7N-06

# 深層強化学習によるロボットの無報酬な環境の探索

妹尾 卓磨† 大澤 正彦†‡ 今井 倫太†

慶應義塾大学理工学部† 日本学術振興会 特別研究員(DC1)‡

## 1. はじめに

未知の環境ではあらかじめ報酬関数や教師データを用意できない．強化学習の分野ではエージェント内部で生成される報酬を内発的動機と呼び，次状態の予測誤差を内発的動機とすることで探索を促すアプローチがある．

深層強化学習の分野では Pathak らが提案している Intrinsic Curiosity Module(ICM) [1]で内部報酬を生成することで無報酬な環境の探索を行っている．また，発達ロボティクスの分野では[2]の研究では内発的動機によってロボットの継続的な行動獲得を行っている．

しかし，[1]の研究では離散的な行動空間のみを扱っており，連続行動空間であるモーター出力へ応用されていない．また，[2]の研究では入力が低次元なベクトルであり，画像などの高次元な入力を扱っていない．

特に内発的動機を用いて画像入力からロボットの探索を行う場合，モーター出力が大きく変化すると慣性力が働いてロボットが転倒，または振動するため次状態の予測や強化学習が困難である．そのため出力を調停することで学習を行える状態遷移を生成する必要がある．

本稿では，出力を調停することでロボットが内発的動機に従って探索を行う探索深層強化学習手法 Arbitrable Deep Intrinsically Motivated Robotic Exploration(ADIMRE)を提案する．著者らは強化学習器の信頼度に応じてモジュールを調停するアンサンブル学習法 Accumulator Based Arbitration Model (ABAM)[3,4]を提案している．ADIMRE は ABAM によって強化学習器の信頼度が低い場合は出力を抑制し，モーター出力が大きく変化するのを抑えることができる．実験では物理シミュレータ Gazebo 上でランダム探索では探索するのが困難な実環境を設定し，ADIMRE を用いることで効率的に探索が行えていることを示す．

A Robot Explores Non-reward Environments By Deep Reinforcement Learning
† Keio University
‡Japan Society for the Promotion of Science, Research Fellow (DC1)

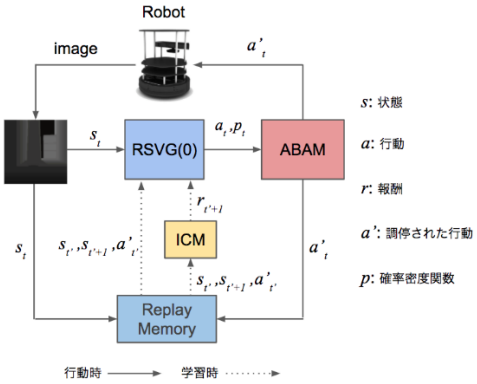


図 1: ADIMRE のモジュール構成

## 2. 関連研究

ロボットに装備されたカメラからの画像は環境全体を観測することのできない Partially Observable Markov Decision Process (POMDP) として考えられる．深層強化学習の分野では LSTM などの再帰的なニューラルネットワークを用いて学習を行うことで POMDP の環境を扱っている[5]．

著者らは先行研究[6]として，ABAM を用いてモーター出力を調停し，ICM で内部報酬を生成することで無報酬な環境の探索を扱っている．しかし，Feed Forward なディープニューラルネットワークで行動を学習しているため，複雑な POMDP 環境の学習は困難である．

本研究では LSTM を用いて学習を行う Recurrent Stochastic Value Gradients (0) (RSVG(0))[7]で行動の学習を行う．RSVG(0)は POMDP 環境で連続行動空間を学習する深層強化学習法であり，本稿では[6]では扱えないランダム探索では困難な環境での学習を扱う．

## 3. Arbitrable Deep Intrinsically Motivated Robotic Exploration (ADIMRE)

図 1 に ADIMRE のモジュール構成を示す．提案手法は ICM を用いて内部報酬を生成し，RSVG(0)の学習を行う．ABAM が出力を調停することで RSVG(0)の出力の ABAM の累積証拠$A_t$が低い場合にモーター出力を抑制し，慣性力によるロボットの転倒や振動を抑えることができる．

