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

Multi-Agent Deep Reinforcement Learning with Multi-Branch Networks Considering Interaction

五藤 強志 *1 Tsuyoshi Goto　平川 翼 *1 Tsubasa Hirakawa　山下 隆義 *1 Takayoshi Yamashita　藤吉 弘亘 *1 Hironobu Fujiyoshi

*1中部大学
Chubu University

When multiple agents are in the same environment, collisions may happen with each other. Because the agents consider their own interests or have a negative effect on other agents. In situation that happens these deadlocks, agents should select the action considering other agents based on multi-agent reinforcement learning which train multiple agents simultaneously. In this paper, we propose the method that trains multiple agents in a network with multi-branch network for this problem. It is possible to train an interaction between agents. In experiment, we build the environment that happens the deadlock between self-driving cars and compare with independent network of each agent. Mereover, we show the behavior of the agent in a deadlock situation.

1. はじめに

強化学習は，エージェントが試行錯誤を通じて価値を最大化するような方策を見つける問題である．また，教師あり学習とは異なり，事前に正解が与えられるのではなく，自身の行動の良し悪しをもとに最適な行動を見つけ出す手法である．

強化学習は教師信号の作成が困難なロボット制御 [Gu 17] やゲーム攻略 [Oriol 17] など様々なタスクに応用されている．近年では，Google の DeepMind が開発したコンピュータ囲碁プログラムである AlphaGo[Silver 16] に強化学習が用いられ，プロの棋士に勝利するなどゲーム攻略において注目された．また，強化学習に深層学習を組み合わせた深層強化学習が主流となっており，Deep Convolutional Neural Network (DCNN)[LeCun 98] と強化学習の代表的な手法である Q 学習 [Watkins 92] を組み合わせた Deep Q-Network (DQN)[Mnih 13][Mnih 15] は，Atari2600 のゲーム攻略において人間を上回る性能を実現した．また，Actor Critic 法と woker と呼ばれる並列で経験を収集する機構を用いた分散学習を組み合わせた手法である Asynchronous Advantage Acotr Critic (A3C)[Mnih 16] は，膨大な数の状態数を持つ連続空間の問題においても有効性を示している．このように，深層強化学習の登場により状態数が膨大なタスクにおいても解決可能となった．

様々なタスクにおける強化学習の応用が期待されているが，実問題へ強化学習を適用する際，環境に複数のエージェントが存在するケースが多く，エージェント同士の相互作用を考慮する必要がある．そこで，マルチエージェント強化学習という手法が用いられる．マルチエージェント強化学習は，複数のエージェントを同時に学習する手法であり，群衆における方策の獲得が可能である．しかし，複数のエージェントが存在する場合，自己利益のみを優先しエージェント同士が衝突することで，学習の停滞や局所解への陥りなど効率的な学習が困難となる．そのため，エージェント同士の衝突を避けて学習を行う仕組みが必要である．本論文では，マルチエージェント強化学習に深層学習を導入する際に，ネットワークを単一としてエージェントごとに複数のブランチを分けて同時学習する手法を提案する．これにより，エージェント同士のインタラクションを考慮しつつ，学習を行うことが可能となる．独自に作成したデッドロックが発生する環境において，各エージェントを独立したネットワークで学習する手法と提案手法を比較し，提案手法の有効性を確認する．また，提案手法の学習済みモデルによる行動の可視化を行い，どのような行動を獲得できたかを確認する．

連絡先:
五藤 強志：tgo96@mprg.cs.chubu.ac.jp
平川 翼：hirakawa@mprg.cs.chubu.ac.jp
山下 隆義：takayoshi@isc.chubu.ac.jp
藤吉 弘亘：fujiyoshi@isc.chubu.ac.jp

2. 関連研究

複数のエージェントを同時に学習し，群衆の方策獲得を図るマルチエージェント強化学習の手法は様々提案されている．主に提案されている手法は 2 つある．1 つ目は，通信機構を導入し，各エージェントのネットワークの内部情報を他のエージェントに渡すことで自エージェント以外を考慮した学習を行う手法である．2 つ目は，環境報酬とは別に学習全体を考慮した報酬を設計することで自エージェント以外を考慮した学習を行う手法である．

