安全運転の持続的な意識づけを促す
運転操作フィードバック手法に関する研究

発表者：機械知能システム学専攻 北川 浩行 (2332034)，k2332034@edu.cc.uec.ac.jp
主任指導教員：中村 友昭 准教授，副指導教員：横井 浩史 教授

1　研究の背景と目的

人間は他者とコミュニケーションすることによって，協調行動を学習することができる．また，自身の状態を記号（言語）で表現し相手に伝達することで，その記号を介して互いの状態を理解し，最適な行動を選択することができる．このように，記号を介してコミュニケーションを取ることで，互いに理解できる共有された記号が創発される過程を，創発コミュニケーションと呼ぶ．この創発コミュニケーションを確率的生成モデルで表現した手法として，谷口らによって提案されたメトロポリス・ヘイスティングス名付けゲーム (MHNG) がある [1]．筆者は MHNG をマルチエージェント強化学習に応用することで，実環境のロボットを用いた協調タスクを実行する手法を提案した [2]．さらに，これを応用したマルチステップの意思決定を伴う協調タスクを実行する手法が提案された [3]．しかし，これらの手法は離散的な状態・行動しか扱うことができないため，複雑なタスクへの応用が困難という問題がある．そこで本稿では，MHNG と深層強化学習を組み合わせることで，より複雑な協調行動を学習できるモデルを提案する．実験では，2 体のエージェントが，MHNG により創発されたメッセージを介してコミュニケーションすることで，協調行動の学習が可能であることを示す．

2　研究の方法

図 1 が提案手法のモデル構造であり，各変数の説明は表 1 の通りである．2 つの Soft Actor Critic モデルをエージェントとし，それぞれがメッセージを介して協調行動を学習する．

2.1　MHNG によるメッセージの創発

2 体のエージェントが協調行動するためには，互いの状態を伝達するための記号が必要となる．ここで，m^t という潜在変数から互いの状態 s_A^t, s_B^t と協調行動の報酬 r_m^t が生成されると仮定する．このモデルでは，m^t は互いの状態の決定に影響を与えるため，この潜在変数 m^t はメッセージと考えることができる．メッセージを創発するために，各ステップのエージェントの状態 s_A^t, s_B^t と協調行動の報酬 r_m^t からメッセージ m^t を推論する．

$$m^t \sim p(\cdot|s_A^t, s_B^t, r_m^t) \tag{1}$$

しかし，式 (1) は自身からは観測できない相手の状態が含まれており，直接計算することができない．そこで，文献 [1] と同様に MHNG を用いることで，互いに独立してメッセージを推論する．まずエージェント A が次式のように，s_A^t と r_A^t をもとに m^{t*} を生成し，B に提案する．

$$m^{t*} \sim p(\cdot|s_A^t, r_m^t) \tag{2}$$

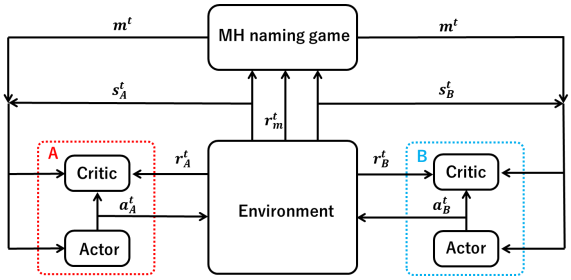

図 1: 提案モデルの概要図

表 1: 各変数の詳細

s_A^t, s_B^t	各エージェントの状態
a_A^t, a_B^t	各エージェントの行動
r_A^t, r_B^t	各エージェントの報酬
r_m^t	協調行動の報酬
m^t	エージェント間でやり取りされるメッセージ

B は提案された m^{t*} を自身の予測に基づき，次式の受理確率に従って受理または棄却する．

$$r = \frac{p(m^{t*}|s_B^t, r_m^t)}{p(m^t|s_B^t, r_m^t)} \tag{3}$$

m^t は現在のメッセージを表す．式 (3) より，A から提案された m^{t*} の受理確率は，B のパラメータのみから計算することができる．つまり，相手の状態を直接観測することなく，メッセージの受理/棄却を判断することができる．

以上の手順を役割を交代しながら繰り返し，最適なメッセージを推論する．このメッセージのやり取りをするコミュニケーションによって，2 体のエージェントが，両者の状態に応じて最適な行動を選択することができる．

2.2　状態とメッセージに基づいた行動決定

各エージェント $i \in \{A, B\}$ は，自身の状態 s_i^t と 2.1 節で推論したメッセージ m^t に基づき，次式から行動 a_i^t を選択する．

$$a_i^t \sim \pi_i(\cdot|s_i^t, m^t) \tag{4}$$

π_i は各エージェントの方策を表す．各エージェントはメッセージによって間接的に相手の状態を知ることができるため，自身と相手の状態に応じた行動を選択することができる．また，選択した行動 a_i^t に対する価値 v_i^t は，次式の行動価値関数 Q_i から算出される．

$$v_i^t = Q_i(s_i^t, a_i^t, m^t) \tag{5}$$

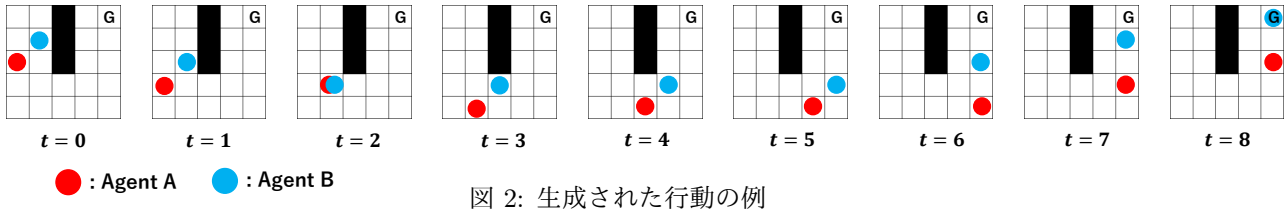

図 2: 生成された行動の例

π_i と Q_i をニューラルネットワークで近似した Soft Actor Critic を用いて，これらの関数を学習する．ネットワークのパラメータ φ_i,θ_i を用いて各エージェントの方策を π_{φ_i}，行動価値関数を Q_{θ_i} と表すと，最小化する目的関数は次式となる．

