КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2024 Т. 16 № 5 С. 1253–1269
DOI: 10.20537/2076-7633-2024-16-5-1253-1269

Ки&М

МОДЕЛИ В ФИЗИКЕ И ТЕХНОЛОГИИ

УДК: 004.89

Адаптивное управление сигналами светофоров на основе обучения с подкреплением, инвариантное к конфигурации светофорного объекта

А. С. Юмаганов[a], А. А. Агафонов[b], В. В. Мясников[c]

Самарский национальный исследовательский университет им. академика С. П. Королёва,
Россия, 443086, г. Самара, Московское шоссе, д. 34

E-mail: [a] yumagan@gmail.com, [b] agafonov.aa@ssau.ru, [c] vmyas@geosamara.ru

Получено 15.04.2024, после доработки — 22.07.2024.
Принято к публикации 09.08.2024.

В работе представлен метод адаптивного управления сигналами светофоров, инвариантный к конфигурации светофорного объекта. Предложенный метод использует одну модель нейронной сети для управления светофорами различных конфигураций, отличающихся как по числу контролируемых полос движения, так и по используемому набору фаз. Для описания пространства состояний используется как динамическая информация о состоянии транспортного потока, так и статические данные о конфигурации контролируемого перекрестка. Для повышения скорости обучения модели предлагается использовать эксперта, предоставляющего дополнительные данные для обучения модели. В качестве эксперта используется метод адаптивного управления, основанный на максимизации взвешенного потока транспортных средств через перекресток. Экспериментальные исследования разработанного метода, проведенные в системе микроскопического моделирования движения транспортных средств, подтвердили его работоспособность и эффективность. Была показана возможность применения разработанного метода в сценарии моделирования, не используемом в процессе обучения. Представлено сравнение предложенного метода с другими известными решениями задачи управления светофорным объектом, в том числе с методом, используемым в качестве эксперта. В большинстве сценариев разработанный метод показал лучший результат по критериям среднего времени движения и среднего времени ожидания. Преимущество над методом, используемым в качестве эксперта, в зависимости от исследуемого сценария составило от 2 % до 12 % по критерию среднего времени ожидания транспортных средств и от 1 % до 7 % по критерию среднего времени движения.

Ключевые слова: управление сигналами светофоров, обучение с подкреплением, подключенные транспортные средства, имитационное моделирование

COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 5 P. 1253–1269
DOI: 10.20537/2076-7633-2024-16-5-1253-1269

Ки&М

MODELS IN PHYSICS AND TECHNOLOGY

UDC: 004.89

Reinforcement learning-based adaptive traffic signal control invariant to traffic signal configuration

A. S. Yumaganov[a], A. A. Agafonov[b], V. V. Myasnikov[c]

Samara National Research University,
34 Moskovskoye shosse, Samara, 443086, Russia

E-mail: [a] yumagan@gmail.com, [b] agafonov.aa@ssau.ru, [c] vmyas@geosamara.ru

Received 15.04.2024, after completion — 22.07.2024.
Accepted for publication 09.08.2024.

In this paper, we propose an adaptive traffic signal control method invariant to the configuration of the traffic signal. The proposed method uses one neural network model to control traffic signals of various configurations, differing both in the number of controlled lanes and in the used traffic light control cycle (set of phases). To describe the state space, both dynamic information about the current state of the traffic flow and static data about the configuration of a controlled intersection are used. To increase the speed of model training and reduce the required amount of data required for model convergence, it is proposed to use an "expert" who provides additional data for model training. As an expert, we propose to use an adaptive control method based on maximizing the weighted flow of vehicles through an intersection. Experimental studies of the effectiveness of the developed method were carried out in a microscopic simulation software package. The obtained results confirmed the effectiveness of the proposed method in different simulation scenarios. The possibility of using the developed method in a simulation scenario that is not used in the training process was shown. We provide a comparison of the proposed method with other baseline solutions, including the method used as an "expert". In most scenarios, the developed method showed the best results by average travel time and average waiting time criteria. The advantage over the method used as an expert, depending on the scenario under study, ranged from 2 % to 12 % according to the criterion of average vehicle waiting time and from 1 % to 7 % according to the criterion of average travel time.

Keywords: traffic signal control, reinforcement learning, connected vehicles, imitation modelling

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 5, pp. 1253–1269 (Russian).

1. Введение

В настоящее задача эффективного управления дорожно-транспортным комплексом продолжает оставаться актуальной проблемой. Ежегодное увеличение числа транспортных средств способствует появлению большого числа пробок (заторов) на дорогах. Эта проблема особенно актуальна для крупных городов. Согласно недавним исследованиям компании «Ингосстрах» [Долгий путь на работу. . . , 2023] лишь 17 % людей, добирающихся на работу на наземном транспорте, не стоят в пробках, при этом 31 % людей сталкиваются с дорожными заторами каждый день. В отчете [Traffic Scorecard, 2023] представлены показатели мобильности населения в наиболее загруженных районах мира, включая временные затраты на совершение транспортных корреспонденций и потребление топлива. Согласно отчету задержки при перемещении могут превышать 100 часов на одного водителя в год, и это число растет. Очевидно, наличие заторов на дорогах способствует существенному увеличению времени, затрачиваемому на движение транспортных средств по своим маршрутам, повышению расхода топлива и увеличению объемов выбросов вредоносных веществ в окружающую среду. Эти эффекты оказывают негативное влияние на стоимость и качество предоставления транспортных услуг. Одним из самых недорогих, доступных и эффективных способов управления транспортными потоками для решения проблемы пробок на дорогах является адаптивное управление светофорными объектами, осуществляющими контроль движения транспортных потоков на перекрестках.

