同じにすることで回避することができている．

3. 提案手法

本研究ではある状態のみの目新しさを測るのではなく，状態の系列に対する目新しさも測るように拡張した「Sequential Intrinsic Reward Generator (SRG)」を提案する．

3.1 深層強化学習とリカレント層

深層強化学習ではリカレント層を用いて状態や行動の履歴を内部状態という形で保持することで，POMDP に対応できることが知られている．実際に深層強化学習で Long Short Term Memory(LSTM)[10] と呼ばれるリカレント層を用いることで，POMDP における有効性を確認した例が存在する [11]．SRG ではこのリカレント層による効果を探索へ組み込む．

3.2 時系列拡張による本質的な違い

目新しさを測る対象をタプルで表現すると，RND と SRG は次のように記述できる．SRG ではこのタプル自体の目新しさを評価する．

$$RND：\{\boldsymbol{s_t}\}$$
$$SRG：\{\boldsymbol{s_{t-l}}, \boldsymbol{s_{t-l+1}}, \cdots \boldsymbol{s_t}\}$$

ここで，l は目新しさを評価する系列長である．MDP では行動を取ることで状態が遷移することが仮定されているので，SRG のタプルに含まれる状態と次の状態の間には暗に行動が含まれていることになる．よって，目新しい行動についても考慮することができる．

状態のみの評価と系列の評価の違いを図 1 を用いて具体的に考えてみる．緑色の長方形は状態を表し，黒色のエッジは両端の状態を相互に行き来することができることを意味する．はじめエージェントは s_0 におり，他の状態は未訪問であるとする．RND は $s_0 \to s_1 \to s_2 \to s_3 \to s_0$ と遷移すると，s_3 を訪問した時までは内部報酬が発生するが，最後に再び s_0 を訪問した時は内部報酬が発生しない．これは既にエージェントが s_0 を訪問済みであるからである．それに対して SRG は一番最初のタプルが $\{s_0\}$ であり，最後のタプルは $\{s_0, s_1, s_2, s_3, s_0\}$ であるため，タプルの内容が異なり，目新しい系列であると言える．よって一番最後に s_0 を訪問した際にも内部報酬が発生する．従って，RND は目新しい状態の訪問に対して内部報酬を生成するのに対し，SRG は目新しい状態遷移に対して内部報酬を生成する．また，SRG は過去の状態や行動の履歴を保持し，その目新しさを測っているため，POMDP の探索に対応することができる．

3.3 Sequential Intrinsic Reward Generator

SRG では上記タプルに対する目新しさを測るために

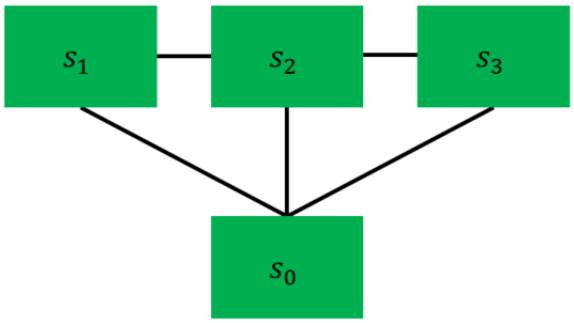

図 1 時系列評価の意義を示す環境の例．
Fig. 1 An environment showing importance of sequence based evaluation.

LSTM を用いる．RND の内部報酬生成器に新たに LSTM を含む DNN を追加し，次のように内部報酬を計算する．

$$\begin{aligned}
\boldsymbol{y_1} &= \hat{f}_1(\boldsymbol{s}) \\
\boldsymbol{t_1} &= f_1(\boldsymbol{s}) \\
\boldsymbol{y_2} &= \hat{f}_2(\phi(\boldsymbol{s})) \\
\boldsymbol{t_2} &= f_2(\phi(\boldsymbol{s})) \\
i &= \alpha \frac{1}{N_1} ||\boldsymbol{t_1} - \boldsymbol{y_1}||_2^2 + (1-\alpha)\frac{1}{N_2}||\boldsymbol{t_2} - \boldsymbol{y_2}||_2^2
\end{aligned}$$

