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

An Experimental Study of Traffic Light Control Using Deep Reinforcement Learning

情報工学専攻 阿保 世聞

Information and System Engineering ABO Sebun

あらまし: 本研究では，深層強化学習の手法である Deep Q-Network を用いた交通信号制御システムを実装し，東京都文京区の交差点を対象とした現実的なシナリオで実験を行う．東京都では交通渋滞の緩和に向けて様々な取り組みを実施している．近年，交通信号制御に対する深層強化学習の有用性が示されている一方で，国内の先行研究では単純化された交差点環境での実験が多く，実際に採用されている現行型制御との比較も十分には行われていない．本研究では，単一交差点，複数交差点，歩行者を導入した交差点を対象として提案手法を固定型制御と現行型制御の 2 つと比較する．

キーワード: 深層強化学習，Deep Q-Network，交通信号制御

1 はじめに

東京都では交通渋滞の緩和に向けて様々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．需要予測信号制御は，車両感知器でリアルタイムに交通量を予測し制御に利用する．

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，人間をしのぐ強力な意思決定エージェントとしての可能性を示している．深層強化学習の応用のなかで，近年は交通信号制御への有用性が示されている [1, 3].

本研究では, 深層強化学習の代表的な手法である Deep Q-Network (DQN) を用いた交通信号制御システムを実装することで車両の待ち時間を短縮することを目指す．国内の先行研究では車両が直進のみを行うものや 1 車線のみで構成された交差点での実験など，単純化された交差点環境での実験が多い．また，実際に採用されている現行型制御との比較も十分には行われていない．そこで本研究では，東京都文京区の交差点で収集したデータに基づく現実的なシナリオを構築し，DQN を用いた制御を実験する．単一交差点，複数交差点，歩行者を導入した交差点での実験を行い，提案手法を固定型制御，現行型制御と比較する．

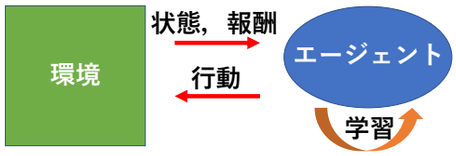

図 1 強化学習のイメージ図

2 強化学習と深層強化学習

2.1 強化学習

強化学習は，学習主体が試行錯誤的に行動を最適化する機械学習の手法である [2]．図 1 に強化学習のイメージ図を示す．強化学習の分野では，学習主体をエージェントとよぶ．エージェントは，環境の一部を状態として感知し，方策に基づいて行動する．環境はエージェントの行動に応じて変化する．それに伴い，エージェントは取った行動に対して報酬を受け取る．さらに，エージェントは変化した環境から次の状態を感知する．このようなエージェントと環境との相互作用の中で，エージェントは受け取る報酬の総和を最大化することを目的として行動を学習する．

強化学習エージェントは，状態空間 $\mathcal{S}$，行動空間 $\mathcal{A}$，報酬関数 $\mathcal{R}$，遷移関数 $\mathcal{P}$ から定義されるマルコフ決定過程の中で学習を行う．時刻 t に状態 s で行動 a をとることで受け取る報酬 r の総和の期待値を価値関数とよび，

$$Q(s,a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right]$$

と定義する．エージェントが自らの経験をもとに価値関数を推定する手法として Q 学習がある．Q 学習エージェントは次の更新式にしたがって，逐次的に価値関数を学習する．

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) \right]$$

ただし，α は学習率，γ は割引率である．

2.2 深層強化学習

深層強化学習は，強化学習エージェントの方策や価値関数をディープニューラルネットワークで近似する手法である．多くの強化学習手法では，状態や状態行動対ご

とに価値関数を保持，学習するため，状態・行動空間が大規模，複雑になると適用が困難になる．深層強化学習は，複雑な入出力関係を学習できる深層学習の特徴を引き継ぎ，従来の強化学習の問題点を解消する．

Deep Q-Network (DQN) は，ディープニューラルネットワークを用いて Q 学習の価値関数を近似する深層強化学習アルゴリズムである．Q 学習の更新式をもとにした教師データを用いてネットワークを学習することで，価値関数を推定する．単純な DQN は学習の収束が保証されておらず，この不安定性を改善するために経験再生，Freezing Target Network，Double Deep Q-Network (DDQN) の 3 手法を利用する．