Помимо стандартного способа управления сигналами светофорного объекта, при котором смена фазы светофора осуществляется согласно заранее установленному расписанию, выделяют адаптивные методы/алгоритмы управления сигналами. Система управления сигналами светофорного объекта называется адаптивной, если выбор фазы светофорного регулирования осуществляется на основе данных о движении транспортных средств в окрестностях регулируемого перекрестка. Классический детерминированный метод адаптивного управления сигналами светофоров, представленный в [Webster, 1958], используя информацию о величине транспортного потока, осуществляет контроль движения на перекрестке путем регулирования длительности всего светофорного цикла и, соответственно, длительности каждой отдельной фазы светофорного цикла. В [Varaiya, 2013] был представлен метод адаптивного управления, заключающийся в минимизации давления на перекрестке, то есть разницы между входящим и исходящим трафиком на перекрестке. В работе [Агафонов, Юмаганов, Мясников, 2022] авторы предложили метод управления, который осуществляет контроль движения транспортных потоков на перекрестке путем выбора следующей фазы светофорного объекта на основе прогнозируемой величины транспортного потока. Прогноз осуществляется по данным о положении и скорости движения транспортных средств на прилегающих к контролируемому перекрестку входящих полосах движения.

Увеличение числа подключенных транспортных средств, способных обмениваться информацией о параметрах движения с транспортной инфраструктурой, разработка автономных транспортных средств, развитие систем мониторинга движения транспорта, увеличение количества различного рода датчиков, детекторов в транспортных системах способствовало существенному увеличению объемов данных, которые могут быть использованы для эффективного управления движением транспортных средств, в том числе путем управления траекториями движения транспортных средств [Agafonov, Yumaganov, Myasnikov, 2019], организации движения беспилотных транспортных средств в форме кластеров [Быков, 2022], а также управления движением на перекрестках. В связи с этим в последнее время широкое развитие получили адаптивные методы управления сигналами светофоров на основе машинного обучения. Такие методы в основном используют обучение с подкреплением (reinforcement learning, RL) для решения задачи управления движением транспортных потоков на перекрестке.

Основные методы обучения с подкреплением можно классифицировать в три категории:

- методы, основанные на оптимизации функции полезности [Ault, Hanna, Sharon, 2020; Wang et al., 2022];
- методы, основанные на оптимизации стратегии (политики) [Lillicrap et al., 2019; Ren et al., 2022];
- методы класса «actor – critic», сочетающие в себе оба подхода: «actor» используется для выбора действия, «critic» — для оценки выбранного действия [Zhang et al., 2022; Mo et al., 2022].

Метод IDQN (Independent Deep Q-Network) [Ault, Hanna, Sharon, 2020] использует подход глубокого Q-обучения для обучения независимых агентов, каждый из которых регулирует движение на отдельном перекрестке. Для описания состояния каждого агента используются следующие факторы: порядковый номер текущей фазы светофорного объекта, общее количество транспортных средств на каждой из входящих полосах движения и сумма их скоростей, количество неподвижных транспортных средств на каждой из полос движения. В [Wang et al., 2022] метод глубокого Q-обучения использовался в сочетании с частично наблюдаемым пространством состояний от детекторов транспортного потока. В [Wei et al., 2022] предложено использовать графовые нейронные сети, что позволяет обмениваться информацией о дорожной ситуации между ближайшими контролируемыми перекрестками. Метод на основе оптимизации стратегии для решения задачи координированного управления на нескольких перекрестках предложен в [Ren et al., 2022]. Метод использует локальную и глобальную координацию между агентами для корректировки награды каждого агента. В работе [Mo et al., 2022] представлен метод адаптивного управления сигналами светофорного объекта, который учитывает наличие на дорогах неподключенных транспортных средств и использует модифицированный вариант метода обучения с подкреплением «actor – critic». В [Wu, Kim, Ma, 2022] авторы исследовали влияние различных способов описания пространства состояний и функций награды на эффективности RL-методов для решения задачи управления на изолированном перекрестке.

В [Guo, Li, Ban, 2019] авторами представлен обзор методов управления дорожным движением с использованием данных от подключенных и автономных транспортных средств (connected and autonomous vehicles, CAV). В первой части обзора были рассмотрены детерминированные и стохастические методы оценки состояния транспортного потока с использованием данных CAV. Вторая часть рассматривает задачи управления с учетом данных CAV и оцененного состояния транспортного потока. Были рассмотрены задачи навигации, управления сигналами светофоров, а также совместного управления транспортными средствами типа CAV и светофорными объектами. Одним из будущих направлений исследований, выделенных авторами, является реализация разработанных методов и алгоритмов для решения задач управления в реальных условиях.

В [Li et al., 2023] авторы выделили перспективные направления исследований в области управления смешанным потоком транспортных средств, включающим подключенные и управляемые водителями транспортные средства, а также рассмотрели современные методы управления светофорами путем синхронизации сигналов светофоров с траекториями движения CAV, методы построения траекторий движения и методы совместного управления светофорами и траекториями движения от уровня перекрестка до уровня всей дорожной сети.

В обзоре [Han, Wang, Leclercq, 2023] авторы представили описание современных стратегий управления сигналами светофоров, сделав упор на выделение проблем, связанных с реализацией методов управления в реальных сценариях дорожного движения и переносом моделей,

обученных в системах моделирования, в реальность. В частности, авторы делают вывод, что использование дополнительной информации о распространении транспортных потоков, например из соответствующих моделей прогнозирования транспортного потока [Прокопцев, Алексеенко, Холодов, 2018], снижает затраты на обучение RL-методов. Кроме того, авторы подчеркивают, что используемые стратегии переноса обученной модели для работы в реальных условиях являются достаточно простыми, требуется как разработка более сложных методов решения задач, так и разработка комплексных методологий сравнительного анализа, позволяющих эффективно оценивать стратегии управления.

