実世界強化学習に向けた Incremental Learning のソーシャルナビゲーションへの適応

○長久 陽斗（九州大学），松本 耕平（九州大学），兵藤 侑樹（九州大学），
富田 湧（九州大学），倉爪 亮（九州大学）

Adapting Incremental Learning Toward Real-World Reinforcement Learning for Social Navigation

○ Haruto Nagahisa (Kyushu Univ.), Kohei Matumoto (Kyushu Univ.), Yuki Hyodo (Kyushu Univ.),
Yuki Tomita (Kyushu Univ.), and Ryo Kurazume (Kyushu Univ.)

Abstract: In recent years, the demand for mobile robots has been rapidly increasing, and the application of deep reinforcement learning (DRL) has attracted significant attention, particularly as a navigation method in dynamic environments that include pedestrians. Conventional DRL approaches typically involve training only in a simulation environment before deployment in the real world. However, discrepancies between simulation and real-world environments can hinder the robot's ability to flexibly adapt to diverse situations. This study aims to develop a method that enables a robot to learn incrementally during real-world operation. A key challenge in this context lies in the constraints of computational resources, as autonomous navigation requires processing information from multiple sensors. To address this issue, this paper proposes a method based on incremental learning that directly utilizes sensor data without buffering, thereby improving the performance of real-world learning and navigation.

1. 緒言

近年，移動ロボットの需要が高まっており，特に歩行者を含む動的環境におけるナビゲーションは重要なタスクであり，深層強化学習を用いた手法が盛んに研究されている．従来の深層強化学習を用いた手法はシミュレーションで事前学習させたものを実世界で運用するといった形が多く，シミュレーション環境と実世界の環境との間には隔たりが存在することから，あらゆる場面で対応することは難しい．そこで，本研究ではロボットを実世界での動作中に逐次的に学習させることを目指す．ここで，課題となる点は計算資源の制約であり，ロボットが自律移動を行うためには様々なセンサを処理する必要があるため限られた計算資源のなかで学習を行う必要がある．そこで，バッファにデータを貯めずにセンサから取得したデータをそのまま学習に用いる Incremental Learning（以降では増分学習と呼ぶ）を用いることによりこの課題に取り組む．増分学習を行う強化学習手法として Action Value Gradient (AVG)[1] を採用する．

また，現実世界で強化学習を行うためには強化学習の探索的行動による安全対策やデータ収集コストが高いことが問題となる．他には，ナビゲーションの強化学習の報酬設定において，学習するエージェントが発見する解に対してバイアスが生じることを防ぐために疎な報酬設定で学習を行いたいが[2]，通常の強化学習手法単独では学習が困難であるという問題がある．そこで，残差強化学習[3] を用いることによってサンプル効率の向上をはかり，学習時の安全性，学習速度の向上と疎な報酬設定でも学習を可能にすることを目指す．

2. Action Value Gradient

AVG は，増分学習の不安定性に対して，正規化手法とスケーリング手法を組み合わせることで対処している深層強化学習手法である．動的環境における予期せぬ変化に対応するために強化学習エージェントにとって実世界で学習を行うことは重要であるがオンボードで利用可能な計算リソースと記憶領域には制限が存在している．以上のことから，リソース制約を満たす小規模なバッチ更新の採用が挙げられるが，深層ニューラルネットワークを活用した実世界学習を可能にする安定した増分学習の開発はまだ未解決な課題である．AVG は，この課題に取り組んでいる手法であり，SAC[4] と類似している手法である．SAC との相違点としては，SAC はオフポリシー型のバッチ処理手法であるのに対して，AVG は増分型オンポリシー手法である．SAC は行動をサンプリングしてリプレイバッファに格納するが AVG は異なり，同じ行動を再利用してアクターの勾配を逆伝播させることはない．さらに，AVG は SAC よりもシンプルな手法であり，安定性向上のためにダブル Q ラーニングやターゲット Q ネットワークを採用していない点が特徴である．他にも AVG は，観測の正規化，Penultimate 正規化，TD 誤差のスケーリング等を行っている．

3. 残差強化学習

残差強化学習とは，事前に作成した方策（以降ではベース方策と呼ぶ）を用いて解決できる部分問題と強化学習によって解決する残差問題に分解する手法である．最終的な方策はこれら 2 つの方策の組み合わせとして表現され以下の式で表される．

$$u = \pi_b(s) + \pi_\theta(s) \tag{1}$$

π_b はベース策で π_θ は強化学習によって作成した方策を表し，s は方策を出力するための状態入力を表す．今回の研究では，ベース方策に ORCA[5] に対して模倣学習させたものを用いる．

