第 29 回日本バーチャルリアリティ学会大会論文集（2024 年 9 月）

Unity ML-Agents を用いた自動運転強化学習おける事故回避シミュレーションの提案

Proposal for accident-avoidance simulation in autonomous driving Reinforcement learning using Unity ML-Agents

増田琉利 [1)]，加藤央祐 [2)]，水谷賢史 [3)]
Ryuto MASUDA, Osuke KATO, and Kenji MIZUTANI

1) 東海大学大学院工学研究科医用生体工学専攻 （〒259-1193 神奈川県伊勢原市下糟屋 143, 4ceym008@tokai.ac.jp）
2) 東海大学工学部医用生体工学科 （〒259-1193 神奈川県伊勢原市下糟屋 143, 0cey1112@mail.u-tokai.ac.jp）
3) 東海大学大学院工学研究科医用生体工学専攻 （〒259-1193 神奈川県伊勢原市下糟屋 143, mk3149@tokai.ac.jp）

概要：自動運転技術は急速に進展しており、国内でも実証実験が行われている。しかし障害物回避や環境変化への適応など課題が残っている。本研究では、まずレースサーキット環境で障害物有りと無しの環境で学習させ、その性能を評価する。次に、サーキット環境で学習させたものと事前学習なしのエージェントを高速道路合流部環境において、他車との衝突回避を学習させ、平均累積エピソード報酬を評価し事前学習の効果を検証した。

キーワード：自動運転，強化学習，サーキット，高速道路

1. はじめに

近年完全自動運転に向けて自動運転技術は急速に拡大している。国内でも 2023 年度にレベル 4 の自動運転車の公道走行「Zen Drive」[1]などが開始され、期待が高まっている。しかし一方で、障害物回避[2]や複雑な環境の変化に対する適応[3]など様々な課題がある。これらの課題を解決するためには、車両が環境を正確に探索し、適切な行動を選択する必要があり、その手段として強化学習が有効である。今回は環境を容易にシミュレートでき交通シナリオをリアルタイムで評価できる Unity ML-Agents[Unity 社]を使用して、自動運転車の強化学習シミュレーションモデルを提案する。

次に道路交通事故の中で高速道路における追突事故は全事故の 42.7%を占めており、追突による事故は多くの後続車両を巻き込む重大な事故を誘発させる。その中でも日本警視庁の調査データ[4]によると、合流時の事故は高い割合を示しており特に速度の違いや車間距離不足、合流のタイミングミスなどが原因で発生する。これらの課題を解決するために高速道路の合流に対しての研究が盛んに行われている。Bouton ら[5]は強化学習カリキュラムを使用することで、密集した交通環境での合流シナリオにおいて、エージェントがより効率的に学習し、訓練中に見たことのない新しい状況にも対応できるようになり、効率的で汎用性の高いポリシーを学習することができる点を強調している。シングルエージェントの手法として Triest ら[6]は、強化学習を使用して、高レベルの意思決定と低レベルな制御器を用いた階層的なアプローチを提案し、他のアプローチ方法よりも低い衝突率で合流操作を実行できることを示した。次にマルチエージェント手法として、S.Chen ら[7]は、出口ランプからの高速道路への出口を目指すため、グラフ畳み込みニューラルネットワーク（GCN）と深層 Q ネットワーク（DQN）を組み合わせた「グラフ畳み込み Q ネットワーク（GCQ）」というモデルを提案している。この手法ではパラメータが少なく、より高速に学習できるとされており、既存のルールベースの手法の結果を有意に上回る結果となった。Y.Chen ら[8]は混合交通環境下での高速道路合流エリアにおける複数車両の協調合流戦略を提案している。高速道路の合流エリアに専用の CAV（Connected Autonomous Vehicles）と人間が運転する車両（HVs）が混在する新しい交通環境に対応する点を新規性としており、エコドライビングを考慮した協調合流モデルの構築を提案している。このようにマルチエージェント型の学習を取り入れることでより高度な自動運転システムの実現が期待される。

