様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １３ 日現在

機関番号：１５３０１
研究種目：研究活動スタート支援
研究期間：2020～2021
課題番号：２０Ｋ２３３２６

研究課題名（和文）抽象度の異なる協調行動を獲得可能なマルチエージェント強化学習

研究課題名（英文）Multi-agent Reinforcement Learning for Cooperative Policy with Different Abstraction

研究代表者
上野　史（Uwano, Fumito）

岡山大学・自然科学学域・助教

研究者番号：３０８８０６８７

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究ではまず，深層強化学習をエージェント同士で入力情報の粒度が異なるマルチエージェント環境に展開し，深層学習によって情報粒度を抽象化していることを分析により明らかにした．また，従来提案した動的環境に追従可能なマルチエージェント強化学習法を深層強化学習に展開することで，入力情報の粒度が異なる複数のエージェントによる迷路問題において最適方策を獲得することを示した．また，動的環境においては，入力情報の粒度が異なる場合，エージェント間で同期的に動くことが難しいため，提案手法の隠れ層に時系列データを学習可能なLSTMを導入し，適切に同期的に協調行動をとる方策を獲得することを明らかにした．

研究成果の学術的意義や社会的意義
本研究により，従来のマルチエージェント強化学習では取り上げられることのなかった入力情報の粒度の異なる状況に対する追従という新たな学問分野を切り開くことができた．また，実問題に即して考えてみても，例えば複数ロボットの協働制御を考えたときに，ロボットごとのセンサの粒度が異なることや，故障などの状況により得られる情報の粒度が変化することは一般的だが，マルチエージェント強化学習ではあまり考えられることがなかったため，実問題における性能がシミュレーションと比べて高くない傾向にあった．本研究成果による方法論で，マルチエージェント強化学習を実問題に応用する上での性能向上に寄与できたと考えられる．

研究成果の概要（英文）：This research analyzed deep reinforcement learning agents' performance in multiagent system with agents having different resolution in input each other to clarify the neural network can abstract the resolution appropriately. Furthermore, this research extended the previous method which enable agents to learn cooperative policy each other in dynamic environment into deep reinforcement learning to result the agents learned a cooperative policy in multiagent maze problem with agents having different resolution in input. At the end, this research introduced LSTM which can learn in time-sequential data into the proposed method to result that the agents can learn synchronously in that maze problem with environment being extended to dynamic one.

研究分野：分散人工知能

キーワード：マルチエージェントシステム　強化学習　ニューラルネットワーク　情報粒度　協調

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

近年，複数エージェントの協調行動獲得のため，目標達成時に得られる報酬を手掛かりに，人が事前に教えることなく，適切な行動を学習する強化学習メカニズムを各エージェントに導入したマルチエージェント強化学習の実用化が始まっている．しかし，マルチエージェント強化学習は，環境状況や他エージェントの行動などの情報が，全エージェントで「同じ粒度」で得られる（あるエージェントは詳細な情報を持つが，他のエージェントは粗い情報を持つなどの違いはない）前提で学習している[1]しかし，実環境を想定した場合，全エージェントで同じ粒度の情報を持つことを保証できない．例えば，カーナビ搭載の全ての車の同時経路最適化を考えると，最適経路は他車の経路選択や道路混雑に影響を受けるため，予想到着時刻はその場所に近いと正確でも，遠くなるとおおよその時刻となるため，全エージェントが同じ粒度で情報を得られない中で，適切な協調（適切な経路選択）が求められる．また，この問題は動的環境（事故発生などの環境変化，ラッシュアワーなどの車の台数の増加等のエージェント数変化）では更に困難になる．

２．研究の目的

本研究では，粒度が異なる情報をヘテロ情報とし，複数エージェントを想定したヘテロ情報の圧縮抽象化とそれに基づく協調行動学習，ならびに動的環境への展開を目的とする．具体的には，ヘテロ情報を圧縮抽象化するエージェントを設計後，そのエージェント間協調行動を実現する強化学習メカニズムを考案し，ヘテロ情報に基づくマルチエージェント学習機構を動的環境上に展開するとともに，その有効性を検証する．

