内発的動機付けと対照学習の改良によるスパースな報酬環境における探索効率向上

Enhancing Exploration Efficiency in Sparse Reward Environments through Refined Intrinsic Motivation and Contrastive Learning

相良博喜 *1　屋藤翔麻 *2　草場壽一 *3
Hiroki Sagara　Shoma Yato　Shuichi Kusaba

*1九州大学 Kyushu University
*2東京工芸大学 Tokyo Polytechnic University
*3東京大学 University of Tokyo

本研究では、報酬が希薄な強化学習環境における探索効率を向上させるため、内発的動機付けと対照学習の改良手法を提案する。内発的動機付けは予測誤差に基づく内部報酬を提供することでエージェントを新奇な状態に導き探索を効率化させる。しかし、既存の手法（Self-supervised Network Distillation, SND）は、エージェントの不規則な動作によるノイズに脆弱であった。これを解決するため、時間的に近接したフレームを正例として扱うことで対照学習を強化するという改良を導入した。これにより、新奇性の検出を向上させ、継続的な探索を実現した。Procgen ベンチマークなどの報酬が希薄な環境での実験では、ベースラインモデルと比較して学習ステップが半分で外部報酬を獲得できるなど改善が確認された。

1. 研究背景・目的

強化学習 (Reinforcement Learning; RL) の分野は近年目覚ましい進歩を遂げている。多様なタスクへの対応、意思決定から複雑なゲームの習得に至るまで、その応用範囲を広げてきた。これらの進展は、計算能力の向上と機械学習手法、特にディープラーニングのブレークスルーによって推進されてきた。しかし、報酬信号が希薄な環境における効率的な探索は依然として大きな課題である。このような環境では、エージェントが複雑な状態空間を効率的にナビゲートし、限られたフィードバックから学習する能力が求められる。この問題に対する有望な解決策として、内発的動機 (Intrinsic Motivation; IM)[2] が挙げられる。IM は、新奇性や驚きなどの内部信号を活用して探索を導き、学習効率を高める。

この分野で注目すべき研究の一つが、Random Network Distillation(RND)[3] である。RND は新奇性検出に基づく手法であり、固定されたランダムなターゲットネットワークと予測モデルを用いる。予測モデルがターゲットネットワークの出力を模倣しようとする際に生じる誤差を新奇性の指標として用い、エージェントに探索の指標となる内発的報酬を提供する。この手法は報酬が希薄な環境でも効率よく学習することが可能であったが、固定されたターゲットネットワークに依存するため、学習が進むにつれて誤差が少なくなり内発的報酬が消失するという課題がある。

こうした課題を克服するために提案されたのが、Self-supervised Network Distillation(SND)[4] である。SND は、ターゲットネットワークを自己教師あり学習によって動的に学習させることで、ターゲットモデルの表現を適応的に進化させ、新奇性検出の精度を向上させる。

本研究では、この SND フレームワークをさらに改良するために、内発的動機の学習方法を改善した。具体的には、対照学習 (Contrastive Learning) を行う際、Noisy TV 問題が発生することを考慮した。この問題は、エージェントが一時的な新奇性に過度に引き付けられ、環境の本質的な新奇性から注意が逸れる現象を指す。既往研究においても、Procgen ベンチマークの Pitfall のようなゲームで、NPC の点滅やランダム動作にエージェントの関心が不適切に引きつけられ、報酬を得ることができないという課題があった [4]。他の環境においても、初期状態におけるキャラクターの無作為な前後移動がノイズとして作用し、エージェントの探索効率を妨げる可能性がある。この問題に対処するため、対照学習を行う際に前後フレームを考慮し、連続するフレームを同一の状態として学習させる手法を導入する。