4. 提案手法のアルゴリズム

本稿では，残差強化学習と AVG を組み合わせた手法（以降では Residual reinforcement learning and AVG

（RAVG）と呼ぶ）を動的環境における移動ロボットのナビゲーションのタスクに適応する．Algorithm1 に RAVG のアルゴリズムを示す．まずは，残差強化学習におけるベース方策として CrowdNav[6][7] の環境において circle crossing のシナリオで歩行者環境は 5 人とし，ロボットと歩行者が ORCA に基づいて行動した際の結果を収集して，その結果を用いて模倣学習させたものを用い，強化学習による方策として AVG で学習させた方策を用いる．行動の出力範囲を $[-1, 1]$ としていて，ベース方策と AVG からの出力には tanh を用いて出力の範囲を $[-1, 1]$ としていることから，残差強化学習として行動を足す組み合わせを決める α の値は 0.5 とする．学習を行う部分については AVG の学習則のまま学習を行う．

Algorithm 1 Residual reinforcement learning and AVG

Require: Base policy π_b
  Initialize: θ, ϕ
  for $n = 0, ..., N-1$ episodes **do**
    Initialize s_0
    while s is not terminal **do**
      Calucate Action $u_t = \alpha\pi_b(s_t) + (1-\alpha)\pi_\theta(s_t)$
      Take action u_t ,observe s_{t+1},r_{t+1}
      Optimize θ, ϕ using AVG
    end while
  end for

5. シミュレーション実験

本稿で提案した RAVG の性能を検証するためにシミュレーションによる 2 つの実験を行った．1 つ目の実験では，増分学習であっても既存の強化学習手法と比較して遜色ない性能を出すことができるかについてとナビゲーションの疎な報酬設定の環境で学習を行うことができるのかについてを検証した．2 つ目の実験は，RAVG で学習を行うことによってベース方策よりも性能を向上させることが可能であるかについてを検証した．実験は CrowdNav のシミュレーション環境で行い，シナリオは circle crossing で問題設定はこちらの論文[8] と同じにしている．

5.1 シミュレーション実験 1

AVG と RAVG，残差強化学習と SAC，PPO[9]，TD3[10] をそれぞれ組み合わせて学習させたものについて比較を行う．残差強化学習を用いたものにはそれぞれのアルゴリズムの名前の前に R をつける．歩行者環境は 5 人で行い，100000 エピソードの学習を行う．その際に 100 エピソード学習を行うごとに 100 エピソードのテストを行い，比較を行う際にはテストを行った際の平均累積割引報酬 (CDR) が最も高いモデルを用いてエピソードごとに歩行者の初期位置を変えて 500 エピソードのテストを行い，ナビゲーションの成功率，衝突率，平均到達時間，CDR を比較する．その際の結果を表 1 に示す．表 1 の結果からナビゲーションの疎な報酬設定の環境においては AVG のみでは学習を行うことができないが残差強化学習を用いることにより学習を行うことが可能になることが判明した．疎な報酬設定の環境では，探索の過程で報酬を獲得することが難しいという点とナビゲーションの環境ではゴールに到達したとしても，歩行者の存在により連続してゴールに到達することは難しく学習により性能を向上させるためには多様な環境に適応する必要があることからサンプル効率がとても低く AVG のみでは学習により性能を向上させることが難しいと考える．しかし，残差強化学習を用いることにより行動のうち半分の要素がゴールに向かう行動を生成することから通常よりゴールに向かう行動が多くなり，それによりサンプル効率の向上につながり，学習を行うことが可能になるのではないかと考える．また，RAVG は成功率と CDR において RSAC と RPPO に比べて高い値を示し，平均到達時間に関しては比較手法のなかで最も良い値を示した．このことから，RAVG は増分学習であるにも関わらず，リプレイバッファありの既存の強化学習と比べて同等程度の性能を示し，指標によっては最も良い値を示すことが判明した．

Table 1 Results of the simulation experiment 1

Method	Success[%]	Collision[%]	Exec. time[s]	CDR
AVG	0.000	0.000	30.0	-0.004
RAVG	0.738	0.252	8.40	0.199
RSAC	0.724	0.244	13.4	0.112
RPPO	0.720	0.002	23.9	0.040
RTD3	0.960	0.018	14.3	0.218