３．研究の方法

本研究は，入力情報粒度の違いに対応するため，従来提案している協調行動学習手法 PMRL (Profit Minimizing Reinforcement Learning)[2]に深層強化学習を導入し，ヘテロ情報を入力とする複数エージェントの迷路問題にてその性能を検証する．深層強化学習法は A3C (Asynchronous Advantage Actor-Critic)[4]を導入し，ヘテロ情報は図 1 のように 4 つの状態を一つものとして観測し，その観測は元となる 4 つの状態から確率的に観測するものを導入する．また，提案手法を PMRL-OM (Profit Minimizing Reinforcement Learning with Oblivion of Memory)[3]にも展開し，動的環境上での有効性も検証する．

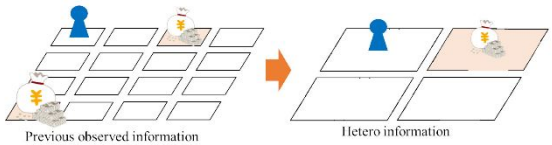

図 1　ヘテロ情報の観測

４．研究成果

まず本研究では，粒度の異なる情報を前提に協調行動学習の実現を目指し，深層強化学習の層構造と入出力情報の抽象化との関係性を検証し考察した．図 2 は迷路問題において，ノード 16 個の隠れ層を 3 層，2 層，1 層，0 層持つ深層強化学習の学習後の層のパラメータを示す．図は左から隠れ層 1 層目，2 層目，3 層目，出力層のパラメータを示し，各列上段から隠れ層を 3 層，2 層，1 層，0 層持つ場合のパラメータを示す．つまり，例えば隠れ層を 3 層持つ深層強化学習の結果は各列の最上段のパラメータを持ち，隠れ層のない場合の結果は出力層の最下段のパラメータのみを持つ．図 2 から，入力層からの距離が等しい隠れ層の重みパラメータはそれぞれ等しく，出力層のパラメータのみが異なることがわかる．以上から，深層強化学習は情報の粒度の差を吸収し協調行動学習を可能にすることがわかった．本成果は国際会議 ISAMSR 2021 にて発表した[5].

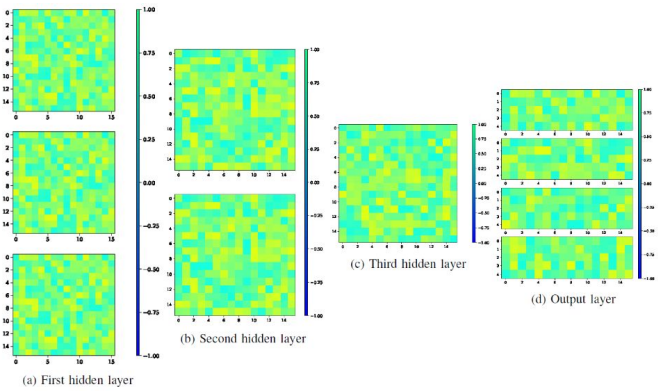

図 2　ネットワークパラメータ

次に，申請者の提案する手法である PMRL および PMRL-OM に深層強化学習を導入し，環境変化を伴う 2 体エージェントにおける迷路問題にてその性能を検証した．具体的には，図 3 に示す環境変化を伴う迷路問題を取り上げ，エージェントはゴール到達時には報酬値 10 を得るが，既に到達済みのゴールへは到達しても報酬値は得られない設定を

