計測自動制御学会東北支部 60 周年記念学術講演会 (2024．12．17)
資料番号 S60.16

DDQN によるエレベーターの協調運動の一考察

A Study of Elevator Cooperative Behavior by DDQN

○高山大生*，岩井俊哉*，米澤直晃*

○ Daiki Takayama*, Toshiya Iwai*, Naoaki Yonezawa*

*日本大学

*Nihon University

キーワード：強化学習（Reinforcement Learning），DDQN（Double Deep Q Network），マルチエージェント(Multi Agent)，協調運動(Cooperative Behavior)，エレベーター制御（Elevator Control）

連絡先：963-8642 福島県郡山市田村町徳定字中河原 1 日本大学 工学部 情報工学科
知能情報処理研究室 岩井俊哉， Tel．：(024)956-8819，Fax．：(024)956-8863

E-mail: iwai. toshiya@nihon-u. ac. jp

1. 緒言

世の中には最短経路問題，配送最適化問題など多くの最適化問題が存在している．深層強化学習はビッグデータを用いた大規模な最適化問題を効率的に解く有力な方法である．エレベーター制御は，利用者が変化する中で待ち時間を少なく利用者を移動させるリアルタイムの最適化問題であり，既存研究 [1)2)] では，深層強化学習によりエレベーターの効率よい配車を学習できることが報告されている．また，深層強化学習の一手法である Double Deep Q Network(以下，DDQN と略記) [3)] では，強化学習における行動選択と学習を別のネットワークに分けて，それぞれの Q テーブルである Q_{target} と Q_{main} を一定間隔で同期させることで精度の良い学習を行うことができる．

本研究では DDQN を用いて，エレベーターの推奨される動きに基づく報酬と利用者の待ち時間に反比例した報酬を利用し，一基のエレベーターと複数基のエレベーターによる利用者の平均待ち時間と目的階への平均到達時間を測定し，複数基のエレベーターの協調制御による効率のよい配車が実現されるか確認する．

2. 研究内容

研究内容を強化学習の項目に分けて説明する．

2.1 エージェント

本研究では，１階から８階まで昇降できるエレベーターを考える．各階にはエレベーターを呼び出すための上階・下階を示すボタン（以下，外ボタン（上），外ボタン（下）と略記）がある．また，エレベーター内には目的階を示すボタン（以下，内ボタンと略記）がある．DDQN のエージェントはエレベーターである

【状態】一基のエージェントの状態は次の 11 種類の変数で決まる．

(1)現在いる階

(2)運動状態として、stop， up， down の３種類，

(3) 現在階，上階及び下階の外ボタン（上）と（下）の点灯状態

(4) 現在階，上階及び 下階の内ボタンの点灯状態

【行動】エージェントの選択できる行動は，運動状態と等しく，stop，up，down の 3 種類とする．

【初期状態】エージェントの初期状態は，現在階が 1 階で，運動状態が stop，外ボタン・内ボタン

は全て消灯とする.

2.2 環境

次の環境を用意した.

【環境】一定の確率でいずれかの階に利用者が現れ，目的階をランダムに選択し，外ボタンを押す.利用者がエレベーターに入ったとき，選択した目的階の内ボタンの階を押す.

2.3 報酬

２種類の報酬を設定する．１つはエレベーター単独の動きに基づく報酬であり，１つは複数のエレベーターの協調運動に関する報酬である．前者を単独行動報酬，後者を協調行動報酬と呼ぶ.

単独行動報酬は，推奨される動きに関する報酬から構成される．推奨される動きとは，次の 6 つのルールに則る動きである．(1)エレベーターは既に乗っている利用者を優先し，異なる方向を目的階とする利用者を混在させず，そのエレベーターには運動方向を目的階とする利用者が乗ることができる．(2)エレベーターは現在の運動方向の近い目的階から停止し利用者を下ろす．(3)エレベーター内の利用者がいなくなったとしても，エレベーターの運動方向に利用者が待っているならば運動方向を変えない．(4)エレベーター内の利用者がいなくなり，エレベーターの運動方向に待っている利用者がいないならば，エレベーターの運動方向と反対方向にいる利用者を乗せるように運動方向を変える．(5)エレベーター内外の利用者がいないならば，エレベーターは止まる．(6)エレベーターは，存在しない階へ移動しない（ここでは，0 階や 9 階）．各時間ステップの行動が，これらの推奨行動に矛盾しないとき正の報酬 1 を与え，矛盾するときに負の報酬 - 1 を与える.

協調行動報酬とは，利用者をより速く目的階に下ろすための報酬であり，次に定義する$RATE$に比例した正の報酬である.

$$RATE = \frac{Ste\ _{ideal}}{Step_{actual}} \qquad (1)$$

ここで，$step_{ideal}$はエレベーター外に利用者が現れてから他の利用者がいない状況で目的階までに要する最短時間である．また，$step_{actual}$は当該利用者が現れてから目的階で降りるまでに実際に要した時間である．$RATE$は利用者がエレベーターを降りたときに得る.