Проведенный обзор подтвердил актуальность решения проблемы управления сигналами светофорных объектов, в частности задачи обучения моделей для их использования в реальных условиях и задачи переноса обучения моделей. Хотя в большинстве случаев RL-методы управления превосходят детерминированный адаптивные методы, модели агентов, обученные этими методами, имеют существенный недостаток. При добавлении нового светофорного объекта в транспортную сеть соответствующего ему агента необходимо предварительно обучить. При этом уже обученные ранее агенты использовать в большинстве случаев не представляется возможным, так как конфигурация добавляемого светофорного объекта может отличаться от конфигурации уже обученных агентов. Представленный в данной работе RL-метод адаптивного управления сигналами светофорного объекта лишен указанного недостатка. В частности, в работе предлагается

- формировать описание пространства состояний RL-агента, инвариантное к конфигурации светофорного объекта, что позволяет использовать больше данных для обучения модели и осуществлять перенос обучения (весов модели) при добавлении нового светофорного объекта в сеть;
- использовать эксперта для предобучения модели агента и в процессе RL-обучения [Hester et al., 2018], что позволяет сократить время обучения. В качестве эксперта используется необучаемый метод адаптивного управления, основанный на нахождении оценки максимального взвешенного потока транспортных средств [Агафонов, Юмаганов, Мясников, 2022].

Работа построена следующим образов. В § 2 представлены основные понятия обучения с подкреплением и описана постановка задачи. В § 3 предложен RL-метод адаптивного управления, описаны используемый метод, способ формирования пространства состояний, пространства действия и награды. В § 4 представлены экспериментальные исследования предложенного метода в реальных сценариях моделирования. В завершении работы представлены выводы и возможные направления дальнейших исследований.

2. Постановка задачи

Задача управления сигналами светофоров на перекрестке с помощью методов обучения с подкреплением (RL-методов) обычно представляется в виде марковского процесса принятия решений, который может быть описан кортежем [Wei et al., 2020]

$$\langle S, A, P, R, \gamma \rangle,$$

где S — это пространство состояний окружающей среды, A — пространство действий, P — вероятностная функция перехода между состояниями, R — функция награды, γ — коэффициент дисконтирования. Рассмотрим подробнее определения элементов указанного выше кортежа.

В момент времени t агент наблюдает состояние окружающей среды $s_t \in S$ (пространство состояний) и совершает действие $s_t \in S$. В результате выполнения этого действия состояние

среды изменяется в соответствии с функцией $P(a_t, s_t; s_{t+1}) = \Pr(s_{t+1} \mid s_t, a_t)$. $R(s_t, a_t; s_{t+1})$ — мгновенная награда, получаемая агентом при переходе из состояния s_t в состояние s_{t+1} в результате выполнения действия a_t. Цель агента — найти такую стратегию (политику) $\pi^* : S \to A$, которая максимизирует ожидаемую итоговую награду, определяемую следующим образом:

$$G_t = \sum_{i=0}^{T} \gamma^i r_{t+i},$$

где T — количество временных шагов, $\gamma \in [0, 1]$ — коэффициент дисконтирования, который управляет важностью будущих наград относительно мгновенной награды.

Один из способов решения проблемы поиска оптимальной стратегии π^* основан на нахождении оптимальной функции полезности $Q^*(s, a)$ (Q-функции). Функция полезности $Q^\pi : S \times A \to \mathbb{R}$ определяется следующим образом:

$$Q^\pi(s, a) = \mathrm{E}[G_t \mid s_t = s, \ a_t = a].$$

Для нахождения оптимальной функции полезности $Q^*(s, a)$ используется уравнение Беллмана:

$$Q^*(s_t, a_t) = \mathrm{E}\left[R(s_t, a_t; s_{t+1}) + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})\right]. \quad (1)$$

Тогда оптимальная стратегия π^* определяется следующим образом:

$$\pi^*(s) = \arg\max_a Q^*(s, a).$$

Таким образом, для решения задачи адаптивного управления сигналами светофорного объекта на перекрестке на основе обучения с подкреплением необходимо определить пространство состояний окружающей среды, пространство действий агента, задать функцию награды и выбрать используемый RL-метод обучения. В следующем разделе последовательно рассмотрим способ формирования описания пространства состояний среды (подпараграф 3.1), используемую функцию награды и пространство действий (подпараграф 3.2), а также используемый метод обучения с подкреплением (подпараграф 3.3).

3. Метод адаптивного управления сигналами светофоров

Для описания предложенного метода управления определим пространство состояний S, пространство действий A, функцию награды R и используемый метод обучения с подкреплением.

3.1. Пространство состояний

В работе предлагается использовать описание пространства состояний окружающей среды (управляемого перекрестка), инвариантное к конфигурации светофорного объекта. Такой подход позволит использовать одну обученную модель агента для управления всеми перекрестками в транспортной сети, использовать больший объем данных / меньшее число эпизодов для обучения модели, осуществлять перенос обучения для настройки управления новыми светофорными объектами.

Для описания пространства состояний в работе предлагается использовать вектор признаков, характеризующий каждую пару полос движения на перекрестке: входящую полосу и соответствующую ей исходящую полосу движения. Причем одной входящей полосе может соответствовать несколько исходящих полос движения, и наоборот. Вектор признаков для одной такой пары полос движения можно условно разбить на две части: динамическую и статическую.

Динамическая часть вектора признаков характеризует наблюдаемое состояние транспортного потока на соответствующей паре полос движения на перекрестке, при этом учитывается состояние транспортного потока по всей длине входящих/исходящих полос движения. В качестве характеристик наблюдаемого состояния транспортного потока на полосе движения *lane* предлагается использовать следующие значения:

- индикатор возможности движения через перекресток по рассматриваемой полосе движения (только входящей) при активной текущей фазе φ светофорного объекта:

$$I_{lane}(\varphi) = \begin{cases} 1, \text{ если проезд разрешен,} \\ 0, \text{ если иначе;} \end{cases}$$

- количество движущихся транспортных средства на рассматриваемой полосе движения MV_{lane};
- количество стоящих транспортных средств на полосе движения HV_{lane};
- суммарное время ожидания транспортными средствами на рассматриваемой полосе движения WT_{lane};
- средняя скорость транспортных средств на рассматриваемой полосе движения AS_{lane}.

