マルチエージェント強化学習を用いたデッドロック問題の解決に関する研究

TP20014 五藤強志　　指導教授：山下隆義

1. はじめに

強化学習を複数のエージェントが存在する実問題に適用する際，エージェントの相互作用を考慮する必要がある．例として複数の車両が行き交う場面での自動運転タスクなどが挙げられる．このようなタスクには，マルチエージェント強化学習 (Multi Agent Reinforcement Learning：MARL) と呼ばれる群衆の行動を学習する手法が用いられる．学習の際に，複数のエージェントが自己利益のみを優先すると，デッドロックが発生するタスクではエージェント同士が衝突し，学習の停滞や局所解への陥りが起こる．本研究では，MARL に深層学習を導入する際に，ネットワークを単一モデルとしてエージェントごとにブランチを分けて同時学習する手法を提案する．自動運転においてデッドロックが発生する環境を対象とし，提案手法の有効性を確認する．さらに，Mask-Attention 機構を導入し，学習によって獲得した行動における判断根拠を可視化し，どのように，デッドロックを回避したのかを確認する．

2. Mask-Attention A3C

Mask Attention A3C(Mask A3C)[2] は，深層強化学習の代表的な手法である Asynchronous Advantage Actor-Critic(A3C)[1] に Attention 機構を導入し，エージェントの判断根拠の視覚的説明を可能とする手法である．MaskA3C は，方策を出力する Policy branch と状態価値を出力する Value branch それぞれに Mask-Attention 機構を導入する．Mask-Attention 機構は，中間層の特徴マップ $F(s_t)$ に対して，アテンションマップ $M(s_t)$ を用いたマスク処理を行う機構であり，Mask-Attention 機構を導入して学習することにより，推論時に，判断根拠の視覚的説明となるアテンションマップが獲得できる．アテンションマップを可視化することにより，ネットワークの注視領域が確認できる．特徴マップに対するアテンションマップを用いたマスク処理を式 (1) に示す．Mask-Attention は，Feature extractor からの特徴マップに対し，$1 \times 1\times$ チャネル数の畳み込み層と sigmoid 関数により獲得している．

$$F'(s_t) = F(s_t) \cdot M(s_t) \tag{1}$$

3. 提案手法

マルチエージェント強化学習には，異なるエージェントがそれぞれの自己利益を求めてしまうことによるデッドロック問題がある．独立したネットワークを複数用いて学習を行う場合，他エージェントの学習を反映する機構がないため自己の学習のみを考慮した行動により，デッドロックの回避行動獲得が遅れてしまう．そこで，単一ネットワークでエージェントごとに Actor-Critic をブランチ分けして同時に学習を行うマルチブランチ構造を提案する．また，エージェント毎の Actor-Critic に Mask-Attention 機構を組み込むことにより，得られた方策に対する視覚的な説明を可能とする．

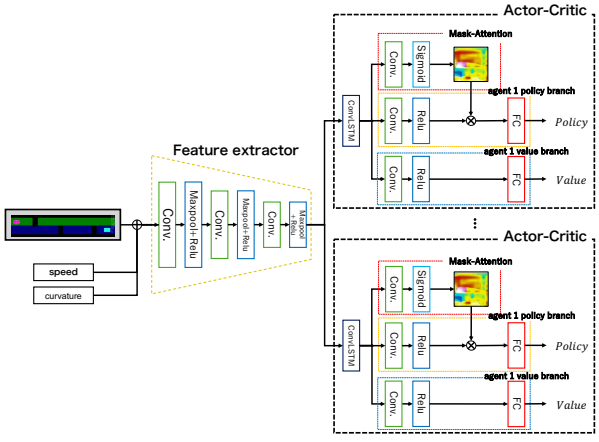

図 1: モデル全体の構造

3.1. マルチブランチ構造

提案手法のネットワーク構造を図 1 に示す．マルチブランチ構造は，Feature extractor を共有し，行動を決定する Actor-Cirtic をエージェントごとにブランチ分けする構造である．各 Actor-Critic は独立しており，それぞれ割り当てられたエージェントの行動を出力する．環境内に存在する学習を行うエージェントの数だけ Actor-Cirtic を用意し，提案手法を用いて全体の制御，学習を同時に行う．提案手法のネットワークは車両付近の鳥瞰画像，速度と曲率を入力とする．提案手法を用いることで，複数の視点を考慮することが可能となり，マルチエージェント環境における相手を考慮した最適な行動を獲得することができる．以下に提案手法の学習の流れを示す．