Shafique ら[2]は従来の PPO 設定と改良 PPO 設定を用いたシングルエージェント学習を 4 つの環境（障害物の有る/無しの単純/複雑なサーキット）で実施し、いずれも改良型 PPO 設定がより良い結果を示すことを報告している。本稿では、まず障害物有りと障害物無しの単純なサーキット環境で安定するまで事前学習を行い、パラメータを引き継いで、高速道路合流部環境を試験対象とし、エージェントに加速車線で十分に加速し、本線で走行する車両に衝突しないように合流する学習を行い、事前学習の効果について報告する。

2. 実験原理

2.1 開発環境

アプリケーションとしては Unity(Unity Technologies)の 2020.3.38f1)を用いた。また強化学習のオープンソースプロジェクトとして Unity ML-Agents(Unity ML-Agents release_19)を用いた。機械学習用のフレームワークとして、Python の Pytorch(python3.8)を使用した。

2.2 エージェントの観察と行動

まずはエージェントの観察について説明する。Ray Perception sensor を利用した Ray cast Observation という観察方法を使用する。実装によりエージェントは壁、障害物などの要素を見分けることができる。図 1 に示すように指定したオブジェクトにヒットすれば赤く反応し、オブジェクトにヒットしていなければ白く表示される。レイの個数やレイをどの方向に照射するかは人間が決定する。次に行動について説明する。エージェントは 3D の car モデルを使用し、1)前進、2)後退、3)左旋回、4)右旋回の 4 種類の行動を実行することができる。

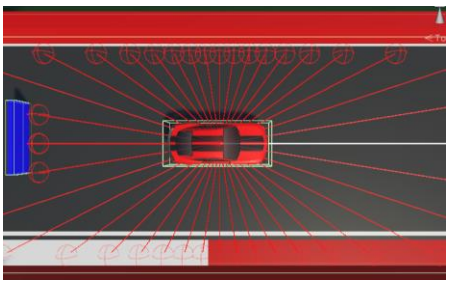

図 1: Ray Perception sensor

2.3 エージェントの報酬設定

エージェントのスクリプトによる報酬設定は強化学習を通してエージェントを訓練させるうえで非常に重要であり、エージェントが取得する同じ「行動」でも「状態」によって結果は大きく異なる。エージェントが適切な結果を得るためには実行を繰り返し、最適化された報酬設定が必要である。以下に使用した報酬設定の概要を記述する。

① 前方方向への速度制限 : 車の速度が指定値以上になった場合、正の報酬を与え、速度を指定値に制限する。これにより速度が適切な範囲に保てるようにする。

② 後退速度の制限 : 車が後退している場合、速度を指定値に制限する。これにより後退が過度になるのを防ぐ。

③ 前方方向への制限 : 前進速度が指定値以上になったら正の報酬を与える。これは車が適切な前進速度を維持するための制御である。

④ 衝突時の処理 : 壁、車、障害物に衝突したときはそれぞれに対応した負の報酬を与える。ただしチェックポイントに到達したときは正の報酬を与える。

3. 実験方法

3.1 サーキット環境でのシミュレーション

まず障害物有りと障害物無しのレース環境を作成した(図 2)。2.4 節に示す報酬設定に基づき、障害物有りと無しのコースを事前学習として学習させる。学習が完了した（一定の報酬値に到達した）時点でサーキットでの学習は終了とした。

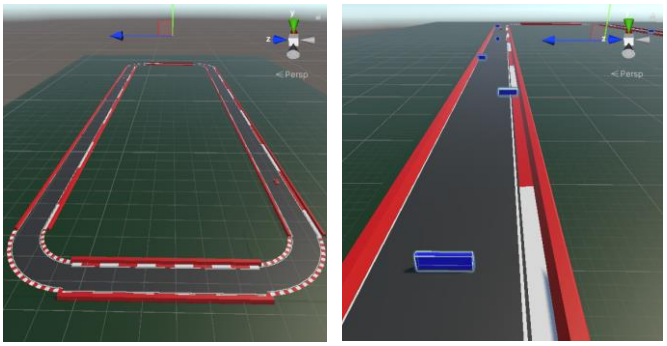

図 2: サーキット環境（左:障害物無し、右:障害物有り）

3.2 高速道路合流部環境でのシミュレーション

今回はシングルエージェント（合流側）での学習とした。エージェントを加速車線で十分に加速させたのち、本線に走行する車両に衝突しないように学習させることを目的とした。高速道路合流部環境の設計を図 3 に示した。高速道路の設計としては本線の幅を 3.5m、白線の幅を 0.15m、加速車線の長さを 200m に設定した。本線車の速度は左車線が 90km/h、右車線が 100km/h とし、エージェントは最大 100km/h 制限とした。具体的な実験内容としては、サーキット環境で障害物有りと障害物無しで、学習が完了したエージェントを用意し、高速道路合流部環境で直接学習させたエージェントとの比較を行う。その際評価方法として、「深層学習フレームワーク」が出力する学習状況の統計情報を可視化するツールである Tensor Board[9]を使用し、移動平均をとった平均累積エピソード報酬を示すグラフを評価する。