3. 数値実験の方法

単独行動報酬と協調行動報酬$RATE$を用いて，一基のエレベーター，二基のエレベーター及び三基のエレベーターについて DDQN により 100 エピソードの学習を行う． エピソードごとに，学習フェーズと検証フェーズを交互に行うこととし，乱数平均のために，同一の実験条件で 10 回の学習を行う.

学習フェーズでは利用者の出現確率をランダムとし，1000 ステップ数でエピソードを終了する．また，行動方策として ε -greedy 方策を用い， ε の値は 0.3 の一定値とする．学習率を 0.0.1，割引率を 0.0.1 として Q 学習を行う.

検証フェーズでは，$prob_{ap}$の値を 0.1，0.2 および 0.4 に設定し，エレベーターを降りた利用者の人数が目標人数num_{goal}に達するか，1 エピソードあたりの最大ステップ数$Step_{verification}$に達した場合に各エピソードを終了する．ここで，一時間ステップごとに利用者が一人出現する確率が小さいほど，利用者の出現する時間間隔が大きくなることを考慮して，$Step_{verification}$を次式で定義した.

$$Step_{verification} = 2 \times \frac{num_{goa}}{prob_{ap}} \qquad (2)$$

ここで，$prob_{ap}$は一時間ステップごとに利用者がいずれかのエレベーターのいずれかの階に一人出現する確率である．また，num_{goal}は目標人数であり 100 人とし，行動方策として greedy 方策を用いた.

学習の評価量として，次に示す〈STEPS〉，〈RATES〉を検証フェーズに計測した．〈STEPS〉は，エピソードが終了するステップ数を10回の学習で乱数平均した測定量である．〈RATES〉とは，式(1)の 10 回の学習で乱数平均した測定量である．〈STEPS〉,〈RATES〉のエピソード推移を測定し，〈STEPS〉,〈RATES〉の推移と$prob_{ap}$の関係性を考察する．また，各$prob_{ap}$の値で〈STEPS〉,〈RATES〉のエレベーターの台数への依存性を比較する.

DDQN におけるパラメータは Q_{main} を更新する間隔を 100 ステップ，Q_{target} に Q_{main} の重みを同期する間隔

を 5 エピソード，バッチサイズを 32，隠れ層サイズを 32，ミニバッチで繰り返す学習回数を 5 回，損失関数を平均二乗誤差とする．

4. 数値実験の結果

$prob_{ap} = 0.1,\ 0.4$ での〈STEPS〉を，それぞれ Fig. 1，2 に示す．図中に，エレベーターの台数が１，２，３台での〈STEPS〉を異なる色で表す．〈STEPS〉はエピソードの増加に伴い収束しており，学習が終了しているといえる．Fig. 1 より，台数が増えることによる〈STEPS〉の違いは明確でない．これは$prob_{ap}$ が 0 に近いとき，1 台のエレベーターだけで少ない待ち時間で利用者を運搬できるため，$prob_{ap}$が小さいと台数による相違が小さくなるからと考えられる．また，Fig. 2 より，台数が増えることで〈STEPS〉がより小さい値で収束しているようにみえるが，台数による〈STEPS〉の差が統計的に優位であるか検証が必要である．

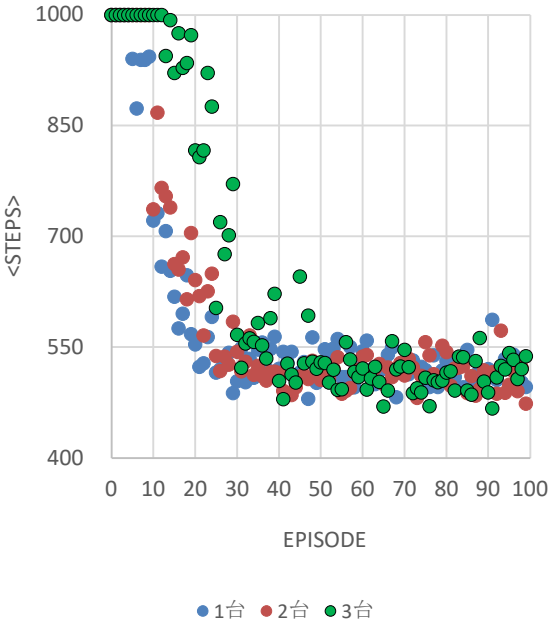

Fig. 1 〈STEPS〉($prob_{ap} = 0.1$)