エージェント同士で通信を行うことで群衆の方策獲得を行う手法として，Sainbayar ら [Sainbayar 16] は，全エージェントの状態と行動のベクトルを入力とし，全エージェントの内部状態を平均した通信ベクトルを全てのネットワークへ入力する機構を導入することで，全エージェントが他エージェントの内部状態を把握し，マルチエージェント環境における他者を考慮した学習を実現している．

報酬設計によってを群衆の方策獲得を行う手法として，Natarajan ら [Natarajan 10] は，逆強化学習を用いることにより，マルチエージェント環境における複数エージェントの相互作用を考慮するような報酬を導き出し，それを用いた学習によりエージェントの衝突回避を実現している．Natasha ら [Natasha 19] は，Social Influence という他者への影響が大きい行動に対する報酬を用意し，3 つのモデルを用いて学習を行うことで他者との協力行動の創発を実現している．1 つ目は，Social Influence を用いて他者へ影響を与える行動を学習する influecer と環境報酬を用いて学習する influencee を学習していく Basic Influence というモデルである．2 つ目は，エージェン

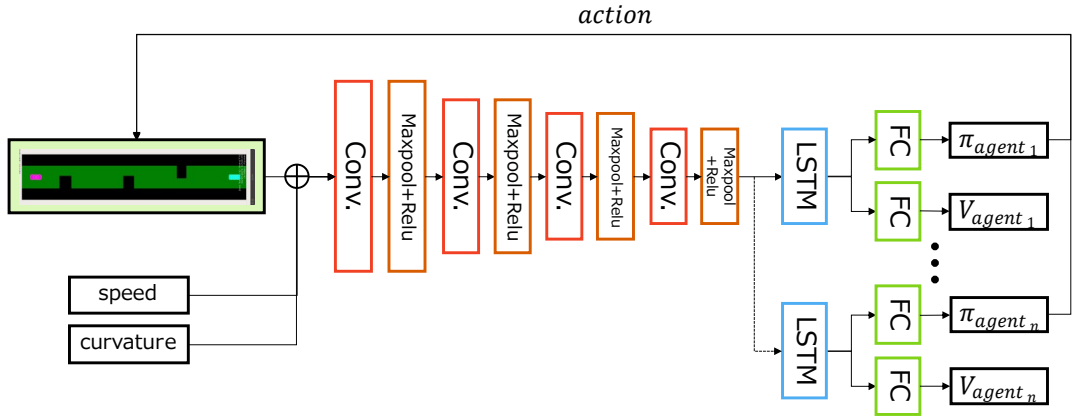

図 1: 提案手法のネットワークモデル

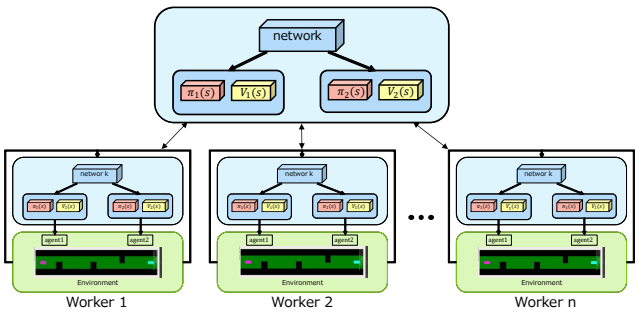

図 2: A3C に基づいた提案手法の学習方法

トに通信機構を導入し Social Influence を用いて他者への通信を学習する Influential Communication というモデルである．3 つ目は相手の行動を予測する機構を導入し，Social Influence を用いて予測機構を学習する Modeling Other Agents というモデルである．

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自己の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ブランチ分けし同時学習を行う手法を提案する．

3.1 ネットワーク構造

提案手法のネットワークモデルを図 1 に示す．ネットワークには車両の鳥瞰画像，速度と曲率を入力する．提案手法の構造は，単一ネットワークでエージェントごとに方策と状態価値の出力層をブランチ分けし，全てのエージェントが畳み込み層を共有するようなネットワークとする．各エージェントはそれぞれ対応したブランチからの出力を制御値とする．これにより，他エージェントの学習を自分の学習に反映し，他エージェントとのインタラクションを考慮した学習が可能となる．また，デッドロックが発生した場合，デッドロックを回避する行動獲得の促進を図る．