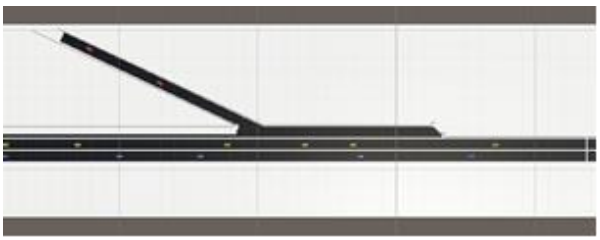

図 3: 高速道路合流部環境

4. 実験結果

Tensor Board によって確認した事前学習環境および高速道路合流部環境での平均累積エピソード報酬の結果をそれぞれ図 4 と図 5 に示しており、いずれも移動平均をとった結果である。

図 5 から赤色のグラフでは、初期段階で負の報酬からスタートし、その後徐々に累積報酬が増加していることや他の 2 つのグラフよりも低い範囲にとどまっていることが読み取れる。緑色のグラフでは、図 4 から障害物無しの環境で、約 28 万 5 千ステップで学習が完了した。高速道路合流部環境への移行後、報酬が急激に向上し、延べ約 50 万ステップで安定していることが読み取れる。障害物無しのサーキット環境での学習が効を奏し、高速道路合流部環境への移行後にすぐに適応できていることが示唆される。図 4 から障害物ありのサーキット環境で、約 50 万ステップで学習が完了した。高速道路合流部環境への移行後、50 万ステップから約 60 万ステップの間で報酬が低いまま安定していることが確認できた。60 万ステップ以降、報酬は向上するが、約 65 万ステップでしばらく安定し、次に約 88 万ステップで再上昇して、最終的には他 2 種類の学習に比べ高い報酬が得られ安定した。

