日本オペレーションズ・リサーチ学会
2021年 春季研究発表会

深層強化学習を用いた交通信号制御システムの実装と評価

(非会員)　中央大学　*阿保世聞　ABO Sebun
01309500　中央大学　高松瑞代　TAKAMATSU Mizuyo

1. はじめに

東京都では慢性的な交通渋滞の緩和に向けて，さまざまな取り組みを実施している．そのうちのひとつが需要予測信号制御の導入である．交差点に到着する車をセンサーで感知して交通量を予測し，リアルタイムに信号制御を行うことで，交通渋滞が発生しないように工夫されている．

深層強化学習は，試行錯誤的に効率的な行動選択を学習する強化学習に，深層学習の技術を組み合わせた手法である．深層強化学習は囲碁 AI や自動運転などで成果を挙げていることで有名である．近年は交通信号制御にも適用され，その有用性が示されている [3, 4].

本研究では，深層強化学習の手法のひとつである Deep Q-Network を用いた信号制御システムを実装し，現実的なシナリオに対して実験する．提案手法を固定型制御および現実の制御と比較し，その性能を評価する．

2. 深層強化学習

深層強化学習は，強化学習に深層学習の技術を組み合わせた手法である [1]．強化学習では，学習主体であるエージェントが，状態 s の観測，行動 a の選択，報酬 r の取得，次の状態 s' の観測という流れを繰り返す中で，受け取る報酬和の最大化を目標として学習を行う．深層強化学習では，強化学習における行動選択にディープニューラルネットワークを用いる．これにより，従来の強化学習では扱うことのできなかった大規模な状態空間をもつ問題にも適用可能となる．

深層強化学習の手法のひとつである Deep Q-Network (DQN) は，価値関数 $Q(s,a)$ に基づく手法である．$Q(s,a)$ は，時刻 t に状態 s で行動 a をとることで得られる報酬和 $r_t + r_{t+1} + r_{t+2} + \cdots$ の期待値を表す．ディープニューラルネットワークに状態 s を入力し，各行動について $Q(s,a)$ を推定する．経験サンプル $\langle s, a, r, s' \rangle$ をもとにニューラルネットワークを更新する．

3. 深層強化学習を用いた交通信号制御

本研究では，DQN を用いた交通信号制御システムを実装する．信号機の表示パターンのひとつをフェーズとよぶ．提案手法では，交差点に配置された信号機制御エージェントが，現在のフェーズと車線上の車両の待ち時間を入力として受け取り，次のフェーズの継続時間を行動として出力する．制御の評価には車両の待ち時間を用いる．

DQN エージェントはフェーズの終了時に，現在のフェーズと車線ごとの車両の待ち時間の総和を，以下の状態 s として受け取る：

$$s = [p_1, \cdots, p_n, w_1, \cdots, w_m].$$

ここで，m は対象とする交差点の車線数であり，n は信号機のフェーズ数である．p_i は，現在のフェーズが i 番目ならば 1，そうでないならば 0 をとる．w_i は i 番目の車線上に存在する車両の待ち時間の総和である．

状態 s を受け取ると，全結合ニューラルネットワークに入力して価値関数 $Q(s,a)$ を計算し，ε グリーディ法を用いて行動 a を選択する．報酬は，指定した継続時間終了後の全車両の待ち時間 w_i を用いて $r = -\sum_{i=1}^{m} w_i$ と定める．

学習を安定させるため，Double DQN (DDQN) と経験再生を用いる．DDQN では，メインネットワークとターゲットネットワークをもつ．それぞれの出力を $Q_{main}(s,a)$, $Q_{target}(s,a)$ とすると，学習で使用する教師データ y は

$$y = r + \gamma Q_{target}(s', \underset{a}{\mathrm{argmax}}(Q_{main}(s', a)))$$

と定義される．ただし γ は割引率である．ターゲットネットワークは数ステップごとにメインネットワークと同期し，それ以外では固定されている．経験再生では，得られた経験サンプル $\langle s, a, r, s' \rangle$ を保存し，教師データ y の計算時に保存したサンプルの中からランダムに選択することでサンプルの偏りを防ぐ．