3.2 学習方法

本論文では，強化学習手法として A3C[Mnih 16] の枠組みを用いる．提案手法を A3C に適用させた場合の構造を図 2 に示す．提案手法のネットワークを各 worker に対応する local network 及び global network として学習を行う．この時，ネットワークは環境内に存在するエージェント数に応じてブランチを分ける．各 worker 環境のエージェントは local network のそれぞれ対応するブランチの出力を制御値として行動する．一定ステップで勾配を global network に送り local network を global network と同期する．

3.3 提案手法の流れ

提案手法を用いて環境内に存在する全てのエージェントを制御し，全体の学習を同時に行う．提案手法の流れを以下に示す．

複数のエージェントが存在する環境においてエージェントの数に応じてブランチを分けた提案手法のネットワークを作成する．環境内に存在する各エージェントの状態を観測し，対応するブランチへエージェント毎に状態を入力する．各ブランチからの出力を用いて対応するエージェントを制御する．各エージェントごとに報酬を蓄積し，対応するブランチ及び畳み込み層の逆伝播を行う．

4. 評価実験

4.1 実験環境

本論文では，自動運転におけるデッドロックを想定した環境を用いる．自動運転環境は，各車両を操作し障害物をよけながら画面端のゴールへ向かって走行する環境である．各自動車は，エージェントとして走行するレーンは固定とする．障害物は車両が 2 台以上通れない程度の間隔を空けてランダムに設置する．

エージェントの報酬設計として，スタート地点から画面端のゴールへの到着による報酬，車両の前進による報酬，左側通行による報酬を正の報酬として設計する．道路外，障害物及び他車両との衝突ペナルティ，急発進や急停車による速度変化のペナルティ，車間距離が一定以下でのペナルティを負の報酬として設計する．各報酬値を表 1，2 に示す．

エージェントの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，最低曲率が 0m/s，最大曲率が 0.25，最低曲率が-0.25 とした．また，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．