連絡先: 相良博喜，九州大学 数理学府，福岡県福岡市西区元岡 744，sagara.hiroki.043@s.kyushu-u.ac.jp

2. 関連研究

強化学習（Reinforcement Learning; RL）は報酬が希薄な環境での探索効率を向上させるために、多くの手法が研究されている。その中でも、内発的動機付け（Intrinsic Motivation; IM）は、生物学的動機付けを模倣したアプローチとして注目されており、新奇性や予測誤差などの内部的な信号を探索の指針として用いる [1],[2]。IM は、「知識ベース」と「能力ベース」の 2 つの主要なアプローチに分類される [5],[6]。知識ベースのアプローチは、探索を通じてエージェントが世界の知識を獲得することに焦点を当て、予測誤差を利用する予測ベースの手法、新奇性を指標とするノベルティベースの手法、情報理論的指標を活用する手法に細分化される [7],[8]。能力ベースのアプローチは、エージェントがスキルを習得し環境内での目標達成能力を向上させることを目的としている [9]。

内発的動機付けに関する包括的な調査では、これらの手法が探索性能を高める上で有効である一方、それぞれの制約や実装上の課題も指摘されている [10],[17]。特に、予測ベースの手法として代表的なランダムネットワーク蒸留（Random Network Distillation; RND）は、固定されたターゲットネットワークと予測モデルの出力誤差を新奇性の指標とすることで探索を促進するが、ターゲットネットワークが静的であるため、学習が進むにつれて内発的報酬が減少し、後半の探索が停滞するという問題がある [3],[10],[17]。一方、好奇心主導型学習を大規模に検証した研究では、新奇性指標や情報理論的アプローチの有用性が示されている [11]。情報理論的アプローチとしては、環境動態モデルの情報ゲインを内発的報酬とする VIME（Variational Information Maximizing Exploration）の例が挙げられる [12]。

自己教師あり学習（Self-Supervised Learning）を活用した手法としては、BYOL-Explore や VICReg などがあり、モン

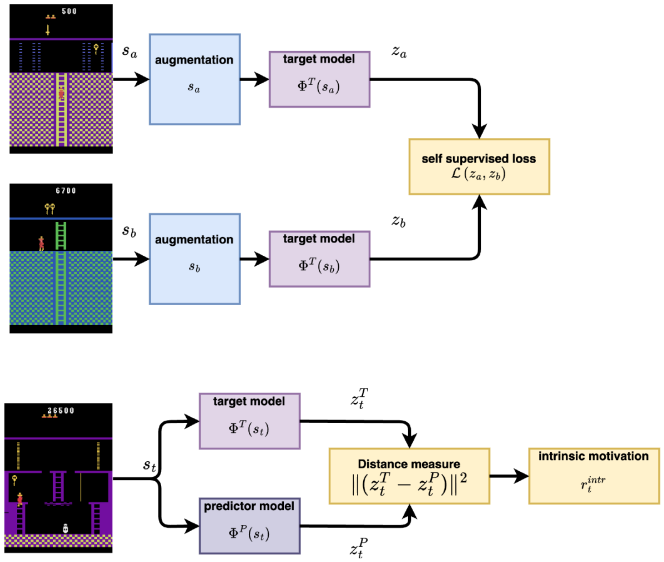

図 1: 自己教師ありネットワーク蒸留（SND）の原理。本手法は 2 つの主要な部分で構成されている。上図: ターゲットモデルを自己教師あり学習にて適切な特徴ベクトルを学習する。下図: ターゲットモデルと予測モデルの誤差から内発的報酬を得る。本図は [4] より転載。