Таким образом, динамическая часть вектора признаков для одной пары входящей (*in*) и исходящей (*out*) полос движения на перекрестке имеет следующий вид:

$$(I_{in}(\varphi),\ MV_{in},\ HV_{in},\ WT_{in},\ AS_{in},\ MV_{out},\ HV_{out}, WT_{out},\ AS_{out}).$$

Для того чтобы модель была инвариантной к типу контролируемого светофорного объекта, то есть могла использоваться на перекрестках различной структуры со светофорными объектами, имеющими различные наборы фаз, помимо описанной выше динамической части вектора признаков, предлагается использовать статические данные о конфигурации светофорного объекта. Для каждой пары «входящий/исходящий потоки движения» (*in* и *out*) формируется вектор (статическая часть вектора признаков), размерность которого равна максимальному количеству основных фаз в светофорном объекте. Под основной фазой понимаются те фазы светофорного объекта, которые не содержат переходных (желтых) сигналов. Для фазы с порядковым номером n_φ значение соответствующего элемента статической части вектора признаков определяется следующим образом:

$$S_{in}^{out}(n_\varphi) = \begin{cases} 0, \text{ если фаза отсутствует;} \\ 1, \text{ если движение с полосы } in \text{ на полосу } out \text{ запрещено;} \\ 2, \text{ если иначе.} \end{cases}$$

При этом порядковый номер n_φ существующей фазы светофорного объекта всегда меньше порядкового номера несуществующей фазы. Используемый способ статического описания пар полос движения, при котором возможно три значения $S_{in}^{out}(n_\varphi)$, позволяет явным образом обозначить количество основных фаз светофорного объекта.

В настоящей работе максимальное количество основных фаз в светофорном объекте полагается равным 5. Объединив соответствующие динамическую и статическую части вектора признаков, получим итоговый вектор признаков для пары входящей (*in*) и исходящей (*out*) полос движения:

$$(I_{in}(\varphi),\ MV_{in},\ HV_{in},\ WT_{in}, AS_{in},\ MV_{out},\ HV_{out},\ WT_{out},\ AS_{out},$$
$$S_{in}^{out}(0),\ S_{in}^{out}(1),\ S_{in}^{out}(2),\ S_{in}^{out}(3),\ S_{in}^{out}(4)).$$

Хотя динамическая и статическая части вектора признаков содержат информацию различной природы, нейронные сети позволяют обнаруживать скрытые закономерности в данных такого вида, в том числе и при решении задач управления светофорным объектом на перекрестке с помощью RL-методов [Wu, Kim, Ma, 2022].

Максимальное количество пар в данной работе полагается равным 20; в случае если таких пар меньше, итоговый вектор признаков дополняется нулями. Таким образом, итоговая размерность вектора признаков, описывающего пространство состояний, равна 280. Стоит отметить, что итоговый вектор признаков формируется путем объединения векторов признаков описанных выше пар полос движения в определенном порядке: начиная с северной входящей полосы и далее по часовой стрелке. Если одна входящая полоса движения связана с несколькими исходящими полосами, то сначала добавляется пара полос, при движении по которой транспортное средство осуществляет поворот направо, затем осуществляет движение прямо и, наконец, осуществляет поворот налево.

3.2. Пространство действий и функция награды

Пространство действий агента, управляющего светофорным объектом, состоит из индексов основных фаз светофорного объекта. Размер пространства действий равен максимальному количеству основных фаз в контролируемых светофорных объектах. Как было указано выше, в настоящей работе максимальное количество основных фаз в светофорном объекте полагается равным 5. Однако светофорные объекты могут иметь меньшее количество основных фаз, чем установленное максимальное число. В результате чего может возникнуть ситуация, при которой RL-метод возвращает порядковый номер фазы, превосходящий максимальное число фаз для рассматриваемого светофорного объекта. Для решения этой проблемы предлагается выбирать индекс устанавливаемой фазы светофора, который соответствует фазе, имеющей максимальное Q-значение среди фаз, доступных рассматриваемому светофорному объекту.

Награду R предлагается рассчитывать на основе вычисления давления перекрестка, которое определяется как разница между количеством транспортных средств, въезжающих на перекресток, и количеством транспортных средств, выезжающих с перекрестка по соответствующим входящим и исходящим полосам движения [Varaiya, 2013]. Таким образом, в качестве функции награды предлагается использовать следующую функцию:

$$R(s_t,\, a_t;\, s_{t+1}) = -P_t^{t+1},$$

где P_t^{t+1} — давление перекрестка за временной интервал $(t,\, t+1)$.

3.3. Метод обучения с подкреплением

Для решения задачи управления сигналами светофорного объекта в качестве метода обучения с подкреплением в работе используется подход двойного Q-обучения (Double DQN) [Van Hasselt, Guez, Silver, 2016], основанный на оптимизации функции полезности. Для решения задачи управления также предлагается использовать представленный в [Hester et al., 2018] подход, который использует опыт эксперта в процессе обучения агента управления (метод Deep Q-Learning from Demonstrations, DQfD). Данный метод был представлен с целью существенного ускорения обучения агента при наличии данных от учителя (эксперта). Часто возникают ситуации, когда отсутствует достаточно точный симулятор для решения какой-либо задачи в реальном мире (управление автомобилем, БПЛА). Тогда агент должен обучаться в реальной среде, в которой его неправильные действия могут иметь серьезные нежелательные последствия. Поэтому в таких ситуация требуется, чтобы агент начинал обучение, имея достаточно хорошую начальную

эффективность решения поставленной задачи. Для этого авторы метода предложили использовать данные об управлении объектом другой системой управления или человеком (экспертные данные) при обучении агента.

В результате процесс обучения агента методом DQfD включает в себя два этапа. На первом этапе осуществляется предварительное имитационное обучение модели. В качестве данных для обучения на данном этапе используются только данные эксперта (учителя). В настоящей работе в качестве экспертных данных предлагается использовать данные, полученные с помощью адаптивного метода управления дорожными сигналами на основе максимального взвешенного потока [Агафонов, Юмаганов, Мясников, 2022] (метод MaxPWFlow).

