2Q-06

マルチエージェント強化学習を用いた効率的未踏領域探査のための共有情報の比較

野中 和典 †
兵庫県立大学 社会情報科学部 †

川嶋 宏彰 ‡
兵庫県立大学 情報科学研究科 ‡

1 はじめに

知能ロボットによる未知の領域の探査は，惑星調査やレスキューなど様々な分野において重要なタスクである．近年，複数ロボットによる探査では，マルチエージェント深層強化学習 (MARL) を用いたアプローチ [1] が主流となっている．一方で，これら MARL を用いた探査においては，各個体がどのような観測情報を持ち，交換すると効率的に探査できるのかは未だ明らかではない．

本研究では，各エージェントが探査することによって得る地図情報や，そのほかの観測情報，通信によって互いに共有する情報の組み合わせを複数パターン用意し，探査効率を比較することで，どのような観測情報の共有が探査を行う上で有効であるかを検証する．

2 効率的かつ効果的な情報共有

2.1 MARL による探査の定式化

本研究では，分散型部分観測マルコフ決定過程 (Dec-POMDP) による定式化 [1] を用いる．以下の記法では，i 番目のエージェントに関するパラメータ p を $p^{(i)}$，全 n エージェントの結合パラメータを $\bar{p} = (p^{(1)}, p^{(2)}, \ldots, p^{(n)})$ で表す．Dec-POMDP は $\langle \bar{S}, \bar{A}, \bar{\Omega}, \bar{O}, R, P, n, \gamma \rangle$ によって定義される．$\bar{S}$ は状態空間，$\bar{A}$ は行動空間，$\bar{\Omega}$ は観測空間であり，観測確率関数は $O^{(i)}(o^{(i)}|s, a^{(i)})$ によって表される．R は共有報酬関数，P は状態遷移確率であり $P(s'|s, \bar{a})$ で表される．n はエージェント数，γ は割引率，エージェント i は θ によってパラメータ化されたポリシー $\pi(a^{(i)}|o^{(i)})$ によって観測 $o^{(i)}$ から行動 $a^{(i)}$ を得る．そして，時刻 t において，割引累積報酬 $J(\theta) = \mathbb{E}\left[\Sigma_t \gamma^t R(s^t, \bar{a}^t)\right]$ をエージェント共同で最適化する．

ここで，本研究における実験環境について述べる．探査のシミュレーション環境として，Grid World 環境のライブラリである Minigrid [2] を用いる．探査ロボットであるエージェントは，$W \times H$ のサイズのマップの中で，上下左右の 4 近傍にのみ動くことができる．マップを構成するグリッドには通過可能領域と障害物の 2 種類が存在する．マップ情報はエージェントにとって最初は未知であり，観測を通じてマップ情報を得ていく（既知となったマップの範囲が探査済みとなる）． エージェントは障害物グリッドを通過することができず，その背後のグリッドを認識することもできない．また，エージェント同士は同じグリッドに重なることができる．

先述の定式化に基づき，本研究での探査タスクの設定を具体的に述べる．行動空間 $\bar{A}$ は同種のロボットによる探査を仮定するため，すべてのエージェントにおいて $\{$前進, 右転回, 左転回, 停止$\}$ とする．観測空間 $\bar{\Omega}$ および観測確率関数 $\bar{O}$ は次節以降で詳述する．全エージェント共有の報酬関数 R は，時刻 t における全エージェントによる探査範囲の集合を Exp_t とするとき，$Exp_t \backslash Exp_{t-1}$ とする．つまり，新たに探査した範囲を報酬として与えるように設計する．これにより，より広い範囲を探査させるように最適化させる．状態遷移確率 P に関しては，本研究では簡略化のため状態遷移は決定的な環境とする．

