# 無報酬な環境での深層強化学習によるロボットの行動獲得

# A Robot Learns How To Behave With No Rewards By Deep Reinforcement Learning

妹尾卓磨 [1] *　大澤正彦 [1,2]　今井倫太 [1]
Takuma Seno[1]　Masahiko Osawa[1,2]　Michita Imai[1]

[1] 慶應義塾大学理工学部
[1] Faculty of Science and Technology, Keio University
[2] 日本学術振興会 特別研究員 (DC1)
[2] Japan Society for the Promotion of Science, Research Fellow (DC1)

**Abstract:** エージェントが他者の情報を探索的に獲得することは HAI において重要だが、報酬の定義できない探索タスクなので、従来の強化学習では困難であった。しかし、Pathak らは無報酬なゲーム環境の探索を行える Intrinsic Curiosity Module (ICM) という手法を提案している。そこで、本研究では他者の情報を探索的に獲得するエージェントを目指して、ロボットの行動獲得を ICM を利用して行い、無報酬な実環境の探索について考察する。

## 1　はじめに

エージェントが学習を行うことでユーザーに対して適応的なインタラクションを実現することができるが、あらかじめ報酬関数の定義、またはデータに対してラベルづけを行う必要がある。特に HAI の分野においては、人によってエージェントとの接し方が異なる場合や報酬関数が定義できないような場合が多いため、エージェントが学習によって適応していくことが困難であった。

機械学習を用いてエージェントが広範囲にわたるインタラクションに適応的になるには、タスクや環境に依存しない報酬関数を定義する必要がある。また、タスクに依存しないためには特徴量の設計を行わず、画像などの高次元な入力から直接学習を行う必要もある。

強化学習における従来研究では、環境からの報酬ではなくエージェント自身で報酬を生成する内発的動機モデルを用いることで、報酬がスパースな環境の探索を行なっている。Schumidhuber ら [10] は状態遷移モデルの予測誤差を内発的動機による報酬とすることで、継続的にエージェントが報酬のスパースな環境で探索を行えるとしている。深層強化学習の分野では Pathak らが Intrinsic Curiosity Module (ICM) [8] を提案しており、強化学習エージェントが ICM の生成する内発的動機によって無報酬な環境での探索を行うことができている。また、発達ロボティクスの分野の研究 [7] でも内発的動機を用いることで、同様に環境からの報酬を与えずにロボットが継続的に行動を獲得できることを示している。

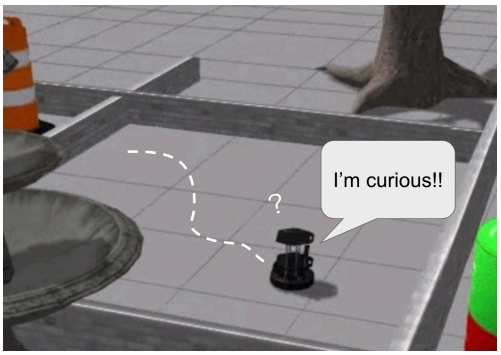


図 1: 内発的動機モデルによって探索を行うロボット

しかしながら、[8] では画像を入力として VizDoom[2] やスーパーマリオブラザーズの探索を行なっているが, 行動は離散的であるため、モーター出力のような連続行動空間を扱っていない。一般的にロボットを用いるような実環境では慣性力が働くため、連続的にモーターの出力を変化させると画像の遷移が予測できず、状態遷移のモデル化や強化学習を行うのが困難である。また、[7] の研究では入力は離散的または低次元の連続値であるため、高次元な入力は扱えておらず、学習が行えるように特徴量を設計する必要があった。

本研究では未知のインタラクションへ適応するエージェントを目指すための最初の段階として、図 1 のような ICM を用いたロボットが無報酬な環境で探索を行うため

*連絡先： 慶應義塾大学理工学部
神奈川県横浜市港北区日吉 3-14-1 26-203
E-mail: seno@ailab.ics.keio.ac.jp

の深層強化学習手法を提案する。著者らは Accumulator Based Arbitration Model (ABAM) [6, 11] を提案しており、複数の強化学習器を信頼度に応じて切り替えることができるアンサンブル学習手法である。ABAM を用いることで強化学習器の信頼度が低い場合にモーター出力を抑制し、急なロボットの動きを軽減することができるため、ICM と強化学習器の学習を容易にすることが期待できる。また、ディープニューラルネットワークで方策関数と価値関数を近似するため、画像入力から End-To-End で学習を行うことが可能である。