ここで，α は状態のみの評価と時系列の評価の内分比である．N_1, N_2 はそれぞれ $\boldsymbol{t_1}, \boldsymbol{y_1}$ と $\boldsymbol{t_2}, \boldsymbol{y_2}$ の次元数を表す．また，$\hat{f}_2, f_2$ が新たに追加した DNN で，f_2 にはリカレント層が含まれていないが，$\hat{f}_2$ に LSTM が含まれている．なお，RND と同じように f_1, f_2 は学習を行わず，$\hat{f}_1, \hat{f}_2$ は学習を行う．LSTM には内部状態として過去に入力された状態が保持されているため，今獲得した状態を入力することで上記タプルに対する出力を行うことができる．しかし，SRG はタプルを時間方向へ拡張するため，DNN の学習が難しくなり，RND と比べて内部報酬が減少しにくくなる．このことは前章で説明した要因 4 による内部報酬が大きいまま維持されることを引き起こす可能性がある．よってこれを防ぐために $N_1 > N_2$ とし，状態を次元圧縮を行う関数 ϕ によって十分学習可能な次元数まで圧縮した上で $\hat{f}_2, f_2$ に入力し，学習を行う．これにより入力の多様性がある程度失われるため，$\hat{f}_2$ の学習がしやすくなる．以上をふまえて SRG のアルゴリズムは図 2 のようになる．図中の e は外部報酬，i は内部報酬，f はエピソードの終端かどうかを表すフラグである．

4. 実験

エージェントが方策を学習する環境として，OpenAI Gym[12] に存在する Atari2600[6] を用いた．その中で本実験では RND の論文で着目されていた探索の困難な環境 (Gravitar, Montezuma Revenge, Pitfall, Private Eye, Solaris, Venture) における RND と SRG の差異について検証する．方策の学習手法は RND で用いられている Proximal

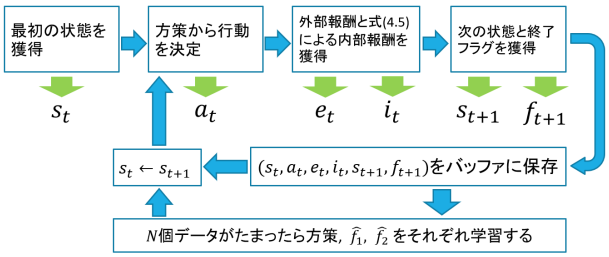

図 2　SRG のアルゴリズム.
Fig. 2　SRG algorithm.

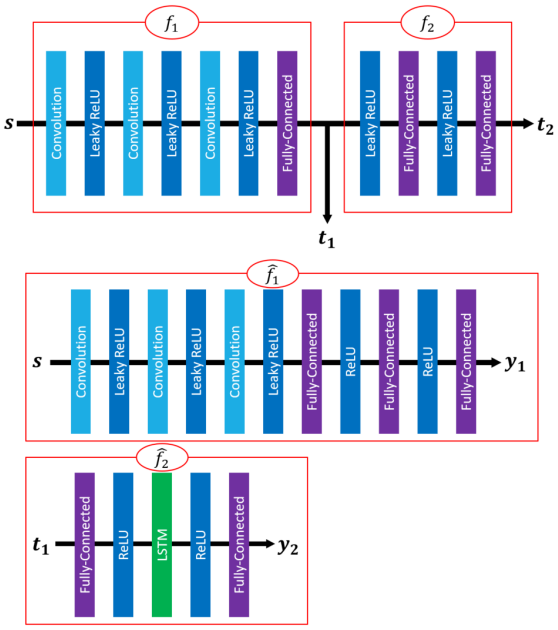

図 3　$f_1, f_2, \hat{f}_1, \hat{f}_2$ の構造.
Fig. 3　$f_1, f_2, \hat{f}_1$ and $\hat{f}_2$ architectures.

Policy Optimization (PPO)[13] を使用した．エージェントが得ることのできる状態は，ゲーム画面を画像として獲得したものである．また，状態の次元圧縮関数 ϕ として f_1 を用いた．f_1 の出力が十分次元圧縮されたものであることが前提となるが，このようにすることで ϕ の計算を省略することができ，アルゴリズム全体を高速化することができる．

本実験における DNN の構造とハイパーパラメータを図 3，表 1 に示す．$f_1, \hat{f}_1$ は RND と同じ構造とした．ハイパーパラメータは基本的に RND と同じものを用いた．

5. 結果と考察

5.1 Graviar

Gravitar は POMDP 性が低く，エージェントがコントロールする物体に強力な慣性が存在するゲームとなっている．よって，右に移動している物体を停止させたい場合，左移動を選択する必要がある．しかし，物体が高速で右に移動している場合は，早い段階で左移動を選択しないと目的の場所で停止させることができない．このように，物体

表 1　ハイパーパラメータ.
Table 1　Hyper parameters.