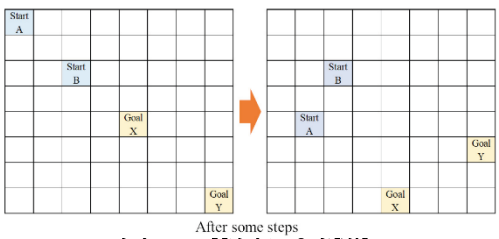

図 3　使用した迷路

おく．図 4 は，各エピソードにおいてすべてのエージェントがゴールへ到達した際に必要としたステップ数を示している．図の縦軸はゴール到達までのステップ数で横軸はエピソード数である．また，青，橙，緑の線はそれぞれ従来法として採用した深層強化学習手法である A3C，および A3C を導入した PMRL と PMRL-OM の結果を示している．環境変化は総ステップ数の半分が過ぎたときにおこるため，学習が進んでいる場合はより遅いエピソードで環境変化が起こる．結果として，PMRL-OM が最も優れた精度を示しており，A3C は環境変化の前後で適切な行動を学習できておらず，PMRL は環境変化後で適切に学習ができていないことがわかる．また，図 5 は 2 体のエージェントの獲得報酬値の推移である．図を見ると A3C と PMRL-OM はどのエージェントも最大の報酬を獲得している一方で，PMRL は右図では最終的に報酬値を獲得できなくなっていることがわかる．以上から，PMRL-OM に深層強化学習を導入することで，動的変化に対応した上で入力情報の抽象化が可能であることがわかる．本成果は国際会議 AROB 2022 にて発表した[6].

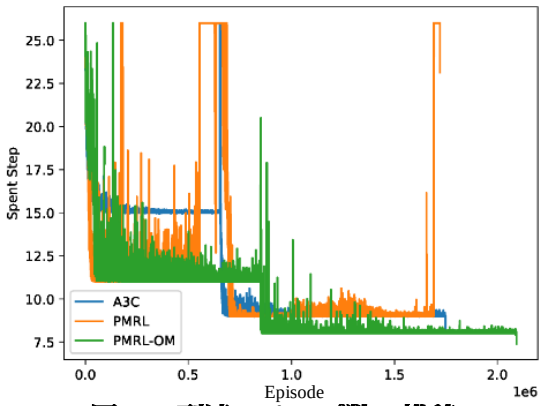

図 4　到達ステップ数の推移

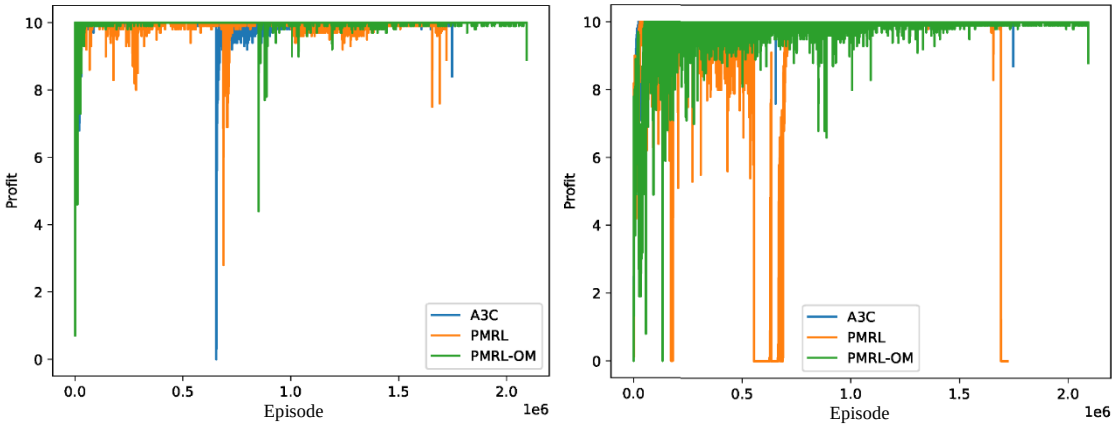

図 5　獲得報酬値（右：エージェント A，左：エージェント B）

また，前述の実験から，ヘテロ情報による動的変化では環境によりその追従性に違いがあり，粒度の荒い観測は状態遷移を繰り返す毎に適切な行動の選択確率が小さくなり，それぞれのエージェントの戦略に時間的なずれを生む PMRL-OM の想定とは異なる状況となっている．そのため手法を動的環境へ展開するため，深層強化学習に拡張した PMRL-OM の隠れ層に，時系列データを扱うニューラルネットワークである LSTM を導入することでそれに追従する新たな手法を提案し，その性能を検証した．具体的には，図 3 の環境変化を伴う迷路問題を取り上げる．問題ではエージェント A と