本論文では、2 章で提案手法を構成する関連研究について説明を行い、3 章で提案手法の詳細について述べる。4 章の実験では ROS と互換性のある物理シミュレータの Gazebo を用いて無報酬な環境を設定し、ロボット視点の画像を入力として提案手法の学習を行う。考察では提案手法を用いることでロボットが無報酬な環境であっても効率的に学習が行えていることを示す。

# 2 関連研究

ロボットの行動獲得を強化学習タスクとして考えるとロボットの取得するセンサー値を $s$、モーター出力を $a$、環境からの報酬を $r$ となる。ロボットのモーター出力は 1 次元以上の連続値である。

本章では提案手法の要素である Deep Deterministic Policy Gradient (DDPG)[3]、Intrinsic Curiosity Module (ICM)[8]、Accumulator Based Arbitration Model (ABAM)[6, 11] の説明を行う。

## 2.1 DDPG [3]

DDPG は Actor-Critic で学習を行う深層強化学習の手法である。[3] では CNN を用いて、画像入力から連続行動空間の学習を行えることが示されている。

2 つのディープニューラルネットワークを用いて方策関数 $\pi$ と行動価値関数 $Q$ を近似し、それぞれのパラメータを $\theta^{\pi}$、$\theta^{Q}$ とする。時刻 $t$ における状態を $s_t$、行動を $a_t$、報酬を $r_t$ とおくと、行動価値関数 Q のパラメータ $\theta^{Q}$ は以下の損失関数 $L_1$ を最小化するように学習を行う。

$$L_1(\theta^{Q}) = \mathbb{E}[R - Q(s_t, a_t|\theta^{Q})]^2 \tag{1}$$

ここで、$R$ は $\theta^{\pi}$、$\theta^{Q}$ と徐々に同期を行うターゲットネットワーク $\theta\pi'$、$\theta^{Q'}$ を導入して以下のように定義する。

$$R = r_{t+1} + \gamma Q(s_{t+1}, \pi(s_{t+1}|\theta^{\pi'})|\theta^{Q'}) \tag{2}$$

また、方策関数のパラメータ $\theta^{\pi}$ は以下の勾配に従って、目的関数 $J$ を最大化するように学習を行う。

$$\nabla_{\theta^{\pi}} J(\theta^{\pi}) \approx \mathbb{E}[\nabla_a Q(s_t, a_t|\theta^{Q})|_{a=\pi(s|\theta^{\pi})} \nabla_{\theta^{\pi}} \pi(s|\theta^{\pi})] \tag{3}$$

学習の安定化のために、DDPG では Deep Q-Network[5] で用いられている Experience Replay で学習を行う。Experience Replay では状態遷移 $e_t = (s_t, a_t, r_{t+1}, s_{t+1})$ を Replay Memory に保存し、ランダムにサンプリングしたものをミニバッチとして学習を行う手法である。

## 2.2 ICM[8]

ICM は逆モデルと順モデルの 2 つのディープニューラルネットワークで構成されている。逆モデルはパラメータを共有した畳み込み層を用いて $s_t$ と $s_{t+1}$ から $a_t$ を推定する。一方順モデルでは、逆モデルの隠れ層の発火状態を $\phi(s_t)$、$\phi(s_{t+1})$ とすると、$\phi(s_t)$ と実際に選択した行動 $a_t$ から $\phi(s_{t+1})$ の推定を行なっている。

内部報酬 $r^i$ は順モデルの出力 $\hat{\phi}(s_t, a_t)$ を用いて以下の式にしたがって定式化する。

$$r_t^i = \frac{1}{2}||\hat{\phi}(s_t, a_t) - \phi(s_{t+1})||_2^2 \tag{4}$$

順モデルの学習と逆モデルの学習は、逆モデルの出力を $\hat{a}$ とすると、それぞれ以下の損失関数 $L_F$、$L_I$ を最小化するように学習を行う。

$$L_I = \frac{1}{2}||\hat{a}_t - a_t||_2^2 \tag{5}$$

$$L_F = \frac{1}{2}||\hat{\phi}(s_t, a_t) - \phi(s_{t+1})||_2^2 \tag{6}$$

[8] の実験ではマルチエージェントで学習を行う Asynchronous Advantage Actor-Critic[4] の報酬生成を ICM を用いて行なっていた。しかし、本研究では単一ロボットの継続的な探索を扱うために前述した DDPG の報酬生成を ICM で行う。