Функция потерь, используемая в процессе обучения, имеет следующий вид:

$$J(Q) = J_{DQ}(Q) + \lambda_1 J_n(Q) + \lambda_2 J_E(Q) + \lambda_3 J_{L2}(Q), \tag{2}$$

где $J_{DQ}(Q)$ — функция потерь для метода Double DQN, $J_n(Q)$ — функция потерь Double DQN на шаге n, $J_E(Q)$ — функция потерь обучения с учителем, $J_{L2}(Q)$ — стандартная функция потерь $L2$ регуляризации, λ_1, λ_2, λ_3 — числовые коэффициенты.

Функция потерь $J_E(Q)$ используется только для данных, полученных от эксперта, в противном случае (при использовании данных, полученных непосредственно с помощью обучаемого агента) соответствующий коэффициент λ_2 равен нулю. Функция потерь обучения с учителем $J_E(Q)$ имеет следующий вид:

$$J_E(Q) = \max_{a \in A}[Q(s, a, \theta) + l(a_E, a)] - Q(s, a_E, \theta),$$

где a_E — действие, выбранное экспертом в состоянии s; $l(a_E, a)$ — граничная функция, которая определяется следующим образом:

$$l(a_E, a) = \begin{cases} 0, & \text{если } a_E = a, \\ m > 0, & \text{если иначе.} \end{cases}$$

Функция потерь $J_E(Q)$ построена так, чтобы Q-значения для действий, не выбранных экспертом, были всегда ниже на величину не меньше m, чем Q-значения действий эксперта в том же состоянии. Таким образом, добавление этой функции потерь приводит Q-значения неисследованных действий к таким значениям, при которых, используя жадную политику, агент может имитировать действия эксперта. Однако в том случае, если в качестве функции потерь на первом этапе обучения будет использоваться только функция $J_E(Q)$, возвращаемые обученной нейронной сетью Q-значения не будут удовлетворять уравнению Беллмана, на основе которого построено Q-обучение (1). Поэтому на всех этапах обучения используются также функции потерь Q-обучения — $J_{DQ}(Q)$ и $J_n(Q)$.

Метод Q-обучения Double DQN в процессе обучения использует две нейронных сети для аппроксимации значений функции полезности — основную (online network) и целевую (target network). При этом целевая сеть имеет ту же архитектуру, что и основная сеть, однако ее весовые коэффициенты формируются путем копирования соответствующих весовых коэффициентов основной сети каждые τ шагов обучения. Целевая нейронная сети используется для оценки выбранного основной сетью действия. Применение двух отдельных нейронных сетей позволяет решить проблему переоценки действий, свойственных базовому нейросетевому методу Q-обучения DQN. Функция потерь для метода Double DQN имеет следующий вид:

$$J_{DQ}(Q) = \left(r + \gamma Q\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a, \theta), \theta_{tr}\right) - Q(s_t, a, \theta)\right)^2,$$

где θ и θ_{tr} — веса аппроксимирующих Q-функцию нейронных сетей: основной и целевой соответственно. Тогда функция потерь Double Q-learning на шаге n задается следующим образом:

$$J_n(Q) = \left(\sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n Q\left(s_{t+n}, \arg\max_a Q(s_{t+n}, a, \theta), \theta_{tr} \right) - Q(s_t, a, \theta) \right)^2 .$$

Используемая в настоящей работе архитектура нейронной сети, аппроксимирующей значения функции полезности, представлена на рис. 1.

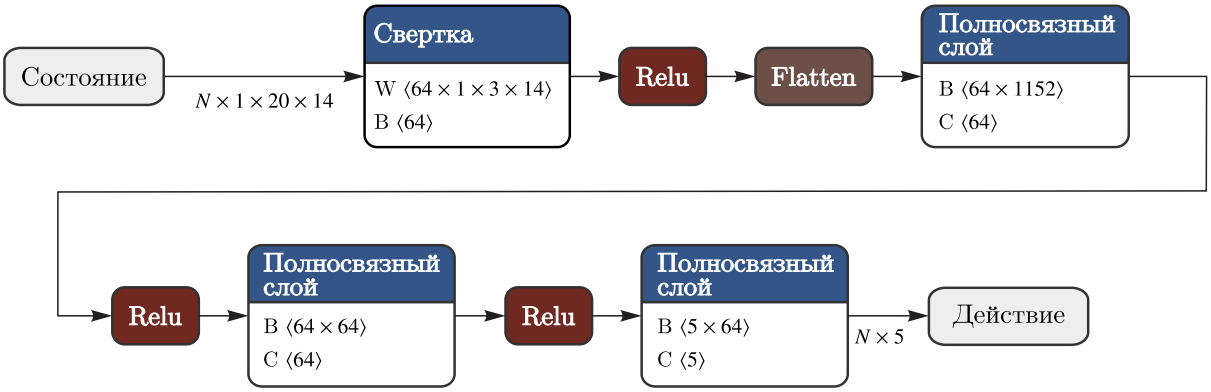

Рис. 1. Архитектура нейронной сети

После предварительного имитационного обучения модели на данных от эксперта на втором этапе осуществляется онлайн-обучение модели на основе метода обучения с подкреплением Double DQN. В качестве функции потерь используется рассмотренная выше функция $J(Q)$ с коэффициентом $\lambda_2 = 0$. В процессе обучения агент, взаимодействуя с окружением, добавляет данные для обучения в хранилище — буфер воспроизведения. Буфер воспроизведения имеет ограниченный размер, и при его полном заполнении старые данные перезаписываются. При этом перезаписываются только данные обучаемого агента, а данные, полученные экспертом, остаются в буфере воспроизведения. В начале второго этапа буфер содержит только экспертные данные. В работе используется буфер воспроизведения с приоритетным воспроизведением опыта [Schaul et al., 2016], в котором вероятность выбора данных для обучения прямо пропорциональна ошибке обучения модели на этих данных.

В рамках разработанного метода каждый агент независимо управляет работой соответствующего светофорного объекта, используя единую (общую) модель агента для управления каждым перекрестком. Поэтому второй этап процесса обучения агента немного отличается от исходного варианта метода DQfD, в котором полагается, что взаимодействие с одним объектом осуществляет один уникальный агент. Псевдокод второго этапа обучения представлен ниже (алгоритм 1). В алгоритме D^{replay} — буфер воспроизведения, TLs — множество светофорных объектов.