5.2 シミュレーション実験 2

10000 エピソードの学習を行い，歩行者環境を 2，5，8，12 と変化させた際の RAVG とベース方策との性能を比較する．それ以外の実験の設定はシミュレーション実験 1 と同様に行う．実験の結果を表 2 に示す．表 2 の結果から，歩行者環境が 2 人のときの衝突率以外の指標において RAVG で学習させた方策の結果のほうが良い結果となっている．このことから，学習によってベース方策よりも性能が向上することから，学習による有効性を確認することができる．

6. 実機実験

シミュレーション実験にて増分学習で学習を行うことが可能であることが判明したため実機を用いて実世界で学習を行い性能が向上するかどうかを検証した．実世界における学習において性能向上が見込めるかということと実世界とシミュレーション環境における隔たりを確認するために，実世界においては実世界で学習させた RAVG，シミュレーション環境で学習させた RAVG，ベース方策を検証し，またシミュレーション環境において実世界で学習させた設定と同じ設定で学習させた RAVG とベース方策を比較した．学習を行うにあたって常に人が行き交う環境でロボットを動作させ学習させることが難しいことから，歩行者環境は CrowdNav のシミュレーション環境を用意してロボットは実世界で動作するというような形で実験を行った．その際のシステムの構成を図 1，実機の構成を図 2 に示す．実験に使用したロボットは，メカナムホイールを用いた全方向移動ロボットでホイールの回転数を用いたオドメトリと 3D-LiDAR からの点群情報と IMU 情報を用いて自己位置推定を行っている．算出した自己位置と CrowdNav からの歩行者情報をモデルに入力し，ロボットに対して速度指令値を送っている．その後，ロボットが次のステップまで行動を行った後に，自己

Table 2 Results of the simulation experiment 2

	Base policy				RAVG			
Number of Pedestrians	Success[%]	Collision[%]	Exec. time[s]	CDR	Success[%]	Collision[%]	Exec. time[s]	CDR
2	0.736	0.046	14.8	0.139	0.872	0.102	10.7	0.236
5	0.474	0.518	9.78	0.044	0.890	0.108	8.54	0.264
8	0.368	0.632	8.81	-0.010	0.700	0.288	8.54	0.171
12	0.164	0.836	8.63	-0.115	0.314	0.686	8.25	-0.029

位置と歩行者情報を更新し RAVG を用いて学習を行い，モデルの更新を行うという流れになっている．

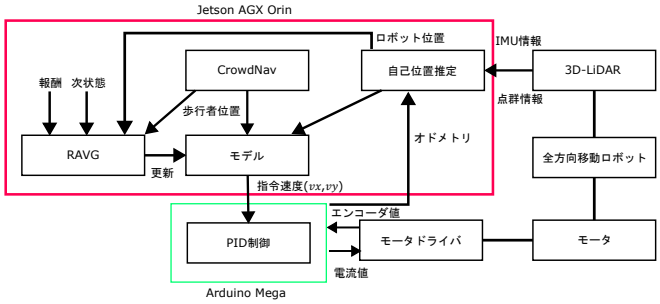

Fig. 1 Softwear configuration

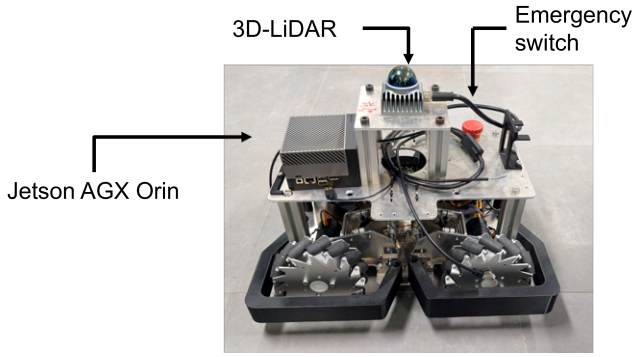

Fig. 2 Hardware configuration

実験の設定としては，歩行者環境は 5 人で行い，3000 エピソードの学習を行う．その際に 100 エピソード学習を行うごとに 10 エピソードのテストを行い，比較を行う際には CDR が最も高いモデルを用いて比較を行う．また，エピソードごとに歩行者の初期位置を変えて 100 エピソードのテストを行い，ナビゲーションの成功率，衝突率，平均到達時間，CDR を比較する．図 3 に実験の環境を示す．スタート位置からゴール位置までの距離は 4m とし環境の縦幅は 5m，横幅は 2.5m とする．ゴールに到達する，人に衝突する，50 秒が経過する，範囲外に進んだ場合は歩行者環境と時間をリセットして，ロボットはスタート位置に戻り再度学習を行う．以下に学習に用いた報酬設定を示す．d_t は時刻 t におけるロボットと周囲の歩行者間の最小距離を表し，p_t^r は時刻 t におけるロボットの位置，p_g はロボットの目標位置を表す．

$$R_t = \begin{cases} -0.25 & \text{if } d_t < 0 \\ (d_t - 0.2) * 0.125 & \text{else if } d_t < 0.2 \\ 1 & \text{else if } p_t^r = p_g \\ 0 & \text{timeout or out of range} \end{cases} \quad (2)$$