ハイパーパラメータ	値
並列環境数	32
α	0.5
a	0.2
外部報酬の割引率	0.999
内部報酬の割引率	0.99
Policy Network の学習率	0.0001
$\hat{f}_1$ の学習率	0.0001
$\hat{f}_2$ の学習率	0.0001
評価する系列長	エピソード長

の速度に応じた行動選択を求められる環境である．敵を攻撃することで正の外部報酬を得ることができ，負の外部報酬は存在しない．図 4 より，移動平均ではあまり差は見られなかったが，2500 あたりのスコアを大量に獲得できるのは SRG のみで，RND ではほとんど獲得できないという結果となった．SRG では行動の目新しさも測るため，様々なタイミングにおける行動を探索することができ，結果的に正しい物体のコントロールを発見することにつながったと考えられる．また，内部報酬はほとんど差異が見られない結果となった．

5.2 Montezuma Revenge

Montezuma Revenge は外部報酬が非常に疎な環境として知られている．様々な部屋と障害物が存在し，特定の部屋に存在するアイテムを獲得することで正の外部報酬を得ることができる．なお，負の外部報酬は存在しない．中には一定間隔で出現と消滅を繰り返す障害物があるような POMDP 性の高い部屋も存在する．このゲームは基本的に部屋の一番奥に 1 つの外部報酬が存在し，それを獲得することでその部屋はほぼ探索終了となる構造になっている．障害物の中には一度触れると消滅するものも存在するため，次に説明する Pitfall よりは広範な探索がしやすい環境となっている．図 5 より，Montezuma Revenge はどちらもしっかりとスコアを獲得できたことが分かる．また，先に説明した POMDP 性の高い部屋に存在する外部報酬を獲得するまでに要した部屋の訪問回数は，SRG が 126 回，RND が 362 回となり，SRG は RND の約 35% であった．このことから SRG による POMDP 性への対処が効果的であったと考えられる．一方内部報酬を見てみると，SRG は途中で急激に増加している個所が複数あることが分かる．そしてその増加はスコアの増加と同期していることが確認できる．このことから，未知の部屋の訪問に対して敏感に反応していることが分かる．

5.3 Pitfall

Pitfall は様々な部屋を訪問しながら迫り来る障害物を

回避しつつお金を集め，その際に正の外部報酬が得られる．また，障害物の回避が Montezuma Revenge と比べてとても難しいものになっており，障害物に接触すると負の外部報酬が与えられる．障害物の中には一定の時間間隔で出現する落とし穴などの POMDP 性が高いものも存在する．さらにお金が落ちている部屋が非常に限られており，Montezuma Revenge よりも正の外部報酬が疎な環境となっているため，とても高度な探索能力が要求される．図 6 を見てみると，RND，SRG ともに全く正の外部報酬を獲得できなかったことが分かる．SRG によって POMDP に対応できたとはいえ，それだけでは十分な探索が行えなかったと考えられる．一方内部報酬は 10000 から 20000 エピソードの間で RND が大きくなるのと同時に，スコアの方で RND は多くの負の報酬を獲得している．このことから探索しようと様々な部屋を訪問することができているが，その分障害物に接触したと考えられる．SRG は内部報酬が単調に減少しているため，あまり多様な部屋を訪問することができなかったがために障害物に接触する機会も少なく，あまり負の報酬を獲得しなかったと考えられる．

5.4 Private Eye

Private Eye は様々な部屋を訪問し，アイテムを回収することで正の外部報酬が得られ，障害物に接触すると負の報酬が与えられるゲームである．この環境は訪問できる部屋と回収できるアイテムに順序関係が存在するため POMDP 性が高い．順序関係が重要なため，元来た部屋に戻らなければならないという状況が多分に存在する環境となっている．図 7 より，SRG は大きな外部報酬を獲得できる頻度が RND よりも高かったことが分かる．Private Eye はアイテムを回収する順序が重要なため，元来た部屋に戻らなければならないという状況が存在するが，理論上 RND は訪問済みの部屋へ戻ると内部報酬は発生しないため，そのような動作を行う動機につながりにくい．しかし，SRG では訪問する順序が違えば内部報酬が発生するため，元の部屋へ戻るという動機につながる．このような特性が結果に強く反映されたと考えられる．一方内部報酬ではあまり差異が見られなかったが，序盤で SRG が 4000 付近のスコアを大量に獲得している段階で SRG のみ内部報酬が上昇していることが確認できる．このことから SRG による探索がスコアの獲得につながっていることが分かる．