## 2.3 ABAM[6, 11]

ABAM は前頭前野の知見に基づいたアンサンブル学習の手法であり、階層関係にある複数のモジュールを信頼度に応じて切り替えることが可能である。

時刻 $t$ で強化学習器の選んだ行動 $a_t$ に対応する行動の選択確率を $p_{a_t}$ とする。$p_{a_t}$ は行動空間が離散的な場合は softmax 関数で算出し、連続的な確率分布で表されている場合は確率密度関数を用いて算出する。各強化学習器に対して累積証拠 $A_t$ を割り当てる。$A_t$ は割引定数 $\gamma$ を導入し、以下の式にしたがって毎ステップ更新される。

$$A_t = \gamma A_{t-1} + p_{a_t} \tag{7}$$

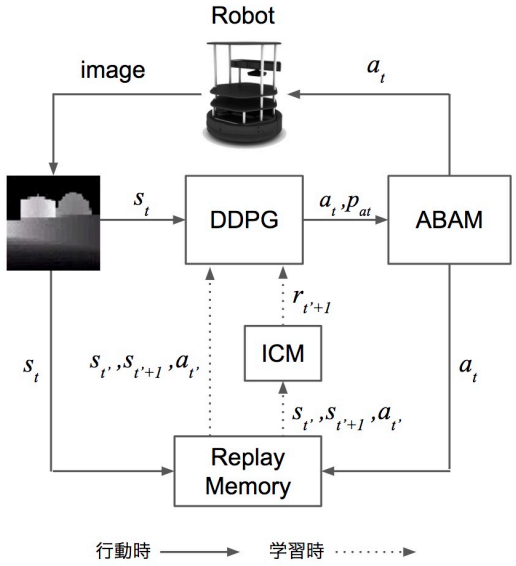


図 2: 提案手法の概要

あらかじめ設定された閾値を $A_t$ が超えている場合、対応する強化学習器の下層にあるモジュールの出力を抑制して、上層のモジュールを優先して使用する。

# 3 無報酬でロボットが探索を行う深層強化学習手法

本論文では、ロボットの無報酬な環境での探索を可能にする手法を提案する。提案手法ではロボットのモーター操作を学習するために、連続行動空間を扱える DDPG を用いて行動の学習を行う。また、本手法は無報酬な環境での継続的な探索を目指しており、Experience Replay を用いてオンラインで学習する DDPG は、オフラインの手法である Trusted Region Policy Optimization [9] よりも学習の設計が容易である。ICM が DDPG の無報酬な環境での探索を促すために内部報酬を生成し、モーター出力を ABAM を導入して抑制することで DDPG と ICM の学習を容易にする。

図 2 に概要を示す。ロボットの取得した画像情報を状態 $s_t$ として DDPG が連続値の行動 $a_t$ および対応する確率密度関数の値 $p_{a_t}$ を出力する。[3] では DDPG は決定的方策となっており、探索のために行動にノイズを付加していた。しかし提案手法では、探索は内部報酬にしたがって行われるため、確率密度関数 $p$ で表される正規分布として確率的に方策を定義する。

行動を出力する度に、ABAM の累積証拠 $A$ を式 7 にしたがって更新する。$A$ が閾値を超えた場合は行動 $a_t$ をそのままモーターの制御に使用する。$A$ が閾値を下回った場合は、急なモーター出力の変化を避けるために $a_t$ を零ベクトルとして出力する。ロボット側では行動が零ベクトルだった場合はモーターの制御を行わない。

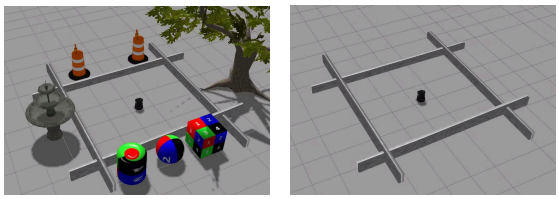

図 3: 実験に使用した環境 (左:Rich, 右:Poor)

一方、Replay Memory は各ステップで $s_t$ と ABAM の出力 $a_t$ を保存する。行動を行うたびに学習を行い、状態遷移 $s_{t'}, s_{t'+1}, a_{t'}$ を入力として ICM が内部報酬 $r_{t'+1}$ を算出し、ミニバッチを作成して DDPG と ICM の更新を行う。