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経験サンプルの中からランダムに選択する．これによりサンプル間の相関を弱めることで，学習が収束しやすくなる．

Freezing Target Network は，目標値計算に用いるネットワークパラメータを固定する手法である．目標値を一定期間固定することで，学習の振動や発散を抑えて学習の安定性を高める．

DDQN は，目標値計算の際に価値関数値を最大化する行動を見つけるネットワークと，その価値関数値を評価するネットワークを分離する手法である．これにより，ある行動の評価値が不当に高くなってしまうという過大評価を抑えることができる．

これらの手法を踏まえた Deep Q-Network アルゴリズムを以下に示す．

Algorithm 1 Deep Q-Network

ネットワーク，経験再生データベースを初期化
状態 s を初期化
エピソードの各ステップに対して繰り返し：
　行動 a を選択：
　　確率 ε でランダムな行動 a を選択
　　それ以外で $a = \underset{a}{\mathrm{argmax}}(Q_{main}(s, a))$
　報酬 r を受け取り次の状態 s' を観測
　経験サンプル $\langle s, a, r, s' \rangle$ を追加
　経験サンプルをランダムに選択
　$y = r + \gamma Q^{target}\left(s', \underset{a'}{\mathrm{argmax}}\, Q^{main}(s', a')\right)$
　教師データを y として θ^{main} を更新
　M ステップごとに更新：$\theta^{target} \leftarrow \theta^{main}$
　$s \leftarrow s'$

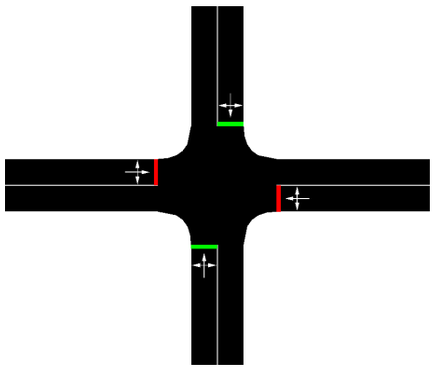

図 2　DQN の構築に使用する単純な交差点

表 1　状態表現の比較

状態表現	1 台あたりの平均待ち時間
待ち時間	12.4 秒
台数	14.6 秒
待ち時間と台数	12.2 秒

3　深層強化学習を用いた交通信号制御

3.1　深層強化学習モデルの設定

本研究では，エージェントを 1 つの交差点で交通信号制御を行うものとして定義する．信号機が表示するパターンのひとつをフェーズとよぶ．エージェントが操作するのはフェーズの継続時間である．

実験には交通流シミュレータ Simulation of Urban Mobility (SUMO) を使用する．SUMO では，定義した道路ネットワークと車両データを用いて交通シミュレーションが行える．

深層強化学習モデルを構築するために，図 2 に示す単純な交差点で，状態空間，行動空間，報酬関数について複数の選択肢を比較する．表 1 に状態表現の比較結果を示す．この結果から状態は現在のフェーズと各車線の車両の待ち時間と待ち台数とする．同様な実験を行った結果，行動は継続時間の集合 {10 秒 , 20 秒 , 30 秒 , . . . , 80 秒 } の中から一つを出力し，報酬は待ち時間の総和とする．

3.2　単一交差点への適用

対象とするのは，図 3 に示す東京都文京区に実在する交差点である．2020 年 12 月 29 日に現地で収集した車両の発生確率, 直進率, 左折率, 右折率のデータを用いて 3600 秒のシナリオを構築した．図 4 に交差点のフェーズを示す．DQN 型制御の比較対象として，フェーズの継続時間を固定した制御（固定型制御）と，実際に使用されていた継続時間の制御（現行型制御）を使用する．