表 1: 正の報酬

	報酬値
ゴール報酬	+10
車両の前進	+0.5
左側通行	+1.5

表 2: 負の報酬

	報酬値
衝突ペナルティ	-10
速度変化ペナルティ	-0.5
車間距離ペナルティ	-3

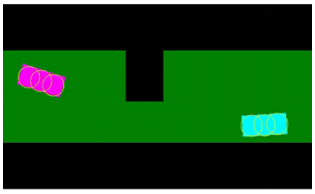

表 3: 衝突回数

	衝突回数
共有なし	997
共有あり	602

図 3: 評価環境

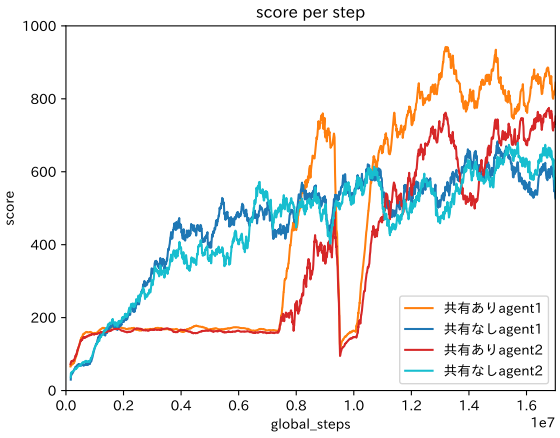

図 4: 学習ステップ数毎のスコア推移

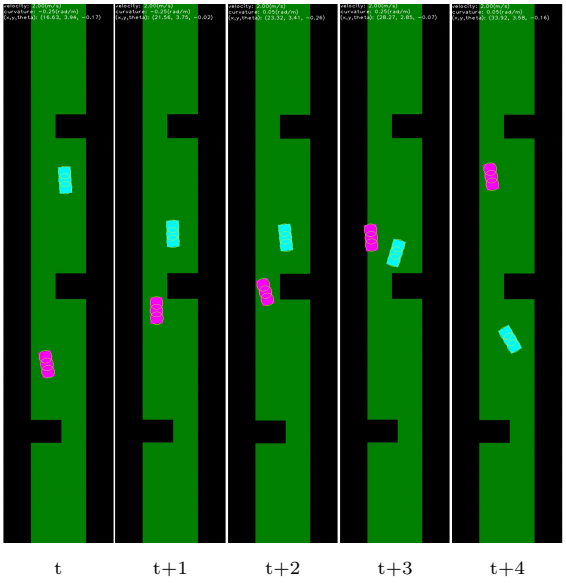

図 5: 提案手法により獲得された行動の可視化結果

4.2 評価指標

前述の自動運転環境におけるステップ数ごとのスコア推移と評価環境における 1000 試行での衝突回数を比較することで，提案手法の有効性を確認する．また，提案手法による学習済みモデルの行動を可視化し，デッドロックによる衝突を回避するエージェントが獲得できているか確認する．比較手法として独立したネットワークを用いて学習する場合 (ネットワーク共有なし)，提案手法 (ネットワーク共有あり) の 2 通りで学習を行う．A3C の worker 数は 7 とし，各手法を 1.0×10^7 ステップ学習する．デッドロックによる衝突回数の比較を行うために，自動運転環境におけるデッドロックを誘発する環境を作成して用いる．作成した評価環境を図 3 に示す．長さを制限した道路に障害物を配置した環境である．終了条件は両端に配置した車両が両車両とも衝突することなく障害物を通り抜けることができた場合に終了とする．

4.3 実験結果

4.3.1 スコア推移による比較

自動運転環境における共有なし agent 1，2 及び共有あり agent 1，2 のステップ数ごとのスコア推移を図 4 に示す．図 4 から，共有ありは共有なしと比較してより高いスコアを獲得した．このことから，デッドロック状態を解消する学習が可能であり，マルチエージェント環境における提案手法の有効性が確認できる．また，共有なし agent1，2 のスコアは両エージェント共に約 500 で差がないのに対し，共有あり agent1，2 でのスコアの差は約 100 となり，agent1 が高いスコアを獲得している．これは，agent2 が agent1 を考慮した学習を行い，デッドロック状態における相手を優先する行動を取ったことによるスコアの低下であると考えられる．このことから，提案手法を用いることにより相手を考慮し，より高い利益を獲得する行動の学習を実現した．

4.3.2 衝突回数による比較

デッドロックを誘発する評価環境における共有なし agent と共有あり agent の衝突回数を表 3 に示す．表 3 から，共有なし agent と比較し共有あり agent の衝突回数が 395 回少ないことが確認できる．このことから，提案手法を用いることでデッドロックによる衝突を回避するエージェントの獲得を実現した．

4.3.3 学習済みモデルを用いた定性的結果

提案手法により獲得された行動を図 5 に示す．ここで赤車両が agent1，青車両が agent2 である．図 5 から，agent2 が t+1 から t+2 まで停止して agent1 が障害物を通り過ぎた後動き出している．これはデッドロックを回避するために agent2 は agent1 が通り過ぎるのを待つという行動を獲得できているためと考えられる．このことから，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の獲得を実現した．

4.4 バック行動を追加した評価実験

4.4.1 実験概要

バック行動を追加して，よりデッドロック状態を回避できるように学習を行う．行動の最低速度を-1m/s に変更し，報酬にバック行動によるペナルティ-3 を加える．バック行動によるペナルティは緊急時のみにバック行動を起こすようにするためである．学習回数は 1.0×10^7 ステップまで行う．学習済みモデルの行動を可視化し，デッドロックによる衝突を回避できているかを確認する．

4.4.2 バック行動を追加した場合における定性評価

バック行動を追加した実験において獲得された行動を図 6 に示す．ここで赤車両が agent1 で青車両が agent2 である．図 6 から t+1 から t+6 までの間にデッドロックが発生し，agent2 が t+4 から t+5 でバック行動を行いデッドロックによる衝突