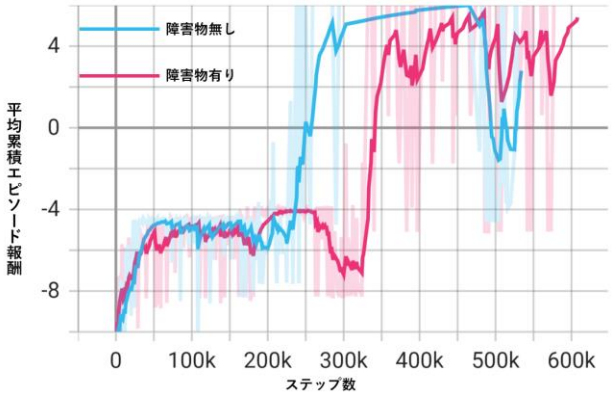

図 4： サーキット環境での平均累積エピソード報酬

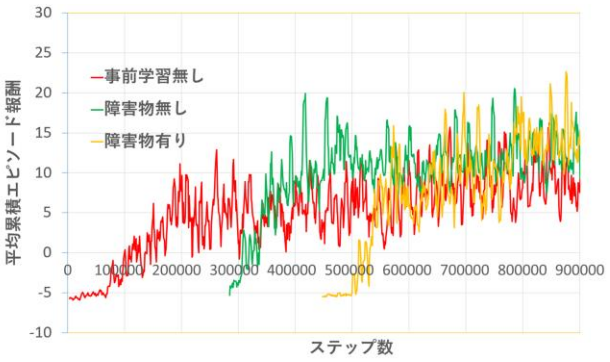

図 5： 高速道路合流部環境での平均累積エピソード報酬

5. 考察

まず事前学習なしでの学習では、初期において適応が遅く、報酬の向上も緩やかであることから、事前学習を行わない場合、エージェントが新しい環境に適応するための時間がかかり、学習効率が低下することを示唆している。次に障害物無しのサーキット環境での学習は、高速道路合流部環境へのスムーズな移行を助けたと考えられる。これにより、エージェントは基本的な走行パターンを学習し、高速道路合流環境での複雑な状況にも迅速に適応できたと考えられる。次に障害物有りのサーキット環境での学習は、特定の回避行動を学習することにより、他の環境への適応が遅れる可能性があることが示唆される。これは、動的な状況に対応するための柔軟性が不足するためであると考えられる。橙色のグラフが 50 万ステップから約 60 万ステップの間で低いまま安定していることは、サーキット環境での学習が過剰であり、オーバーフィッティング[10]が発生していることを示唆している。しかし最終的には最も高い報酬を達成することから、より複雑な環境での事前学習が、長期的な報酬の向上に寄与する可能性があることを示唆している。

6. むすび

本研究では、レースサーキット環境で障害物の有無を条件として学習したエージェントと事前学習なしのエージェントを高速道路合流シミュレーションで比較し、他車との衝突回避における平均累積エピソード報酬や損失関数を評価して、事前学習の効果を検証した。その結果、障害物無しのサーキット環境での事前学習は初期の適応を助け、高速道路環境での学習効率を向上させた。一方、障害物有りのサーキット環境での事前学習は初期適応が遅れたものの、最終的には高い報酬を達成した。これにより、事前学習がエージェントの学習効率と柔軟性を高める重要な要素であることが示唆された。G.Shi ら[11]は、人間の運転介入を PPO アルゴリズムに統合することで、自律運転システムのトレーニング効率と安全性が大幅に向上することを実証し、将来的な自律運転システムの開発において、人間と機械の協調が重要な役割を果たすことを示唆している。今後は、学習が停滞しているときに、人間の運転データを使用した模倣学習を導入することにより、学習のトレーニング効率や安全性の向上を目指していく必要がある。

参考文献

[1] 永平寺町役場. “自動運転「ZEN drive」” .2023, https://www.town.eiheiji.lg.jp/200/206/208/p010484.html.

[2] F. Shafique, T. Naeem, A. Hussain: “Path Tracking and Obstacle Avoidance Control Using Deep Reinforcement Learning for Autonomous Vehicles,” IEEE proceedings, 2023.

[3] P. Mihalcea: “Machine Learning Centralized Traffic Control System for City Intersections,” Undergraduate, Vol.18, pp.1-23, 2023.

[4] 警察庁. “交通事故統計情報のオープンデータ” ,2023, https://www.npa.go.jp/bureau/traffic/bunseki/info.html

[5] M.Bouton, A.Nakhaei, D.Isele, K.Fujimura, M.Kochefer: “Reinforcement Learning with Iterative Reasoning for Merging in Dense Traffic,” IEEE proceedings, 2020.

[6] S.Triest, A.Villaflor, J.M.Dolan: “Learning Highway Ramp Merging Via Reinforcement Learning with Temporally-Extended Actions,” IEEE, pp.1595-1600, 2020.

[7] S. Chen, J. Dong, P. Ha, Y. Li, S. Labi, "Graph Neural Network and Reinforcement Learning for Multi-agent Cooperative Control of Connected Autonomous Vehicles", Computer-Aided Civil and Infrastructure Engineering, Vol. 36, pp. 838-857, 2021.

[8] Y. Chen, W. E., X. Wang, C. Wang, "Multi-vehicle Cooperative Confluence Strategy in Freeway Merging Area under New Mixed Traffic Environment", 2022 5th International Conference on Intelligent Autonomous Systems (ICoIAS), pp. 358-363, 2022.

[9] 布留川英一.Unity ML-Agents 実践ゲームプログラミング.株式会社ボーンデジタル出版, 2022.

[10] A. J. M. Muzahid, S. F. Kamarulzaman, M. A. Rahman, "Comparison of PPO and SAC Algorithms Towards Decision Making Strategies for Collision Avoidance Among Multiple Autonomous Vehicles", ICSECS-ICOCSIM, Vol.4, pp.200-205, 2021.

[11] Gaosong Shi, Qinghai Zhao, Jirong Wang, Xin Dong, "Research on reinforcement learning based on PPO algorithm for human-machine intervention in autonomous driving", Electronic Research Archive, Vol. 32, No.4, pp. 2424-2446, 2024.