5.5 Solaris,Venture

これらの環境では SRG と RND にほとんど差異は見られなかった．これは環境の POMDP 性が低く，順序関係も存在しないため，SRG の特性が有効でなかったためであると考えられる．

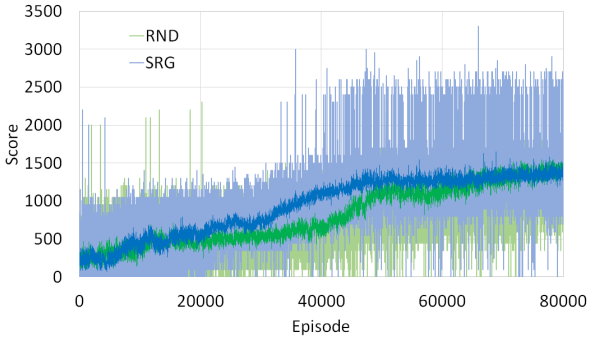

図 4 Gravitar におけるスコア (外部報酬) の推移.
Fig. 4 A learning curve of score in Gravitar.

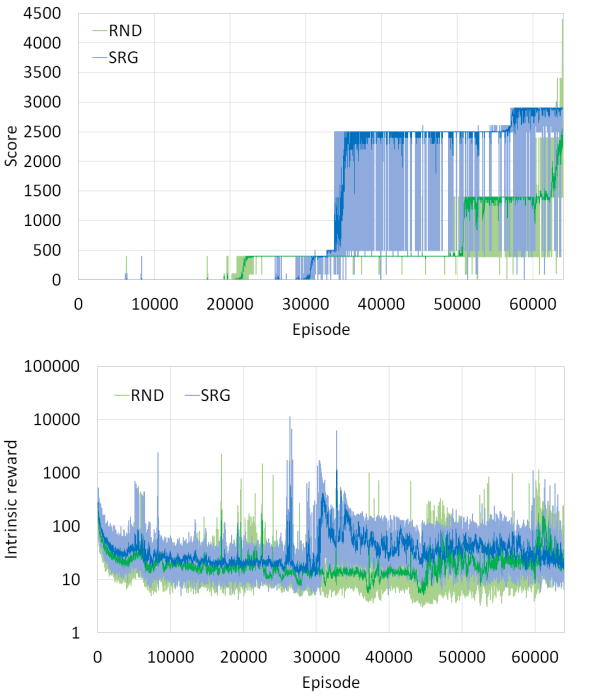

図 5 Montezuma Revenge における結果．上がスコア (外部報酬)，下が内部報酬の推移を表す．
Fig. 5 Learning curves of score and intrinsic reward in Montezuma Revenge.

6. おわりに

本研究では得られた系列の目新しさを測る内部報酬生成器を提案した．提案手法では目新しい状態のみに着目して探索を行う従来手法を，目新しい系列に着目して探索を行うものに拡張し，同時に目新しい行動についての考慮と POMDP の探索へ対応することができるようになった．Atari2600 の Pong の報酬設定を変更したものと，RND の論文で着目されていた探索の困難な環境における差異を検証した結果，提案手法で拡張された効果が得られる環境で RND との差異を確認することができた．