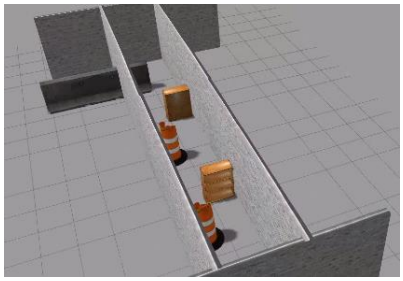

図 2:実験に使用した Gazebo 上の環境

RSVG(0)の方策は正規分布で表されており，行動価値を出力する価値ネットワークと正規分布の平均値を出力する方策ネットワークで構成されている．各時刻 $t$ で方策ネットワークの出力に基づいて正規分布を決定し，確率密度関数 $p_t$ に従って行動を決定する．選択された行動 $a_t$ の確率 $p_t(a_t)$ に従って ABAM で以下の累積証拠 $A_t$ の計算を行う．

$$A_t = \gamma A_{t-1} + p_t(a_t) \quad (1)$$

$\gamma$ は定数で与えられる割引率であり，各時刻で累積証拠の割引を行う．$A_t$ の値が与えられた閾値を超えている場合は強化学習器が信頼できるとして $a_t$ を用いてモーター制御を行い，閾値を下回っている場合はモーター出力を抑制する．

学習は Experience Replay で行う．時系列で学習を行うため，Replay Memory からランダムにエピソード中の経路の一部をサンプリングしてミニバッチを作成し，各時刻 $t$ で学習する[5]．

## 4. 実験

### 4.1 実験条件

ADIMRE を用いて，図 2 で示した物理シミュレータ Gazebo 上の環境で学習をおこなった．500 ステップを 1 エピソードとして，エピソードが終了するとスタート地点から再びエピソードを開始する．ロボットには Turtlebot を用いて入力 $s_t$ には装備されているカメラの深度画像を使用した．行動 $a_t$ は前進方向と回転方向を[1,-1]の範囲で表す 2 次元ベクトルである．評価として LSTM を用いない[6]のモデルと ABAM を除いたモデルとの比較を行った．実験条件を表 1 に示す．

表 1. 実験条件

| ABAM / LSTM | LSTM あり | LSTM なし |
|---|---|---|
| ABAM あり | ADIMRE | -LSTM [6] |
| ABAM なし | -ABAM | -ABAM-LSTM |

すべての条件において方策関数の分散は 0.01 で固定し，ABAM の閾値を 3.0，γ を 0.5 とした．各条件で 1000000 ステップの学習を行い，探索範囲について比較を行った．

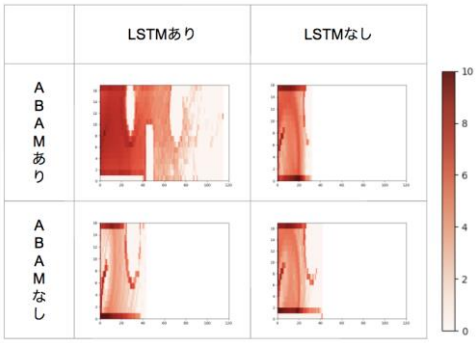


図 3:探索範囲のヒートマップ

### 4.2 実験結果

図 3 に各条件での探索範囲を log スケールで表したものを示す．ADIMRE がもっとも広い範囲を探索しており，他の条件では一つ目の障害物までしか到達できていない．幅の狭い環境では壁に詰まった際に入力画像が黒一色となってしまうので，時系列で状態遷移を学習しないと探索できないからである．

また，-ABAM の条件でも限られた範囲しか探索できていないため，ABAM によって ADIMRE では学習可能な状態遷移が生成できているといえる．

## 5. おわりに

本稿では提案手法がランダム探索では困難な POMDP 環境において探索を行えていることを示した．今後の課題として ABAM のハイパーパラメータは自動決定できないため，動的にパラメータを決定する手法を検討していく．

## 参考文献

[1] Pathak et al. (2017) “Curiosity-driven exploration by self-supervised prediction”. ICML.
[2] Oudeyer et al. (2007) “Intrinsic motivation systems for autonomous mental development”. IEEE transactions on evolutionary computation.
[3] Osawa et al. (2017) “Accumulator based arbitration model for both supervised and reinforcement learning inspired by prefrontal cortex”. ICONIP.
[4] 妹尾 et al. (2017) “Accumulator based arbitration model dqn: 複数モジュールを調停した深層強化学習法”. 神経回路学会全国大会.
[5] Hasuknecht et al. (2015) “Deep recurrent q-learning for partially observable mdps”. CoRR
[6] 妹尾 et al. (2017) “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”. HAI シンポジウム.
[7] Hess et al. (2015) “Memory-based control with recurrent neural networks”. arXiv preprint, arXiv:1512.04455