# 4 評価実験

## 4.1 実験設定

提案手法を ROS と互換性のある物理シミュレーション環境である Gazebo を用いて図 3 に示す 2 つの環境で学習を行い、ABAM なしのモデルとの比較、および異なる環境ごとの比較を行なった。また、ベースラインとして学習を行わずにランダムに動くモデルとの比較も行なった。環境 Rich は仕切りの外側に四方で異なる物体を配置した。環境 Poor では仕切りの外側には物体がなく、四方で同じ景色となっている。実験設定を表 1 に示す。ロボットには Turtlebot を用いて、ABAM の

表 1: 実験設定

| ABAM \ 学習 | ICM Rich | ICM Poor | Random |
|---|---|---|---|
| ABAM あり | ABAM-Rich | ABAM-Poor | ABAM-Random |
| ABAM なし | Naive-Rich | Naive-Poor | Naive-Random |

割引率 $\gamma$ を 0.5、閾値は 0.6 で学習を行なった。入力は図 4 のような Turtlebot に取り付けられた $42 \times 42$ の大きさの深度カメラの画像を使用した。

行動は Turtlebot の進行方向と回転方向を 2 次元ベクトルで表して、それぞれ $[-1.0, 1.0]$ の値をとる。また、ICM と DDPG のディープニューラルネットワークの構成はそれぞれ [8] と [3] に記載されているものと同じにした。

各設定で 10 万ステップまで学習を行い、各ステップ $t$ におけるロボットの位置と内部報酬の値を記録した。

図 4: 入力の深度画像例

## 4.2 実験結果

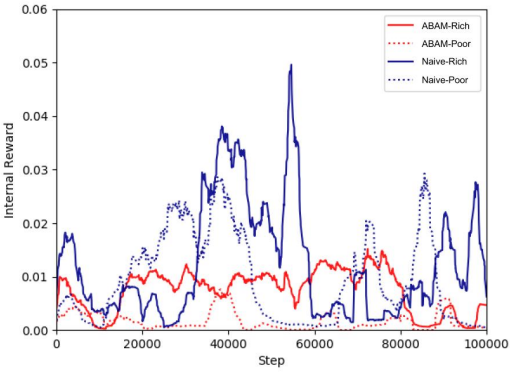


図 5: 各設定における内部報酬の変化

図 5 に各設定の内部報酬の変化を示す。ABAM なしの Naive-Rich と Naive-Poor は、ABAM ありの ABAM-Rich、ABAM-Poor と比べて比較的大きな値を示している。また、環境ごとの差という点では ABAM-Rich は ABAM-Poor よりも高い値を示している。

次に環境を $40 \times 46$ のグリッドに分割し、ロボットが訪れた回数を log スケールでヒートマップとして可視化したものを図 6 に示す。また、表 2 にはグリッド中で訪れた場所の割合を示した。ABAM-Rich がもっとも多くの場所を訪れており、ABAM-Rich、ABAM-Poor が ABAM なしの Naive-Rich、Naive-Poor よりも広範囲の場所を訪れたことが表 2 で示されている。さらに、ABAM ありの場合はランダムな探索よりも優れており、ABAM がない場合にはランダムな探索よりも狭い範囲しか探索ができなかった。また図 6 から、移動経路を比較すると、ABAM がある場合は円を描くように移動しており、ABAM がない場合は直線的に移動した様子が確認できる。