テズマの復讐や ProcGen 環境のような高難易度タスクにおいても有効な探索を実現している [13],[14]。さらに、エージェントが環境を内部モデルとして捉える「世界モデル（World Models）」の概念が近年注目を集めている [15][16]。世界モデルでは、エージェントが環境のダイナミクスを学習し、内部的な予測や計画、さらには潜在表現を用いた目標生成を通じて探索を促進することが可能となる [10],[17],[18],[19],[20],[21],[22]。こうした世界モデルの枠組みにおいても、内発的動機付けはエージェントが未知の状態を積極的に探索するための強力な駆動力となる。一方、内発的動機付け手法においては、環境中の純粋なノイズや人間のエージェントにとって有益でない情報（いわゆる ”Noisy TV”）が過剰に興味を引いてしまう問題が指摘されている [2],[16]。この問題に対処するためには、タスク達成や学習効率の観点からノイズを制御する仕組みや報酬設計が必要であるとされる。強化学習アルゴリズムとしては、方策勾配法の一種である PPO（Proximal Policy Optimization）が、実装の容易さと安定性から広く利用されており、内発的動機付け手法との組み合わせによりスパース報酬環境でも効果的な学習が期待される [23]。

3. 提案手法

3.1 SND の基本的な手法

SND は、内発的動機付けを活用し、報酬が希薄な環境における探索効率を向上させる手法である。具体的には、強化学習における報酬に、内発的報酬 r^{intr} と外的報酬 r^{ext} の合成を用いて、新奇性の高い行動へエージェントを誘導する。

$$r_t = r_t^{\text{ext}} + \eta \cdot r_t^{\text{intr}}$$

ここで、η は内発的報酬の重み付け係数である。

内発的報酬は、予測モデル Φ_P とターゲットモデル Φ_T の間の誤差を内発的報酬として利用する [図 1]。この誤差は以下

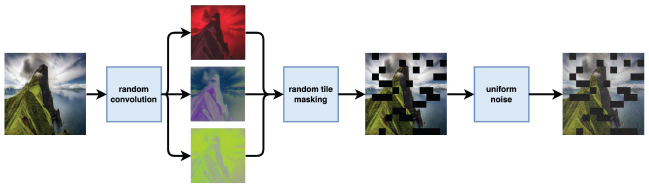

図 2: SND-V のデータオーグメンテーション方法 本図は [4] より転載。

の式で定義される：

$$r_t^{\text{intr}} = \|\Phi_T(s_t) - \Phi_P(s_t)\|_2^2$$

ここで、s_t は現在の状態を表し、$\|\cdot\|_2^2$ は L2 ノルムの 2 乗を意味する。この誤差は、新奇性を評価する指標として利用され、新しい状態ほど大きな値を取る。予測モデルは、以下の損失関数で定義される通り、ターゲットモデルの出力を模倣するよう学習が行われる。

$$L_P = \frac{1}{N}\sum_{s\in S}\|\Phi_T(s) - \Phi_P(s)\|_2^2$$

つまり、RND のようにターゲットモデルが固定されたアルゴリズムの場合、最終的には内発的報酬が収束してしまい内発的報酬が発生しなくなる。SND では、自己教師あり学習を利用してターゲットモデルを動的に更新し、状態表現の分散性を向上させる。本研究では既往論文にて提案されている SND-V というアルゴリズムを用いた。

3.2 SND-V

SND-V は対照学習をベースとした新奇性検出手法である。類似する状態の特徴を近づけ、異なる状態を分離する損失関数を用いることで、特徴空間を効果的に構築する。また、状態にノイズを付与することで、ロバストな特徴表現を学習することができる。

SND-V における損失関数の定義は以下の通りである：

$$L_T = \sum_n (\tau_n - \|Z_n - Z_n'\|_2^2)^2$$

ただし、$\tau_n = 0$ の場合、Z_n および Z_n' は同じ状態に対してノイズを付与した状態に対する特徴ベクトルを表す。$\tau_n = 1$ の場合、Z_n および Z_n' は全く異なる状態、異なるエピソード、異なるステップに対してノイズを付与した状態に対する特徴ベクトルである。