В работе используются следующие рекомендуемые авторами [Hester et al., 2018] значения рассмотренных параметров: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10^{-5}$, $m = 0{,}8$, $\tau = 10^4$.

4. Экспериментальные исследования

Для оценки эффективности и работоспособности предложенного метода был проведен ряд экспериментальных исследований в системе микроскопического моделирования транспортных потоков SUMO (Simulation of Urban MObility) [Lopez et al., 2018]. Сравнение разработанного метода проводилось со следующими методами:

Алгоритм 1. Модифицированная версия второго этапа обучения агента методом DQfD

Входные данные: D^{replay}, θ, θ_{tr}, τ, TLs
for $t \in \{1, 2, \ldots\}$ **do**
for $tl \in TLs$ **do** # Цикл по светофорам
Выбор действия для каждого tl на основе текущей функции полезности
$a_{tl} = \arg\max_a(Q(s_{tl}, a, \theta))$
end for
# Одновременное выполнение выбранных действий на каждом светофорном объекте;
for $tl \in TLs$ **do** # Цикл по светофорам
Получение нового состояния среды s'_{tl} и награды r_{tl};
Сохранение $(s_{tl}, a_{tl}, r_{tl}, s'_{tl})$ в D^{replay};
Получение выборки из D^{replay};
Обучение текущей нейронной сети на полученной выборке с использованием (2);
end for
end for
if $t \bmod \tau == 0$ **then**
$\theta_{tr} = \theta$
end if

- используемым в качестве эксперта адаптивным методом управления сигналами светофоров MaxPWFlow [Агафонов, Юмаганов, Мясников, 2022];
- методом независимого глубокого Q-обучения IDQN (Independent Deep Q-Network) [Ault, Hanna, Sharon, 2020], в котором один агент управляет одним светофорным объектом;
- классическим методом управления сигналами светофорного объекта FixedTime, при котором фазы светофора сменяются согласно заранее установленному расписанию независимо от текущей дорожной ситуации.

В качестве критериев оценки эффективности сравниваемых методов управления сигналами светофоров были выбраны следующие: среднее время ожидания транспортных средств и среднее время движения по маршруту (в секундах).

Экспериментальные исследования проводились с шагом симуляции, равным 1 секунде, при общем времени симуляции 3600 секунд. Сравнение всех моделей проводилось на одинаковых данных — выборке из десяти эпизодов моделирования. При этом эпизоды отличались между собой значением параметра seed системы SUMO, который влияет на начальные положения транспортных средств на сегментах транспортной сети и динамику их движения. Экспериментальные исследования проводились на четырех сценариях имитационного моделирования, основанных на сценарии TAPAS Cologne [TAPASCologne. . . , 2024]:

- сценарий cologne1 (сценарий содержит один регулируемый перекресток; общее количество транспортных средств в сценарии — 2014);
- сценарий cologne3 (транспортная сеть представляет собой короткий участок одного из шоссе транспортной сети города Кёльна и содержит 3 регулируемых светофорами перекрестка различной структуры; общее количество транспортных средств в сценарии — 2856);
- сценарий cologne8 (транспортная сеть представляет собой участок транспортной сети города Кёльна и содержит 8 регулируемых светофорами перекрестков с различной структурой; общее количество транспортных средств в сценарии — 1740);

- сценарий cologne8_high (транспортная сеть идентична транспортной сети сценария cologne8, но увеличена транспортная нагрузка на сеть; общее количество транспортных средств в сценарии — 4214).

На рис. 2 представлены транспортные сети, используемые в указанных выше сценариях.

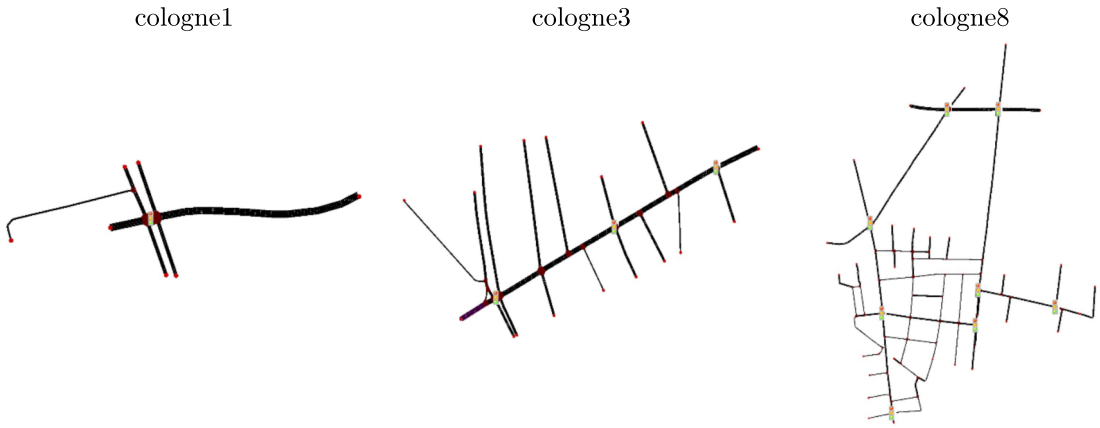

Рис. 2. Транспортные сети используемых сценариев моделирования движения

На первом этапе экспериментальных исследований оценивалась сходимость методов обучения с подкреплением: предложенного метода (метод-эксперт — MaxPWFlow) и метода IDQN. На рис. 3 представлены кривые обучения методов на сценарии cologne3. Как видно из полученных результатов, методы были успешно обучены и достигли стабильных значений среднего времени движения. При этом предложенному методу потребовалось значительно меньше эпизодов для обучения (примерно в 10 раз). Стоит отметить, что на рис. 3 представлены результаты оригинальной реализации метода IDQN, которая не использует опыт эксперта.