1. エージェントの数だけブランチ分けしたモデルを作成
2. 各エージェントの状態を対応するブランチへ入力
3. エージェント毎に報酬を蓄積
4. エージェントごとに損失関数を計算
5. 対応するブランチ及び畳み込み層に誤差を伝播

3.2. Mask-Attention の導入

提案手法は，図 1 に示すように，エージェントごとの Actor-Critic に Mask-Attention 機構を導入する．Mask-Attention 機構は，中間層の特徴マップに対して，Attention を用いたマスク処理を行う機構であり，学習によってアテンションマップが獲得可能となる．アテンションマップを可視化することにより，エージェントの注視領域を表示することができる．本研究では，エージェントの行動に対する注視領域を確認するため，Policy ブランチのみに Mask-Attention 機構を導入する．また，入力画像に対する時空間情報が欠落しないよう各 Actor-Critic には，ConvLSTM を用いる．

4. 評価実験

デッドロックが生じる 4 シーンを対象にした評価実験を行う．

4.1. 実験環境

本研究では，自動運転環境において，デッドロックが生じる 4 シーン (図 2) を対象とする．シーン 1，3，4 は表 1，シーン 2 は表 2 に示す報酬ルールに基づいて学習を行う．すべての環境は，走行レーンを正しく走行し，他車に衝突すると失敗となる環境となっている．すべての環境において，制御するエージェント数を 2 とする．行動は速度増減及び曲率増減変化なしの組み合わせで合計 9 通りであり，最大速度が 3m/s，最低速度が-1m/s，最大曲率が 0.25，最低曲率が-0.25 とする．

4.2. 実験概要

各シーンの環境において学習を行い累積報酬の比較を行う．比較手法は，通常の強化学習である A3C，MARL における従来手法である MADDPG，提案手法の計 3 つである．それぞれの手法ごとに，シーン 1 では，2.0×10^7 ステップ，シーン 2 では，2.0×10^7 ステップ，シーン 3 で

表 1: 報酬ルール 1

	報酬値
ゴール報酬	+10
中間地点に近づく	+2
衝突ペナルティ	-10
走行ゾーン離脱ペナルティ	-10
車間距離によるペナルティ	-3

表 2: 報酬ルール 2

	報酬値
ゴール報酬	+10
車両の前進	+0.5
左側通行	+1.5
衝突ペナルティ	-10
速度変化によるペナルティ	-0.5
車間距離によるペナルティ	-3

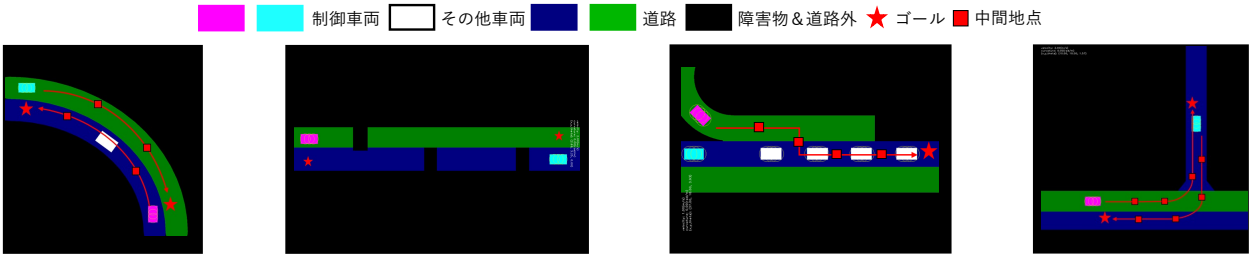

図 2: 各シーン毎の学習環境：片側一車線で，見通しの悪い道路に路駐車両が停車しているシーン (シーン 1)，走行車線に障害物があり，対向車を避けて走行するシーン (シーン 2)，幹線道路に合流するシーン (シーン 3)，狭い出入り口でのシーン (シーン 4)

表 3: 各シーンにおける獲得した累積報酬による比較 (100 エピソードの平均)

	シーン 1		シーン 2		シーン 3		シーン 4	
	エージェント 1	エージェント 2	エージェント 1	エージェント 2	エージェント 1	エージェント 2	エージェント 1	エージェント 2
提案手法	**961.5**	**580.2**	**997.0**	**994.4**	**785.0**	**585.3**	**773.4**	**589.6**
A 3 C	324.7	275.3	76.1	84.8	195.0	140.8	134.8	137.1
MADDPG	170.0	165.1	109.1	120.0	109.5	126.7	154.4	134.0