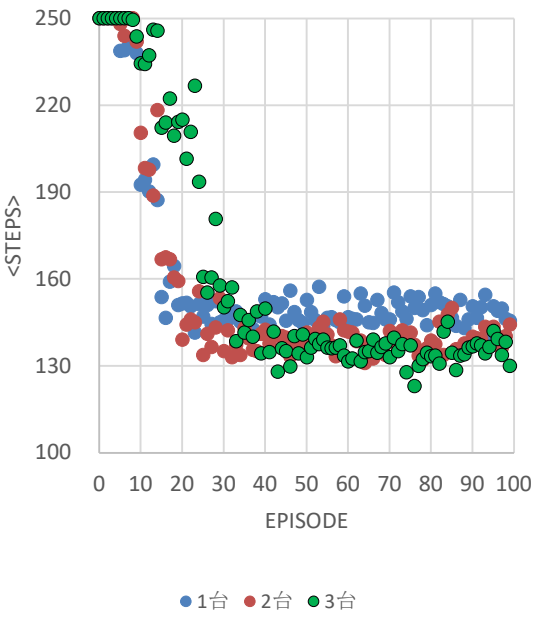

Fig. 2 〈STEPS〉($prob_{ap} = 0.4$)

$prob_{ap} = 0.1,\ 0.4$ での〈RATES〉を，それぞれ Fig. 3, 4 に示す．図中に，エレベーターの台数が１，２，３台での〈RATES〉を異なる色で表す．Fig. 3, 4 より，３種類のエレベーター台数の〈RATES〉は，エピソードの増加に伴い収束しており，学習が終了しているといえる．Fig. 3 より，台数が増えることによる〈RATES〉の違いは明確でない．これは$prob_{ap}$が 0 に近いとき，1 台のエレベーターだけで少ない待ち時間で利用者が運搬できるため，$prob_{ap}$が小さいと台数による相違が小さくなるからと考えらえる．Fig. 4 より，１台に比べて２，３台での〈RATES〉は大きな値で収束している．これは$prob_{ap}$が大きく混雑しているとき，１台より複数台での利用者の運搬が待ち時間を低減するためと考えられる．

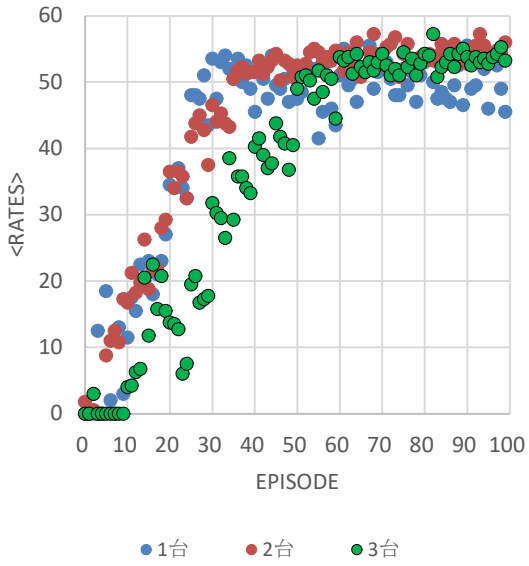

Fig. 3 〈RATES〉($prob_{ap} = 0.1$)

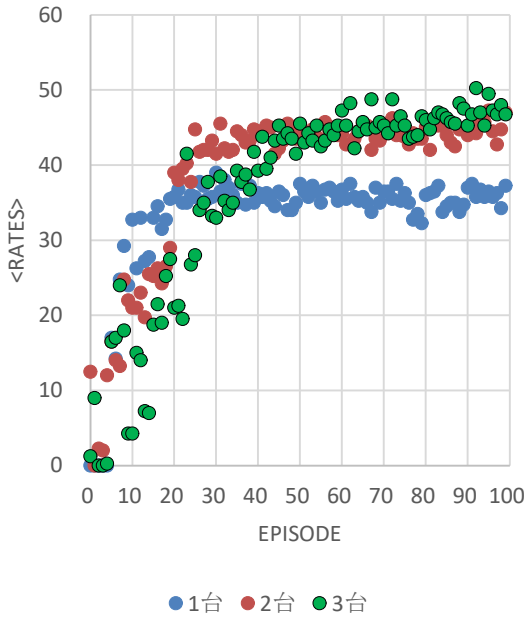

Fig. 4 〈RATES〉($prob_{ap} = 0.4$)

5. 考察

本研究では，利用者が混雑している状況では複数台のエレベーターを使うことで利用者の待ち時間を低減する運搬を学習できた．

6. 今後の課題

本研究では，協調行動を行ったことが明確に示されていないため，一時間ステップごとの各エレベーターの現在階を測定し，エレベータの行動を解析することで協調行動の有無を確認したい．

また，エレベーターの動作に必要な電力量など他の要素を報酬に加えて，混雑状況による適当なエレベーターの台数を検討する．

参考資料

1) Mateusz Wojtulewicz *et al.*, Application of Reinforcement Learning in Decision Systems: Lift Control Case Study, 14, 569(2024).

2) 吉田航他，深層 Q 学習によるエレベータ制御最適化のための Q-Network 構造の検討，第 34 回人工知能学会全国大会論文集(2020).

3) Hado van Hasselt *et al.*, Deep Reinforcement Learning with Double Q-learning, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 2094–2100, (20.15).

[3] 小高知宏，『強化学習と深層学習 —C 言語によるシミュレーション—』，オーム社，8—12 頁，20.17 年 10 月 20 日．