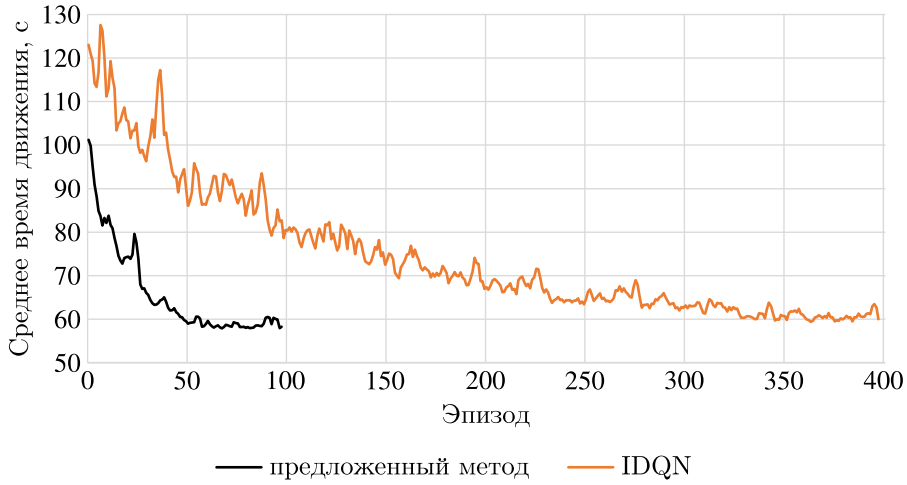

Рис. 3. График обучения RL-методов в сценарии «cologne3»

На втором этапе экспериментальных исследований было проведено сравнение методов в трех сценариях моделирования. Результаты представлены в таблице 1.

Как показывают результаты, предложенный метод в большинстве случаев позволяет улучшить эффективность метода-эксперта (MaxPWFlow) по рассматриваемым критериям, однако уступает методу IDQN на сценарии cologne1. Худший результат в каждом из рассмотрен-

Таблица 1. Оценка эффективности методов управления в сценариях моделирования по критериям среднего времени ожидания и среднего времени движения (в секундах)

Метод	cologne1		cologne3		cologne8	
	Среднее время ожидания	Среднее время движения	Среднее время ожидания	Среднее время движения	Среднее время ожидания	Среднее время движения
FixedTime	30,36	67,5	25,06	76,48	27,98	117,27
MaxPWFlow	9,57	47,15	9,24	59,57	4,25	89,90
IDQN	9,42	45,24	8,79	59,02	4,97	90,11
Предложенный метод	9,37	46,1	8,8	58,71	4,58	89,74

ных сценариев показал классический (неадаптивный) метод управления сигналами светофоров FixedTime.

На следующем этапе экспериментального анализа оценивалась эффективность обученных моделей при изменении объема трафика (примерно в 3 раза) в сценарии. Модели, обученные на сценарии cologne8, применялись для решения задачи управления в сценарии cologne8_high, то есть для тестирования использовались данные, отсутствующие в обучающей выборке. Результаты представлены в таблице 2.

Это объясняется тем, что в методе IDQN обучается отдельная модель для каждого светофорного объекта, в отличие от универсальной модели, инвариантной к структуре светофорного объекта, используемой в предложенном методе.

Таблица 2. Оценка эффективности методов управления в сценарии с повышенным объемом трафика

Метод	cologne8_high	
	Среднее время ожидания	Среднее время движения
FixedTime	61,55	152,17
MaxPWFlow	15,01	102,15
IDQN	13,15	92,66
Предложенный метод	13,97	95,32

Из-за увеличения объема трафика в сценарии увеличились средние показатели времени движения и времени ожидания. Однако используемые подходы являются достаточно эффективными для решения задачи управления и показывают лучшие результаты по сравнению с методами адаптивного управления MaxPWFlow.

Следующий этап экспериментальных исследований посвящен исследованию возможности применения модели на новой транспортной сети, не участвующей в процессе обучения, отличающейся структурой транспортных потоков. Так как конфигурация светофорных объектов в сценарии cologne3 уникальна относительно светофоров сценария cologne8, используемая ранее модель нейронной сети представленного метода была дополнительно обучена на этом сценарии, при этом в процессе обучения использовались данные экспертного метода MaxPWFlow. Модели агентов метода IDQN также были обучены на сценарии cologne3. В качестве нового сценария используется сценарий cologne1, транспортная сеть которого содержит светофорный объект, имеющий идентичную конфигурацию с одним из светофорных объектов транспортной сети сценария cologne3. Такая постановка экспериментов позволяет сравнить методы на новом сценарии без их обучения, что особенно важно для метода IDQN. Результаты экспериментальных исследований представлены в таблице 3.

Как видно из полученных результатов, предложенный метод позволяет повысить эффективность решения задачи управления по сравнению с базовыми методами по всем критериям

Таблица 3. Оценка эффективности методов управления на сценариях с различной транспортной сетью (структурой транспортных потоков)

Метод	cologne1		cologne3	
	Среднее время ожидания	Среднее время движения	Среднее время ожидания	Среднее время движения
FixedTime	30,36	67,5	25,06	76,48
MaxPWFlow	9,57	47,15	9,24	59,57
IDQN	33,1	70,84	8,79	59,02
Предложенный метод	11,51	49,56	9,02	58,38

оценки эффективности на сценарии cologne3. На новом сценарии cologne1 предложенный метод ожидаемо уступает методу MaxPWFlow примерно на 19 % по критерию среднего времени ожидания и на 1 % по критерию среднего времени движения, однако полученный результат существенно превосходит результат метода FixedTime. Метод IDQN на новом сценарии показал худший результат, уступив даже классическому методу управления сигналами светофоров. Таким образом, предложенный метод может быть использован на новом сценарии без дополнительного обучения на нем (в отличии от метода IDQN).

На заключительном этапе экспериментальных исследований был проведен анализ влияния процесса дополнительного обучения модели предложенного метода на результаты ее работы на разных сценариях. Используемая на втором этапе экспериментальных исследований модель предложенного метода была дополнительно обучена на сценарии cologne1 (экспертные данные получены методом MaxPWFlow). Полученная модель была протестирована на каждом из рассмотренных ранее сценариев. Модели метода IDQN были обучены на соответствующем исследуемом сценарии. Полученные результаты по критерию среднего времени ожидания представлены в таблице 4, по критерию среднего времени движения — в таблице 5.