ノイズの付与方法は既往研究と同様に [図 2]、下記 3 つのフィルタを p=0.5 の確率で適用する。

・ 一様ランダムノイズ：ピクセル値に $[-1, 1]$ の範囲でノイズを追加。

・ ランダムマスキング：画像タイルをランダムにマスクする（タイルサイズは 2、4、8、12、16 ピクセル）。

・ ランダム畳み込みフィルタ

3.3 提案する改良手法

前章で述べた通り、Noisy TV 問題とはエージェントが時系列的なノイズに不適切に関心を持ち、探索が非効率化する現象を指す。

例えば、初期状態におけるキャラクターの無作為な前後移動がノイズとして作用し、エージェントの探索効率を妨げる可能

性がある。この問題に対処するため、対照学習を行う際に前後フレームを考慮し、連続するフレームを同一の状態として学習させる手法を導入する。これにより、初期状態でのランダムな動作や視覚的なノイズをエージェントの興味の対象から外し、より本質的な違いに基づく状態に関心を移すことが期待される。具体的には、SND-V に下記の改良を加えた。

1. 同一環境および STEP が $-N$ から $+N$ ステップの範囲内にある $2N+1$ 個の状態からランダムに選択する。
2. 選択した状態にノイズを加え、類似状態ペアを構成する。
3. 非類似状態の選択は既往研究と同様にランダムに行う。

このアプローチでは、既往研究の SND-V が採用していた完全に同一の状態ペアからノイズを加える方式よりも多様な状態表現を学習できる。特に周期的に変化する状態を同一状態と捉え、無駄な内発的報酬の発生を抑えることが期待される。これにより、状態空間全体でより精度の高い新奇性検出が可能となり、エージェントが短期的なノイズではなく本質的な環境変化に基づいて行動できるようになる。

4. 実験・考察

本研究では、Self-supervised Network Distillation(SND)[4] の条件設定を引き継ぎ、提案手法の有効性を検証した。具体的には、ProcGen ベンチマークの Climber 環境において、提案手法を適用し、探索効率および報酬取得の改善を評価した。

本実験では、ペアの選択範囲は N=1、つまり前後ステップの状態を選択する方法を採用した。 [図 3] が本手法の外部報酬の推移である。既往研究では、2.6×10^7 ステップで初めて外部報酬を得たが、提案手法では 1.3×10^7 ステップ、つまり半分以下のステップ数で外部報酬を獲得していることがわかる。また、内発的報酬 [図 4] に関しても、違いがある。0.1 から 0.8×10^7 ステップの序盤において、既往研究では内発的動機のピークは現れないが、提案手法では内発的動機のピークが現れていることがわかる。

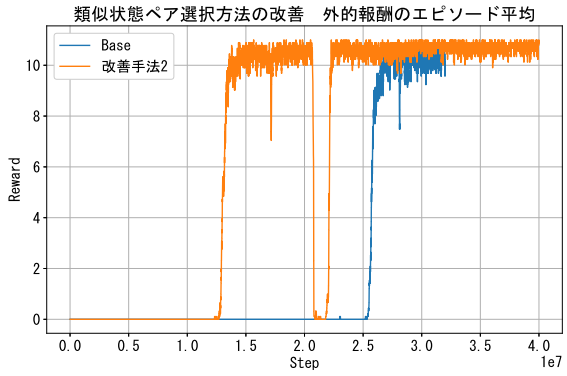

図 3: 外的報酬の推移

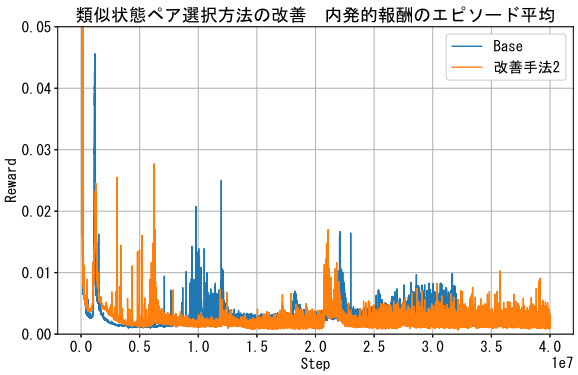

図 4: 内発的報酬の推移

学習序盤で高い内発的動機を得ることができた要因として、エージェントが適切な行動を選択することができず摂動状態に陥ることがあるが、提案手法では大きな状態変化を得ることができ、内発的動機が発生しやすくなっていることが考えられる。これにより学習序盤の探索効率が向上し、外部報酬を得るまでの時間が短くなっていると考えられる。