2.2 エージェント間の相互作用

エージェントが行動決定に用いることのできる情報 $\bar{\Omega}$ および $\bar{O}$ については以下が考えられる．

1. エージェントの時刻 t での観測情報
2. エージェントの内部（記憶）情報
3. 他エージェントから通信により得られる情報

1. はカメラやセンサ等で取得した周辺情報であり，取得後は 2. へ結合される．2. に関しては次節で述べる．3. のタイミングとその内容は様々なものが考えられるが，本研究では簡略化のため他のエージェントが持つ 2. が直ちに共有され，自エージェントの 2. へ結合されるものとする．

Shared Information Comparison for Efficient Exploration of Unexplored Areas Using Multi-Agent Reinforcement Learning

† Kazunori Nonaka, School of Social Information Science, University of Hyogo
‡ Hiroaki Kawashima, Graduate School of Information Science, University of Hyogo

図 1　本研究で用いるマップ表現．エージェントがそれぞれ内部情報に持つ．

2.3　エージェントの内部（記憶）情報

エージェントは得た情報を図 1 で示すようなマップ表現として保持するものとする．実験で用いるマップの記憶情報を以下に示す．

(a) 位置座標マップ　自他エージェントの位置およびその向き
(b) 探査済みマップ　エージェントが探査した範囲
(c) 障害物マップ　(b) の範囲内における障害物位置
(d) 軌跡マップ　自他エージェントが動いた軌跡
(e) 短期的履歴　上記すべての 1 時刻前の情報を保持することで変化情報を捉えやすくする．

2.4　観測情報の取得と処理

前節で述べた観測情報について，より精度を向上させるために以下の設定を用いる．

視界（観測範囲）　レンジセンサを想定し，自分を中心とした周囲を観測できるものとする．
前処理　観測情報を自分の進行方向が上になるよう回転させるとともに，自エージェントが中心となるように平行移動させる．

3　実験

アーキテクチャ　まず，3 層の CNN による特徴抽出器で 2.3 節のマップ表現を 12×12 の特徴表現に変換する．そして，3 層の行動デコーダにより特徴表現を 2.1 節で述べた 4 種の行動のカテゴリカル分布として出力する．パラメータの学習には，代表的な強化学習アルゴリズムである PPO を MARL に適用した MAPPO [3] を用いた．
学習条件　本研究では同種の探査ロボットを想定しているため，複数エージェント間でパラメータ共有を行い，単一ポリシーで学習を行った．12 個の部屋に区切られた 25×25 のマップを用い，バッチサイズ 8，エピソードの最大ステップ数 150，総ステップ数 10,000,000，エポック数 6，学習率 0.0005 とした．

表 1　網羅率の比較．括弧内は標準偏差

条件	(i)	(ii)	(iii)
網羅率	0.74 (0.09)	**0.80 (0.05)**	0.77 (0.07)

探査時および学習時の情報共有に関する比較

情報を共有しながら探査することの有効性を検証するため，以下の条件で実験を行った．
(i) 1 体のエージェントで学習し，3 体のエージェントが情報共有せずに探査
(ii) 1 体のエージェントで学習し，3 体のエージェントが情報共有しながら探査
(iii) 3 体のエージェントで情報共有して学習し，3 体のエージェントが情報共有しながら探査

用いた観測情報は (a) から (e) である．試行ごとにランダムに部屋の数と大きさが変わり，100 回の試行で探査可能領域に対するエージェントが探査した割合（網羅率）の平均を比較する．

表 1 より，探査時の情報共有は効果がある一方で，学習時から情報共有する (iii) の条件では，エージェントが同じところに留まり旋回を続ける傾向が見られた．未探査領域の探査行動がより効率よく進むよう報酬設計等を改善することを今後の課題とする．

謝辞　本研究の一部は科研費 JP21H05302 および JST ムーンショット型研究開発事業 JPMJMS2238 の補助を受けて行った．

参考文献

[1] Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, and Yu Wang. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. *AAMAS*, 2023.
[2] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, 2023. abs/2306.13831.
[3] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative, multi-agent games. *NeurIPS*, 2022.