本研究では POMDP に対応した提案手法で Pitfall を学習させたが，結局スコアを獲得することができなかったた

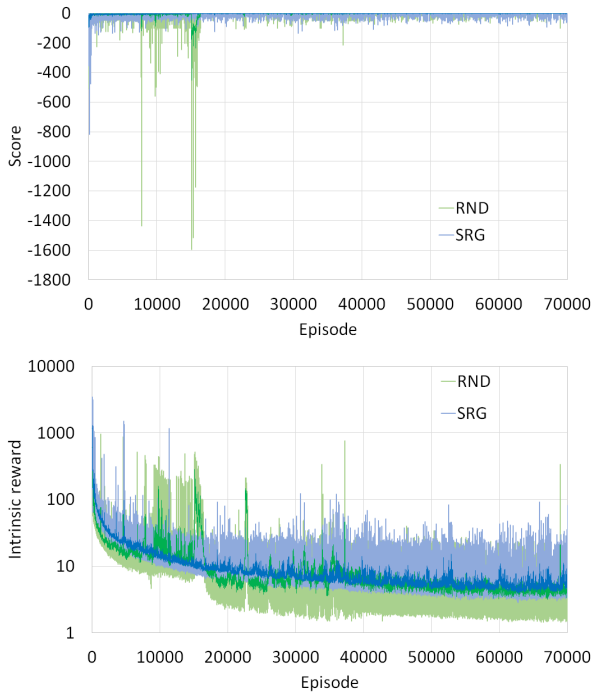

図 6 Pitfall における結果．上がスコア (外部報酬)，下が内部報酬の推移を表す．

Fig. 6 Learning curves of score and intrinsic reward in Pitfall.

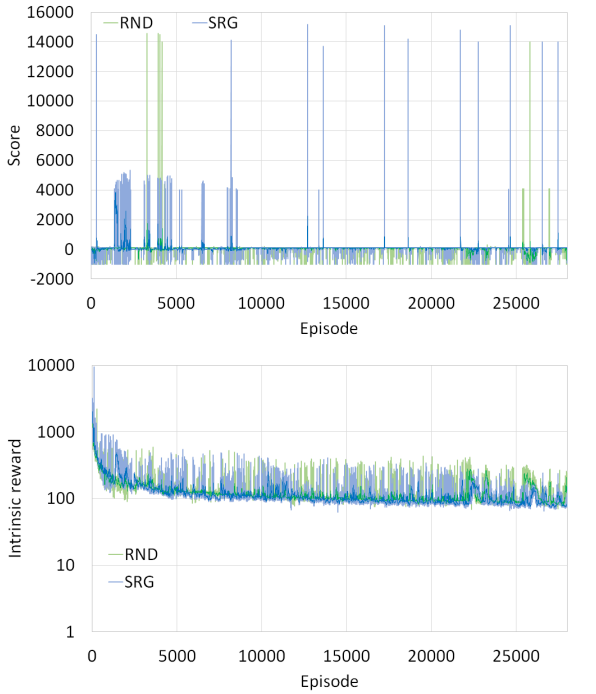

図 7 Private Eye における結果．上がスコア (外部報酬)，下が内部報酬の推移を表す．

Fig. 7 Learning curves of score and intrinsic reward in Private Eye.

め，まだ他にも探索に必要な要素が存在すると考えられる．従って今後の課題としては，Pitfall などの既存手法で解くことのできない環境におけるエージェントの挙動をより詳しく検証し，現状の探索能力に足りない要素を考察することが挙げられる．

参考文献

[1] Pieter Abbeel, Adam Coates, Andrew Y.: Autonomous Helicopter Aerobatics through Apprenticeship Learning, *International Journal of Robotics Research*, pp. 1–31 (2010).

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei.: ImageNet Large Scale Visual Recognition Challenge, arXiv:1409.0575v3 [cs.CV] (2015).

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.: Deep Residual Learning for Image Recognition, arXiv:1512.03385v1 [cs.CV] (2015).

[4] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M.: Playing Atari With Deep Reinforcement Learning, *NIPS Deep Learning Workshop* (2013).

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D.,Wierstra, D., Legg, S. and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, pp. 529–533 (2015).

[6] Brockman, Greg, Cheung, Vicki, Pettersson, Ludwig, Schneider, Jonas, Schulman, John, Tang, Jie, and Zaremba, Wojciech.: Openai gym. arXiv:1606.01540 (2016).

[7] David Silver et al.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, pp. 484–489 (2016).

[8] Andrew G. Barto.: Intrinsic Motivation and Reinforcement Learning, *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47 (2012).

[9] Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov.: Exploration by Random Network Distillation, arXiv:1810.12894 [cs.LG] (2018).

[10] Sepp Hochreiter, Jurgen Schmidhuber.: LONG SHORT-TERM MEMORY, *Neural Computation*, Vol. 9, Issue. 8, pp. 1735–1780 (1997).

[11] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu.: Asynchronous Methods for Deep Reinforcement Learning, arXiv:1602.01783 [cs.LG] (2016).

[12] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, Wojciech Zaremba.: OpenAI Gym, arXiv:1606.01540 [cs.LG] (2016).

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov.: Proximal Policy Optimization Algorithms, arXiv:1707.06347 [cs.LG] (2017).