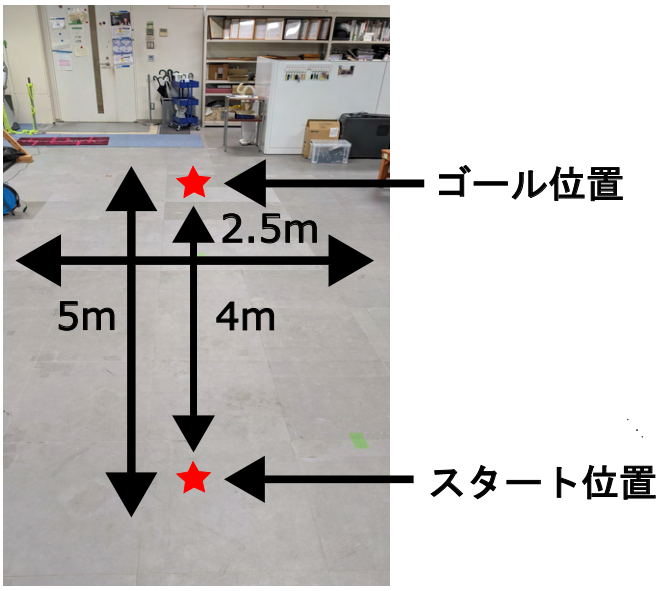

Fig. 3 Environment for the real-world experiment

実験の様子を図 4 に結果を表 3 に示す．実世界で学習させたものは Method のまえに R がつきシミュレーション環境で学習させたものは Method のまえに S がついている．表 3 の結果から実世界の環境とシミュレーション環境との隔たりがとても大きいことがわかる．シミュレーションで学習させたものは実世界で動作させた際にシミュレーションと比較して全ての指標において性能が下がっている．これは，シミュレーション環境においては速度司令値が与えられた際に次のステップにおいては理想的にその速度で動作した場合の位置に存在するが，実世界においては速度司令値が与えられたあとにそこから目標速度に対する制御を行うことからすぐにその速度に到達することができず，また路面の環境によっても摩擦等が違うことから場所によっても制御の追従速度は一定でない．また，動的環境におけるナビゲーションにおいては，素早い動作の切り替えが必要であることからも実世界とシミュレーション環境の隔たりが大きく影響したと考える．また，実世界で学習させた RAVG の結果はシミュレーションで検証した RAVG の結果には及ばないが実世界で動作