## 4.3 考察

### 4.3.1 ABAM の評価

図 6 より、ABAM ありの場合では、ABAM なしの場合よりも環境に依らずに広い範囲を探索することが

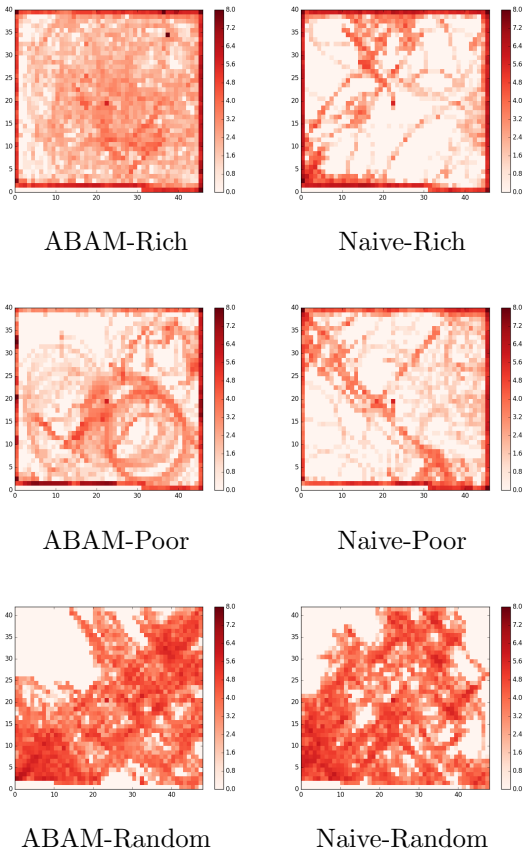


図 6: $40 \times 46$ のグリッド中で訪れた場所のヒートマップ

できた。これは、累積証拠が閾値を超えない場合はモーター出力が抑制されるため、急な出力の変化を抑えることができたので状態遷移がモデル化しやすくなったためと考えられる。図 5 に示した通り、内部報酬の値が ABAM ありの場合に比べて ABAM なしの方が大きな値を示している。内部報酬は ICM における予測誤差であるため、図 5 からも ABAM によって状態遷移がモデル化可能になったと言える。

また、ABAM がある場合とない場合で移動の様子が異なっていたのは、回転方向には慣性の力が働かないため、ABAM がない場合は逆向きの回転がかかると直ちに方向が変わるので、慣性の働く直進方向の移動が多かった。一方で ABAM がある場合では、累積証拠の閾値を越えるために一貫した行動選択が必要なため、確率的に逆向きの回転方向の出力があっても直ちには進行方向が変わらず、同じ回転方向へ移動することができた。

### 4.3.2 環境の比較

ABAM-Rich の内部報酬の値が ABAM-Poor の内部報酬の値よりも大きかったのは、図 1 のようにオブジェクトが四方で異なるため、ICM が収束せず DDPG の行

表 2: $40 \times 46$ のグリッド中で訪れた場所の割合

| 設定 | 割合 |
|---|---|
| ABAM-Rich | 92.2% |
| ABAM-Poor | 77.2% |
| ABAM-Random | 74.3% |
| Naive-Rich | 54.0% |
| Naive-Poor | 66.6% |
| Naive-Random | 73.5% |

動が継続して変化したためである。ABAM-Poor ではオブジェクトが少なく、ICM の学習が容易なため学習が進むと内部報酬の値が減少し続けた。しかし、ABAM なしの場合ではどちらの設定でも内部報酬の傾向に多きな差はなかった。これは、ABAM なしの場合ではモーターの出力が継続して変化するため、どちらの設定でも ICM のモデル化が行われなかったからである。

## 5 おわりに

ABAM と ICM を用いることで、画像入力からロボットの無報酬な環境での探索を行えることが示せた。また、ABAM を用いることで慣性力が働くようなロボットの行動や状態遷移の学習が容易になることが示せた。本研究では DDPG を用いて行動の学習を行なったが、ロボットの主観視点の画像を入力とする場合は、POMDP として一般的に扱われるため、DRQN[1] のように LSTM を使用することで性能を向上させられると考えられる。また、単純な予測誤差ではなく、[7] で行なっているような学習進捗を報酬として学習を行うことで振る舞いが変わる可能性がある。以上を考慮して今後も実験を継続していきたい。

## 参考文献


[1] Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. *CoRR, abs/1507.06527*, 2015.

[2] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*, pp. 1–8. IEEE, 2016.

[3] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *ICLR*, 2016.

[4] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 2015.

[6] Masahiko Osawa, Yuta Ashihara, Takuma Seno, Michita Imai, and Satoshi Kurihara. Accumulator based arbitration model for both supervised and reinforcement learning inspired by prefrontal cortex. *ICONIP*, 2017.

[7] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, Vol. 11, No. 2, pp. 265–286, 2007.

[8] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *ICML*, 2017.

[9] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1889–1897, 2015.

[10] J urgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior (SAB90)*, 1991.

[11] 妹尾卓磨, 大澤正彦, 今井倫太. Accumlator based arbitration model dqn: 複数モジュールを調停した深層強化学習手法. 神経回路学会全国大会, 2017.