$$J_\pi(\varphi_i) = E[\alpha \log \pi_{\varphi_i}(a_i^t|s_i^t, m^t) - Q_{\theta_i}(s_i^t, a_i^t, m^t)] \quad (6)$$

$$J_Q(\theta_i) = E[\frac{1}{2}(Q_{\theta_i}(s_i^t, a_i^t, m^t) - \hat{Q}(s_i^t, a_i^t, m^t)^2] \quad (7)$$

ただし

$$\hat{Q}(s_i^t, a_i^t, m^t) = r_i^t + \beta r_m^t + \gamma E[V(s_i^{t+1}, m^{t+1})] \quad (8)$$

$$V(s_i^t) = E[Q_{\theta_i}(s_i^t, a_i^t, m^t) - \alpha log \pi_{\varphi_i}(a_i^t|s_i^t, m^t)] \quad (9)$$

である．また，α はエントロピー正則化への重み，β は協調行動の報酬の重み，γ は割引率である．それぞれの目的関数が最小となるように，勾配降下法によって以下のようにパラメータを更新する．

$$\varphi_i \leftarrow \varphi_i - \lambda_{\varphi_i} \nabla_{\varphi_i} J_\pi(\varphi_i) \quad (10)$$

$$\theta_i \leftarrow \theta_i - \lambda_{\theta_i} \nabla_{\theta_i} J_Q(\theta_i) \quad (11)$$

ただし，$\lambda_{\varphi_i}, \lambda_{\theta_i}$ は学習率である．

3 現在の結果

提案手法を用いて 2 体のエージェントが協調行動を学習できるかを検証した．図 3 のグリッドワールド空間で，互いが衝突を回避しながらゴールへ到達することを目標とする移動タスクを行った．

3.1 モデルの学習

本実験では，オフラインで事前に取得した学習データを用いてモデルを学習させた後に，協調行動が可能か検証した．まず，各エージェントのスタート地点を数ヶ所設定してランダムに行動させ，一方がゴールに到達したらスタート地点に戻すことを繰り返し，計 9840 個の学習データ $[s_i^t, a_i^t, s_i^{t+1}, r_i^t, r_m^t]$ を取得した．次に，取得したデータから，互いの状態と協調行動の関係を表現するメッセージ m^t を MHNG によって推論した．推論されたメッセージも加えたデータ $[s_i^t, a_i^t, s_i^{t+1}, r_i^t, r_m^t, m^t]$ を用いたミニバッチ学習により，方策 π_i と行動価値関数 Q_i を学習した．

3.2 協調行動の生成

学習されたパラメータを用いて，両エージェントが協調してゴールに到達できるかを検証した．まず各エージェントを移動可能グリッドにランダムに配置し，各状態に対するメッセージを MHNG によって推論した．次に，状態と推論したメッセージを各 Actor に入力し，出力された行動を実行して次の状態を決定した．以降は同様に，各状態に対するメッセージの推論と行動生成をゴールに到達するまで繰り返した．この手順を 1000 回繰り返し，スタートからゴールするまでの衝突回数の合計を算出した．実験結果を表 2 に示す．表 2 より，メッセージなしの場合と比較して，衝突回数が約 75% 減少していることが分かる．また，スタートからゴールまでの生成された行動の例を図 2 に示す．図 2 より，途中で一度衝突したものの，そこからエージェント A が衝突を避ける方向へ移動し，その後再びゴールに向かっていることが分かる．このことから，両エージェントが状況に応じて協調行動しつつ，ゴールに向かう方策を学習できていることが確認できる．

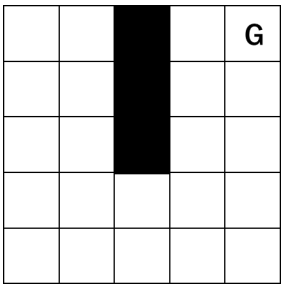

図 3: 実験環境

表 2: 実験結果

	メッセージ m あり	メッセージ m なし
衝突回数 (回)	66	252

4 まとめ及び今後の取り組み

本稿では，MHNG と深層強化学習を組み合わせることで，協調行動の学習が可能なモデルを提案した．実験では，エージェント同士がメッセージを介したコミュニケーションにより，互いの状態に応じて衝突を回避した行動を学習できることを確認した．今回は状態・行動が離散なタスクで評価したが，今後は状態・行動が連続なタスクへ適用することを考えている．また，潜在状態空間モデルへ拡張し，実世界のロボットタスクに適応することを目標としている．

参考文献

[1] Tadahiro Taniguch et al., “Emergent Communication through Metropolis-Hastings Naming Game with Deep Generative Models”, arXiv: 2205.12392, 2022.

[2] 江原広人, 中村友昭, 谷口彰, 谷口忠大, “分散的ベイズ推論としてのマルチエージェント強化学習と記号創発”, 言語処理学会第 29 回年次大会, 2023

[3] Tomoaki Nakamura, Akira Taniguchi, Tadahiro Taniguchi, “Control as Probabilistic Inference as an Emergent Communication Mechanism in Multi-Agent Reinforcement Learning”,arXiv:2307.05004, 2023