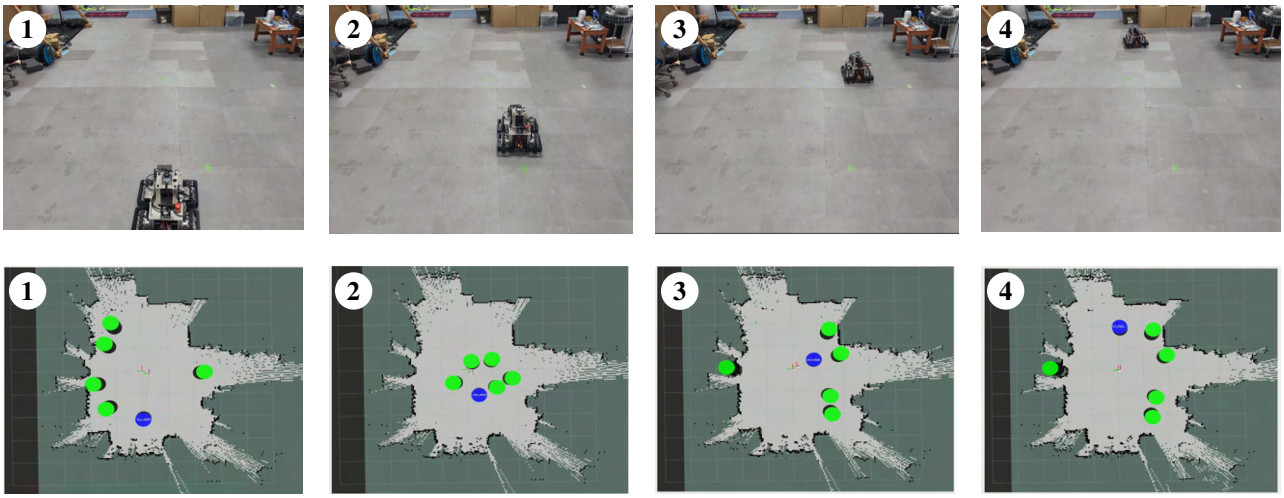

Fig. 4 Scenes of the real-world experiment

させたものの中では成功率と CDR の指標において最も良い結果を示した．以上のことから，実世界とシミュレーション環境との隔たりから実世界学習の必要性を確認することができ，今回の実験から RAVG を用いて実世界で学習を行うことが可能であることが確認できた．

Table 3 Results of the real-world experiment

environment	Method	Success[%]	Collision[%]	Exec. time[s]	CDR
real	R-RAVG	0.300	0.670	10.2	-0.054
real	S-RAVG	0.210	0.780	5.93	-0.073
real	S-Base policy	0.150	0.610	13.2	-0.113
simulation	S-RAVG	0.880	0.120	4.51	0.453
simulation	S-Base policy	0.680	0.320	11.8	0.097

7. 結言

本稿では増分学習を用いて強化学習を用いたナビゲーションが可能であるかを検証した．具体的には，ナビゲーションの疎な報酬設定での環境でも残差強化学習を用いることにより，学習を行うことができることを確認し，AVG を用いることによって増分学習であっても他のリプレイバッファを持った既存の強化学習と比べて遜色ない性能を出すことができることが判明した．また，RAVG を用いて学習を行うことにより，ベース方策よりも性能を向上させることができることから学習を行うことの有用性を確認することができた．そして，実機実験を行うことにより，実世界の環境でも増分学習を行うことができることが確認できた．今後の展望としては，環境の難しさによって性能向上に差が存在することから，拡散モデル等の表現力の高いモデルを採用することにより難しい環境でも性能が向上することができるように取り組みたいと考える．また，今回の実機実験では限られたエピソード数でしか学習を行っていないが理想的には常に学習を続けるロボット AI の実現を目指していることから，より長期間に渡って学習を行ったとしても学習が破綻することなく続けられることを検証したい．

参考文献

[1] G. Vasan, M. Elsayed, S. A. Azimi, J. He, F. Shahriar, C. Bellinger, M. White, and R. Mahmood: Deep Policy Gradient Methods Without Batch Updates, Target Networks, or Replay Buffers, Advances in Neural Information Processing Systems (NeurIPS), (2024), pp. 845–891.

[2] G. Vasan, Y. Wang, F. Shahriar, J. Bergstra, M. Jägersand, and A. R. Mahmood: Revisiting Sparse Rewards for Goal-Reaching Reinforcement Learning, Reinforcement Learning Journal (RLJ), 4, pp. 1841–1854 (2024).

[3] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine: Residual Reinforcement Learning for Robot Control, Proceedings of the International Conference on Robotics and Automation (ICRA) (2019), pp. 6023–6029.

[4] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine: Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, Proceedings of the International Conference on Machine Learning (ICML) (2018), pp. 1856–1865.

[5] J. van den Berg, S. J. Guy, M. C. Lin, and D. Manocha: Reciprocal *n*-Body Collision Avoidance, Proceedings of the International Symposium of Robotics Research(ISRR) (2009), pp. 3–19.

[6] C. Chen, Y. Liu, S. Kreiss, and A. Alahi: Crowd-Robot Interaction: Crowd-Aware Robot Navigation With Attention-Based Deep Reinforcement Learning, Proceedings of the International Conference on Robotics and Automation (ICRA) (2019), pp. 6015–6022.

[7] Y. Chen, C. Liu, B. E. Shi, and M. Liu: Robot Navigation in Crowds by Graph Convolutional Networks With Attention Learned From Human Gaze, IEEE Robotics and Automation Letters, 5.2, pp. 2754–2761 (2020).

[8] 兵頭 侑樹，松本 耕平，富田 湧，倉爪 亮: 動的環境における学習ベースおよびルールベースの切り替え手法を用いた移動ロボットナビゲーション, 第 42 回 日本ロボット学会学術講演会 (2024), 1E3–04.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov: Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[10] S. Fujimoto, H. van Hoof, and D. Meger: Addressing Function Approximation Error in Actor-Critic Methods, Proceedings of the International Conference on Machine Learning (ICML) (2018), pp. 1582–1591.