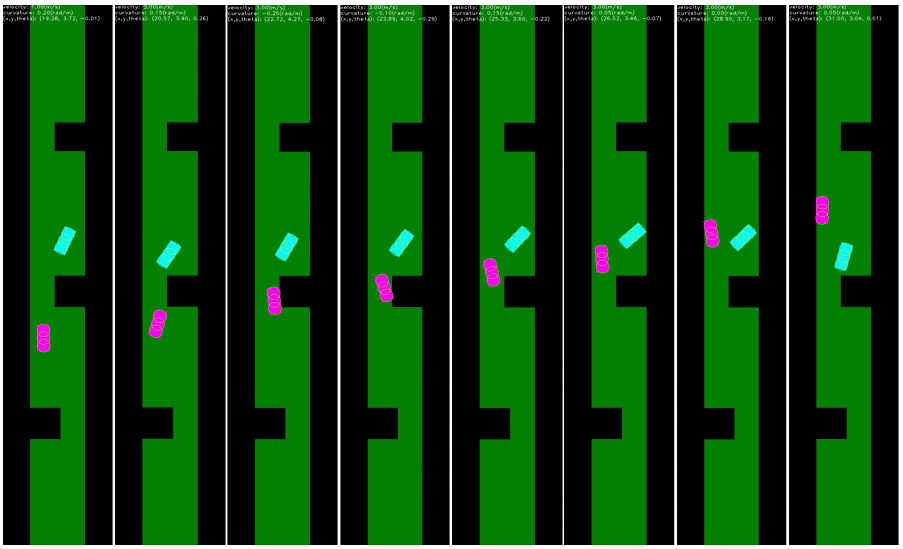

図 6: バック行動を追加した場合の可視化結果

を回避していることがわかる．このことから，バック行動を追加しても相手を考慮した学習を行うことができ，より衝突を避けるエージェントの獲得を実現した．

5. おわりに

本論文では，他エージェントの学習を考慮してデッドロックを回避する行動を獲得する手法を提案した．提案手法では，単一ネットワークでエージェントごとに方策と状態価値の出力層をブランチ分けし同時学習を行うことで，デッドロック状態を解消し学習の効率化を実現した．また，自動運転環境における学習済みモデルの行動の可視化により，相手を考慮した方策を持つエージェントの獲得を示した．今後の予定としては，他環境における提案手法の評価や，報酬以外を考慮するモデルの実現などが挙げられる．

謝辞

本研究は，総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム (SIP) 第 2 期/自動運転 (システムとサービスの拡張)「自動運転技術 (レベル 3，4) に必要な認識技術等に関する研究」(管理法人：NEDO) によって実施されました．

参考文献

[Gu 17] Gu, S., Holly, E., *et al.*: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, in *International Conference on Robotics and Automation*, pp. 3389–339 (2017)

[LeCun 98] LeCun, Y., Bottou, L., *et al.*: Gradient-based learning applied to document recognition, *IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998)

[Mnih 13] Mnih, V., Kavukcuoglu, K., *et al.*: Playing Atari with Deep Reinforcement Learning, in *National Institute for Physiological Sciences* (2013)

[Mnih 15] Mnih, V., Kavukcuoglu, K., *et al.*: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Mnih 16] Mnih, V., Badia, A. P., *et al.*: Asynchronous Methods for Deep Reinforcement Learning, in *International Conference on Machine Learning*, pp. 1928–1937 (2016)

[Natarajan 10] Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., and Shavlik, J. W.: Multi-Agent Inverse Reinforcement Learning, in *The Ninth International Conference on Machine Learning and Applications*, pp. 395–400 (2010)

[Natasha 19] Natasha, J., Angeliki, L., *et al.*: Social Influence as Intrinsic Motivation for Multi-Agent Deep ReinforcementLearning, in *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049 (2019)

[Oriol 17] Oriol, V., Timo, E., *et al.*: StarCraft II: A New Challenge for Reinforcement Learning, *CoRR*, Vol. abs/1708.04782, (2017)

[Sainbayar 16] Sainbayar, S., Rob, F., *et al.*: Learning multiagent communication with backpropagation, in *Advances in neural information processing systems*, pp. 2242–2252 (2016)

[Silver 16] Silver, D., Huang, A., *et al.*: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587, p. 484 (2016)

[Watkins 92] Watkins, C. J. and Dayan, P.: Q-Learning, *Machine learning*, Vol. 8, No. 3-4, pp. 279–292 (1992)