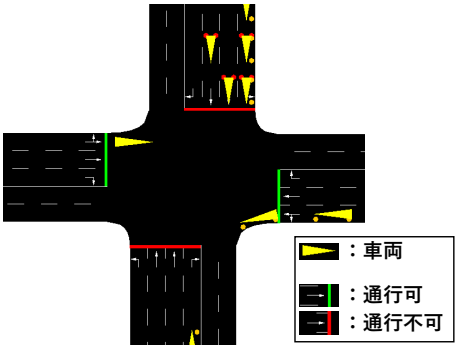

図 1: 対象とする交差点（14 車線）

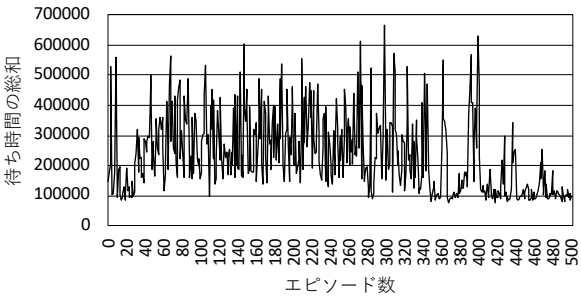

図 2: 車両待ち時間の総和の推移

4. 数値実験

本実験では，交通流シミュレータ Simulation of Urban Mobility (SUMO) [2] を使用する．対象とするのは，合計 14 車線からなる文京区の交差点である（図 1）．車両は 1 秒ごとに確率的に発生し，直進，左折，右折が可能である．フェーズは

- 南北方向からの全通行が可能
- 南北方向からの右折のみが可能
- 東西方向からの全通行が可能
- 東西方向の右折のみが可能

の 4 つである．エージェントの行動は 10, 20, . . . , 80 秒の中から選択する．3 つの中間層をもつネットワークを使用し，1 エピソードごとにターゲットネットワークを更新する．

3600 秒のシナリオを 1 エピソードとして，500 エピソードのシミュレーションを行った結果を示す．フェーズの継続時間を固定した制御（固定型）と，実際に使用されていた継続時間の制御（現行型）を DQN エージェントの比較対象とする．車両の発生確率と現行型制御の継続時間は，2020 年 12 月 29 日の 13 時から 14 時の記録をもとに設定した．

表 1: 平均待ち時間の比較

	固定型	現行型	提案手法
平均待ち時間	41.28 秒	34.30 秒	32.48 秒

図 2 に提案手法のエピソードごとの待ち時間の総和を示す．図 2 をみると，約 400 エピソードで学習の収束を確認できる．表 1 には固定型，現行型，提案手法の車両 1 台あたりの平均待ち時間をまとめている．固定型では全フェーズの継続時間が 30 秒の場合が最良であったため，これを採用している．提案手法に示す結果は，学習終了後のモデルで価値関数にグリーディな行動選択をしたものである．提案手法では平均待ち時間が固定型より約 9 秒短い．また，現行型と比較すると，待ち時間を約 1.8 秒短縮することに成功している．

今回の実験では，学習の収束後にも待ち時間が大きくなることがあった．これは，深層強化学習の一般的な問題として挙げられる不安定性が原因であると考えられる．優先度付き経験再生やバッチ正規化などの技法を取り入れることで問題点の改善につながると期待される．

参考文献

[1] 伊藤多一，今津義充，須藤広大，二ノ平将人，川崎悠介，酒井裕企，魏崇哲: 現場で使える！Python 深層強化学習入門—強化学習と深層学習による探索と制御, 翔泳社, 2019.

[2] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker: Recent Development and applications of SUMO—Simulation of urban mobility, *Int. J. Adv. Syst. and Meas.*, vol. 5, no. 3/4, pp. 128–138, 2012.

[3] X. Liang, X. Du, G. Wang and Z. Han: A deep reinforcement learning network for traffic light cycle control, *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1243–1253, 2019.

[4] E. van del Pol and F. A. Oliehoek: Coordinated deep reinforcement learners for traffic light control, *In NIPS'16 Workshop on Learning, Inference and Control of Multi-Agent Systems.*