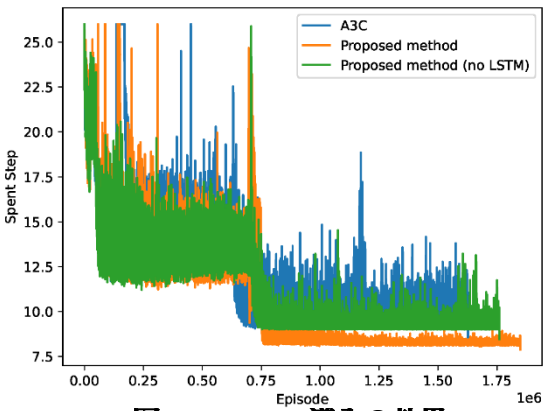

図 6　LSTM 導入の効果

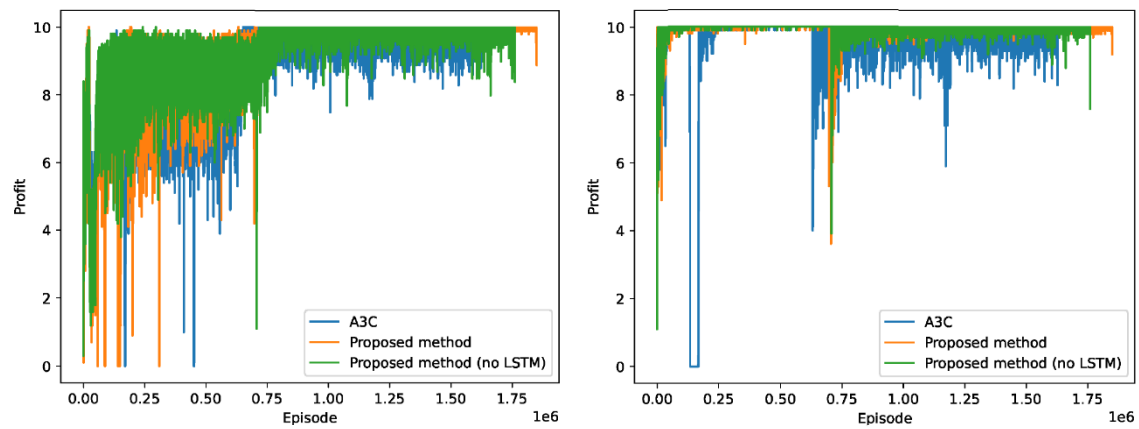

図 7　LSTM による獲得報酬値の変化（右：エージェント A，左：エージェント B）

B 双方が 16 個のノードの隠れ層を一つ持つ構造の深層強化学習器を持ち，その性能の違いをはかった．図 6 はエージェントが全ゴールへ到達できた時の最短のステップ数を示す．Proposed method が提案手法，Proposed method (no LSTM)が深層強化学習を導入した PMRL-OM を示す．これを見ると分かる通り，提案手法によって性能が向上していることがわかる．また，図 7 は同実験における 2 体のエージェントの獲得報酬値の推移である．図を見るとどのエージェントも最大の報酬を獲得しているが，収束性では提案手法が勝っていることがわかる．またこれにより，A3C 及び深層強化学習を導入した PMRL-OM ではステップ数が提案手法よりも大きいにも関わらず，報酬値をどのエージェントも獲得しており，最適な政策でゴールに到達できていないことがわかる．以上から，提案手法により動的環境においてヘテロ情報を持つエージェント同士の協調行動学習を実現した．

最後に，図 8 は前述の実験において提案手法の LSTM の挿入位置を変化させたときの結果である．hidden，policy，value はそれぞれ，隠れ層，政策の出力層の前，状態価値の出力層の前に配置した際の結果である．これを見ると，policy は環境変化後の性能が悪く，value は環境変化前の性能が悪いが，hidden はどちらの性能も最もよく，入力層の隠れ層として利用することが良いことがわかる．これは本来の目的であるヘテロな入力情報をノード数で抽象化し，それ時系列的に処理することが実現しているためである．本成果は国際会議 ICAART 2022 にて発表した[7].