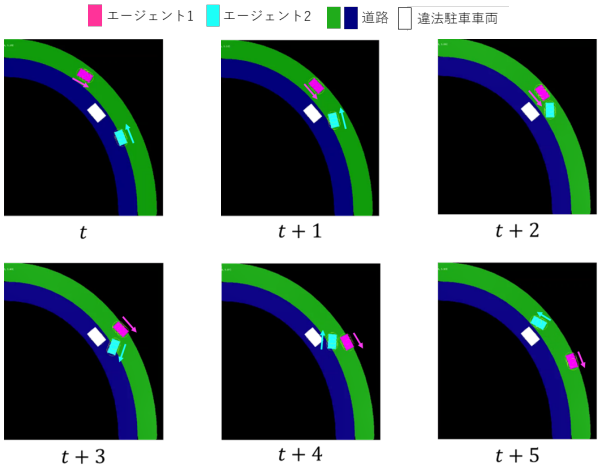

図 3: 行動の可視化結果：矢印は各車両の進行方向

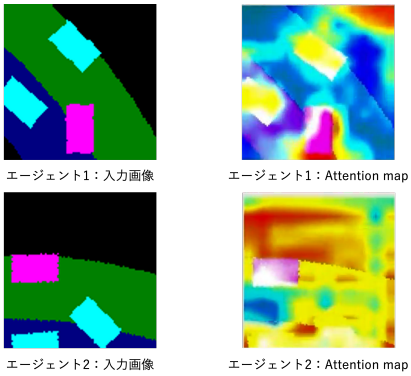

図 4: アテンションマップの可視化結果

は，5.0×10^6 ステップ，シーン 4 では，5.0×10^7 ステップ学習を行う．すべての環境において，エピソードの終了条件はエージェントが道路外，障害物及び他車両に衝突した場合と 1.0×10^5 ステップ経過で終了とする．

4.3. 累積報酬による比較

各シーン 100 エピソード評価を行った場合の平均累積報酬を表 3 に示す．表 3 から，全てのシーンにおいて提案手法を用いたエージェントが最も高い報酬を得ることができており，A3C 及び MADDPG と 300 点以上のスコア差がある．これは，提案手法のみデッドロックを回避することで，報酬を増やすことに成功しているため累積報酬が大きく向上している．逆に，A3C 及び MADDPG では，デッドロック状態を回避できず，局所解へ陥ってしまっているため (累積) 報酬が低下している．これらのことから提案手法の有効性が確認できる．

4.4. 提案手法のモデルを用いた定性的評価

シーン 1 におけるデッドロック発生時のエージェントの行動とエージェントごとのアテンションマップを図 3，4 に示す．ここで赤車両がエージェント 1 であり，青車両がエージェント 2 である．図 3 から，エージェント 2 が t+1 から t+3 まで待機し，エージェント 1 が通り過ぎると動き出している．これより，デッドロックを回避するためにエージェント 2 はエージェント 1 が通り過ぎるのを待つという相手を考慮してデッドロックを回避する行動を獲得できていると考えられる．また，図 4 から，デッドロックが発生した場面において，エージェント 1 は周囲の道と先の道を注視し，エージェント 2 は，自身と他車両を注視していることが確認できる．このことから，エージェント 1 は相手を気にせず，先に道を進むエージェントであり，エージェント 2 は相手と自分を注視し，安全運転するエージェントであると考えられる．これらのことから，提案手法を用いることで，デッドロックが発生した場合，相手を考慮してデッドロックによる衝突を回避するエージェントを獲得することが可能となる．また，MARL において，相手と異なる視点を持ち，異なる行動を行うことで，デッドロック問題を解決しているということが確認できた．

5. おわりに

本研究では，他エージェントの学習を考慮するマルチエージェント強化学習の手法を用いることで，デッドロックが発生する環境における問題を解決した．自動運転においてデッドロックが発生する環境で提案手法の有効性を示した．また，Mask-Attention を用いてエージェント毎の方策を確認したところ，相手を考慮してデッドロックを回避する行動を獲得していることも確認できた．今後の予定としては，より多様な環境における最適な行動の獲得や，より相手を考慮し，協調行動を行うモデルの実現などが挙げられる．

参考文献

[1] V. Mnih , *et al.*, “Asynchronous methods for deep reinforcement learning”, ICML, 2016.

[2] H. Itaya, *et al.*, “Visual Explanation using Attention Mechanism in Actor-Critic-based Deep Reinforcement Learning”, IJCNN, 2021.

研究業績

[1] 五藤強志 等, “インタラクションを考慮したマルチブランチネットワークによる深層強化学習”, 人工知能学会全国大会, 2020.