5. まとめ

本研究では、内発的報酬調整と対照学習の改良により、報酬が希薄な環境での探索効率を向上させる手法を提案し、Climber での実験においてその効果を実証した。今後は、より難易度の高いゲーム環境 (Montezuma's Revenge, Pitfall など) での検証を進め、提案手法の汎用性を評価する必要がある。また、将来的な研究の方向性として、外発的報酬のスケーリングパラメータの最適化など、より効果的な内発的動機付けの実現に向けた検討が必要である。

参考文献

[1] Sutton, R. S., Barto, A. G. Reinforcement Learning: An Introduction. MIT Press (1998).

[2] P. -Y. Oudeyer, F. Kaplan and V. V. Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. IEEE Transactions on Evolutionary Computation, vol. 11, no. 2, pp. 265-286 (2007).

[3] Yuri Burda and Harrison Edwards and Amos Storkey and Oleg Klimov. Exploration by random network distillation. International Conference on Learning Representations (2019).

[4] Matej Pecháč, Michal Chovanec, Igor Farkaš. Self-supervised network distillation: An effective approach to exploration in sparse reward environments. Neurocomputing. Volume 599 (2024).

[5] Ryan, R. M., Deci, E. L. Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology, 25(1), 54-67 (2000).

[6] Deci, E. L., Ryan, R. M. Intrinsic Motivation and Self-Determination in Human Behavior. Springer Science and Business Media (1985).

[7] Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., Munos, R. Unifying Count-Based Exploration and Intrinsic Motivation. Advances in Neural Information Processing Systems, vol. 29 (2016).

[8] Aubret, A., Matignon, L., Hassas, S. A Survey on Intrinsic Motivation in Reinforcement Learning. arXiv:1908.06976 (2019).

[9] Pathak, D., Agrawal, P., Efros, A. A., Darrell, T. Curiosity-Driven Exploration by Self-Supervised Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).

[10] Yuan, M. Intrinsically-Motivated Reinforcement Learning: A Brief Introduction. arXiv:2203.02298 (2022).

[11] Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., Efros, A. A. Large-Scale Study of Curiosity-Driven Learning. arXiv:1808.04355 (2018).

[12] Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F., Abbeel, P. VIME: Variational Information Maximizing Exploration. Advances in Neural Information Processing Systems (NeurIPS) (2016).

[13] Grill, J. B., Strub, F., Altché, F., et al. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems (NeurIPS) (2020).

[14] Seo, Y., Shin, J., Lee, K. State Entropy Maximization with Random Encoders for Efficient Exploration. Advances in Neural Information Processing Systems (NeurIPS) (2021).

[15] Ha, D., Schmidhuber, J. World Models. arXiv:1803.10122 (2018).

[16] Schmidhuber, J. A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) (1991).

[17] Barto, A., Singh, S., Chentanez, N. Intrinsically Motivated Reinforcement Learning. Advances in Neural Information Processing Systems (NeurIPS) (2004).

[18] Santucci, V. G., Baldassarre, G., Mirolli, M. Learning to Play with Intrinsically-Motivated, Self-Aware Agents. arXiv:1802.07442 (2018).

[19] Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P. World Discovery Models. arXiv:1902.07685 (2019).

[20] Shyam, P., Beer, J., Gordon, G. J. Active World Model Learning with Progressive Curiosity. arXiv:2007.07853 (2020).

[21] Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J. Learning Latent Dynamics for Planning from Pixels. arXiv:1811.04551 (2019).

[22] Corcoll, M., Mesnier, A., Martin, J., Girgin, S. Discovering and Achieving Goals via World Models. arXiv:2110.09514 (2021).

[23] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. Proximal Policy Optimization Algorithms. arXiv:1707.06347 (2017).