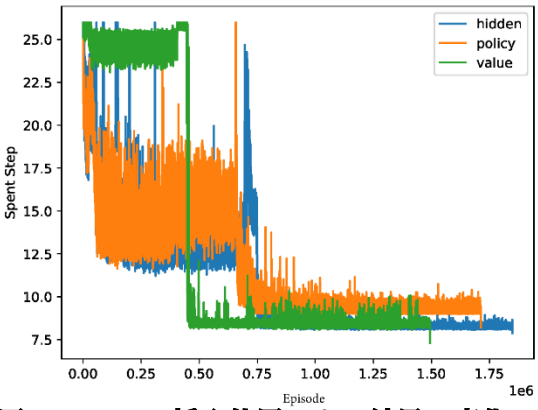

図 8　LSTM 挿入位置による結果の変化

参考文献

[1] Raileanu, R., et al., Modeling Others using Oneself in Multi-Agent Reinforcement Learning, ICML 2018, pp.4257-4266, 2018.

[2] Uwano, F., et al., Multi-Agent Cooperation Based on Reinforcement Learning with Internal Reward in Maze Problem, SICE JCMSI, pp. 321-330, 2018.

[3] Uwano, F., et al., Utilizing Observed Information for No-Communication Multi-Agent Reinforcement Learning toward Cooperation in Dynamic Environment, SICE JCMSI, pp. 199-208, 2019.

[4] Mnih, V., et al., Asynchronous Methods for Deep Reinforcement Learning, ICML 2016, pp. 1928-1937, 2016.

[5] Uwano, F., A Cooperative Learning Method for Multi-Agent System with Different Input Resolutions, ISAMSR 2021, online, September 2021.

[6] Uwano, F., Policy-oriented Goal Selection in Multi-Agent Reinforcement Learning for Dynamic Environments without Communication, AROB 2022, online, January 2022.

[7] Uwano, F., LSTM-based Abstraction of Hetero Observation and Transition in Non-Communicative Multi-Agent Reinforcement Learning, ICAART 2022, online, February 2022.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 上野 史，北島 瑛貴，髙玉 圭樹	４．巻 36
２．論文標題 多次元意見共有モデル上のシグモイド関数に基づく誤報防止アルゴリズム	５．発行年 2021年
３．雑誌名 人工知能学会論文誌	６．最初と最後の頁 B-KB2_1-12
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.36-6_B-KB2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　3件）

１．発表者名 上野　史
２．発表標題 マルチエージェントシステムにおける協調行動の抽象度と深層強化学習器の関係性の考察
３．学会等名 第48回知能システムシンポジウム
４．発表年 2021年

１．発表者名 上野　史
２．発表標題 動的環境におけるマルチエージェント強化学習―不完全な情報から集団を動かす仕組み―
３．学会等名 第6回岡山大学AI研究会（招待講演）
４．発表年 2021年

１．発表者名 上野　史
２．発表標題 A Cooperative Learning Method for Multi-Agent System with Different Input Resolutions
３．学会等名 4th International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR 2021)（国際学会）
４．発表年 2021年

１．発表者名 上野　史
２．発表標題 Policy-oriented Goal Selection in Multi-Agent Reinforcement Learning for Dynamic Environments without Communication
３．学会等名 27th International Symposium on Artificial Life and Robotics (AROB 2022)（国際学会）
４．発表年 2022年

１．発表者名 上野　史
２．発表標題 LSTM-based Abstraction of Hetero Observation and Transition in Non-Communicative Multi-Agent Reinforcement Learning
３．学会等名 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)（国際学会）
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関