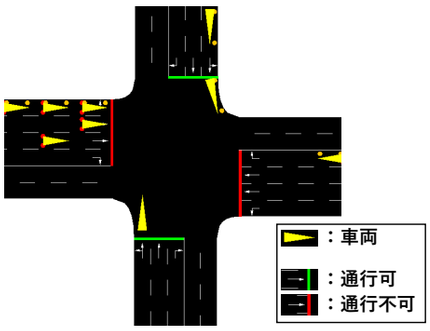

図 3　単一交差点

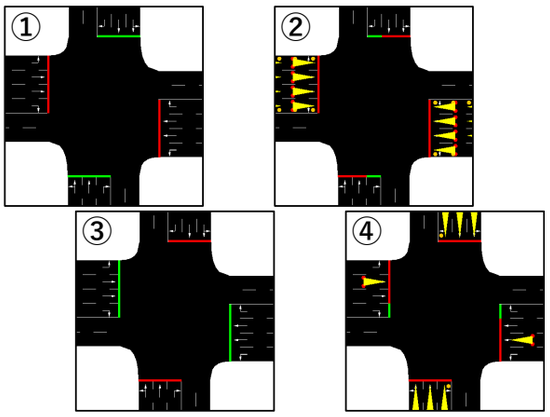

図 4　対象交差点のフェーズ

表 2　平均待ち時間の比較

	平均待ち時間
固定型	39.83 秒
現行型	34.30 秒
DQN 型	32.48 秒

表 2 には固定型制御，現行型制御，DQN 型制御の車両 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 20 秒の場合が最良であったため，これを採用している．DQN 型に示す結果は，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．DQN 型では平均待ち時間が固定型より約 7.4 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．

3.3　複数交差点への適用

図 5 に対象とする東京都文京区の 2 つの交差点を示す．エージェント間で協調させるため，状態伝達モデル，行動伝達モデル，報酬共有モデル，起点・従属モデルの 4 種類を実装した．状態伝達モデルと行動伝達モデルでは，それぞれ他のエージェントの状態，行動を自らの状態に追加する．報酬共有モデルは，他の交差点の車両の待ち時間も報酬に含める．起点・従属モデルは，起点エージェントの制御に合わせて従属エージェントの制御を決定する．具体的には，従属エージェントは起点エージェントのフェーズを受け取ることで，起点エージェントが配置された交差点から車両が流入するタイミングでその車線に青信号を表示するフェーズに切り替える．

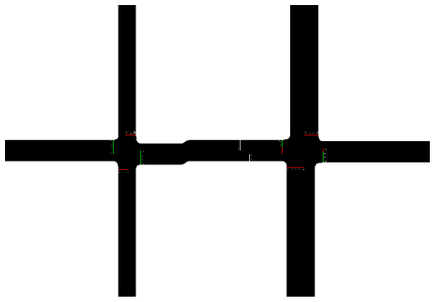

図 5　複数交差点

表 3　平均待ち時間の比較

	平均待ち時間
固定型	81.74 秒
現行型	48.17 秒
DQN 型（独立）	60.07 秒
DQN 型（状態伝達）	61.19 秒
DQN 型（行動伝達）	82.88 秒
DQN 型（報酬共有）	60.31 秒
DQN 型（起点・従属）	47.79 秒

表 3 にモデルごとの車両の平均待ち時間を示す．起点・従属モデルを利用した DQN 型制御が現行型制御をわずかに上回り，最良の結果となっている．他の 3 つのモデルはいずれも協調動作を獲得することができず，単一交差点で使用した独立 DQN 型制御と同等の性能にとどまっている．原因はいくつか考えられるが，特にエージェントが互いに変化し続けることが挙げられる．エージェントは他のエージェントに関する情報を受け取り学習を行う．しかし他のエージェントもまた学習を行うため，他エージェントに関する情報が時間とともに常に変化し，結果的に学習が安定しないという問題が起こりうる．一方で，起点・従属モデルのように明示的にエージェント間に協調を行わせることで，交差点間のスムーズな交通を実現する効率的な制御を学習することができたと言える．

4　歩行者を考慮した交通信号制御

図 6 に対象とする歩道付きの交差点を示す．単一交差点で示した交差点に，0.1 の確率で歩行者が発生する．歩行者の待ち時間を状態と報酬に追加した歩行者版 DQN 型制御を実験する．

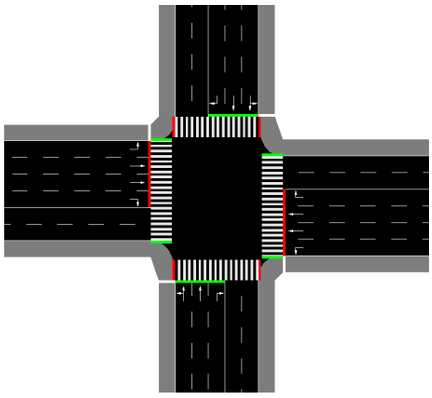

図 6　歩道付きの交差点

表 4　平均待ち時間の比較

	車両	歩行者
固定型（10 秒）	329.53 秒	17.19 秒
固定型（40 秒）	236.24 秒	52.63 秒
現行型	59.14 秒	29.78 秒
DQN 型（車両のみ版）	199.54 秒	77.50 秒
DQN 型（歩行者版）	64.97 秒	25.28 秒

表 4 に制御手法ごとの車両と歩行者の平均待ち時間を示す．車両，歩行者の両方で実用的な結果を示したのは現行型制御と歩行者版 DQN 型制御であった．歩行者版 DQN 型制御は現行型と比較すると，車両の待ち時間は約 5.8 秒大きくなっているが，歩行者の待ち時間は約 4.5 秒小さくなっている．車両のみを考慮した DQN 型制御と比較すると，車両，歩行者の両方で大きく上回る結果を示しており，歩行者を考慮することの重要性がうかがえる．

図 7 に，フェーズごとの選択された行動の割合を示す．車両の右折のみが許されるフェーズ 2, 4 では最短の継続時間を選択することがほとんどであり，通行できる車両が限られたフェーズの継続時間を最小限に抑えようとしていると考えられる．また，車両の直進，左折，右折，歩行者の通行が許されるフェーズ 1, 3 でも，現行型制御と比べると短い継続時間が選択されていた．歩行者は歩道上で一度に多数の通行が可能であるため，フェーズの継続時間を短くすることで歩行者の通行を促し，待ち時間を抑えることができたと考えられる．車両は停止状態から速度を出すためには加速の時間が必要になり，フェーズの継続時間が短いと多くの車両にこの加速のための時間がかかるため，車両の待ち時間については現行型制御を下回る結果になったと考えられる．

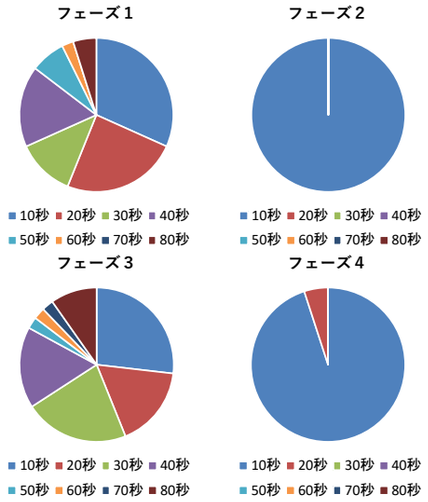

図 7　図 4 の各フェーズで選択された行動の割合

5　結論

本研究では，Deep Q-Network を用いた交通信号制御システムを実装し，東京都文京区の交差点を参考にした現実的なシナリオで実験した．単一交差点では，現行型制御よりも待ち時間を短縮した．複数交差点では，起点・従属モデルを利用した DQN 制御が，明示的に協調を定義することで現行型制御よりも待ち時間をわずかに短縮した．歩行者を含む交差点では，歩行者の待ち時間を状態と報酬に含めた結果，現行型制御よりも車両の待ち時間は大きくなったが，歩行者の待ち時間を短縮することに成功した．実験全体を通して，東京都の現行型制御が DQN を用いた交通信号制御に匹敵する性能をもつことがわかった．

今後の課題は，さらに実用的な交通信号制御を目指し，より大規模な環境で実験を行い，その環境に適したモデルに改良することである．

参考文献

[1] 神崎陽平, 佐藤季久恵, 高屋英知, 小川亮, 芦原佑太, 栗原聡: Deep Q-Network を用いたマルチエージェントによる交通信号制御システムの提案, 第 32 回全国大会人工知能学会, 2018.

[2] R.S. Sutton and A.G. Barto: 強化学習, 三上貞芳, 皆川雅章 (訳), 森北出版, 2000.

[3] 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.*