Таблица 4. Оценка эффективности предложенного метода после дополнительного обучения по критерию среднего времени ожидания (в секундах)

Метод	cologne1	cologne3	cologne8	cologne8_high
FixedTime	30,36	25,06	27,98	61,55
MaxPWFlow	9,57	9,24	4,25	15,01
IDQN	9,42	8,79	4,97	13,13
Предложенный метод	9,48	8,35	4,77	13,87

Таблица 5. Оценка эффективности предложенного метода после дополнительного обучения по критерию среднего времени движения (в секундах)

Метод	cologne1	cologne3	cologne8	cologne8_high
FixedTime	67,50	76,48	117,27	152,17
MaxPWFlow	47,15	59,57	89,90	102,15
IDQN	45,24	59,02	90,11	93,09
Предложенный метод	46,43	58,39	90,1	94,74

Как видно из полученных результатов, процесс дополнительного обучения не оказал отрицательного влияния на эффективность разработанного метода: полученные ранее результаты были незначительно улучшены. Таким образом, модель нейронной сети предложенного метода, последовательно обученная на сценариях cologne8_high, cologne3 и cologne1, в трех случаях из четырех показала лучший результат среди рассмотренных методов по критерию среднего вре-

мени движения и в двух случаях — по критерию среднего времени ожидания, при этом только в одном сценарии по одному критерию уступил экспертному методу. Кроме того, используемое описание пространства состояний, инвариантное к конфигурации светофорного объекта, позволяет использовать предложенный метод в различных сценариях движения и объемах трафика, в том числе для управления новыми светофорными объектами без необходимости обучения модели с нуля.

5. Заключение

В работе представлен метод адаптивного управления сигналами светофоров на основе метода двойного Q-обучения с подкреплением, инвариантный к конфигурации светофорного объекта. Для решения задачи управления светофорными объектами различной структуры (как относительно конфигурации перекрестка, так и относительно рассматриваемого цикла фаз) предлагается формировать описание пространства состояний агента управления из двух частей: динамической, описывающей движение транспортных средств по каждой полосе движения на перекрестке, и статической, описывающей конфигурацию светофорного объекта. Предложенный способ формирования векторного представления состояния окружающей среды позволяет использовать одну модель агента для управления светофорными объектами различного вида в транспортной сети.

Для повышения скорости обучения модели и сокращения требуемого объема данных для сходимости модели (сокращения эпизодов моделирования и/или времени сбора данных в реальных условиях) предлагается использовать эксперта, предоставляющего дополнительные данные для обучения модели. В качестве эксперта используется необучаемый метод адаптивного управления, основанный на нахождении оценки максимального взвешенного потока транспортных средств. Использование предложенного подхода позволяет выполнять перенос обученной модели из среды моделирования в реальную дорожную сеть и сократить время, необходимое для подстройки модели к новым условиям среды.

Результаты проведенных экспериментальных исследований показали эффективность и работоспособность разработанного метода. Важным преимуществом разработанного метода перед другими RL-методами, включая используемый для сравнения метод IDQN, является применение одной модели нейронной сети всеми агентами, контролирующими светофорные объекты. Как показали результаты экспериментальных исследований, такой подход позволяет, обучив один раз модель нейронной сети на множестве регулируемых перекрестков различной конфигурации, достаточно эффективно применять ее в новой среде без дополнительного обучения.

Дальнейшие исследования могут быть направлены на исследование эффективности разработанного метода на крупной городской транспортной сети, в состав которой входит большое количество светофорных объектов разного вида. Кроме того, в настоящей работе рассмотрен только один метод управления сигналами светофоров, используемый в качестве эксперта. Соответственно, другим направлением дальнейших исследований является сравнение эффективности применения различных методов адаптивного управления сигналами светофоров в качестве экспертов.

Список литературы (References)

Агафонов А. А., Юмаганов А. С., Мясников В. В. Адаптивное управление дорожными сигналами на основе нейросетевого прогноза максимального взвешенного потока // Автометрия. — 2022. — Т. 58, № 5. — С. 85–97.

Agafonov A. A., Yumaganov A. S., Myasnikov V. V. Adaptive traffic signal control based on neural network prediction of weighted traffic flow // Optoelectronics, Instrumentation and Data Processing. — 2022. — Vol. 58, No. 5. — P. 503–513. (Original Russian paper: *Agafonov A. A., Yumaganov A. S., Myasnikov V. V.* Adaptivnoe upravlenie

dorozhnymi signalami na osnove nejrosetevogo prognoza maksimal'nogo vzveshennogo potoka // Avtometriya. — 2022. — Vol. 58, No. 5. — P. 85–97.)

Быков Н. В. Моделирование кластерного движения беспилотных транспортных средств в гетерогенном транспортном потоке // Компьютерные исследования и моделирование. — 2022. — Т. 14, № 5. — С. 1041–1058.

Bykov N. V. Modelirovanie klasternogo dvizheniya bespilotnyh transportnyh sredstv v geterogennom transportnom potoke [A simulation model of connected automated vehicles platoon dynamics in a heterogeneous traffic flow] // Computer Research and Modeling. — 2022. — Vol. 14, No. 5. — P. 1041–1058 (in Russian).

Долгий путь на работу: «Ингосстрах» исследовал, сколько времени люди тратят на дорогу и как коротают время в пробках. — [Электронный ресурс]. — https://www.ingos.ru/company/news/2023/14641835-ed3d-4edf-4690-08dbd07f61fc (дата обращения: 19.01.2024).

Dolgij put' na rabotu: «Ingosstrah» issledoval, skol'ko vremeni lyudi tratyat na dorogu i kak korotayut vremya v'probkah [Long way to work: Ingosstrakh studied how much time people spend on the road and how they pass the time in traffic jams]. — [Electronic resource]. — https://www.ingos.ru/company/news/2023/14641835-ed3d-4edf-4690-08dbd07f61fc (accessed: 19.01.2024; in Russian).

Прокопцев Н. Г., Алексеенко А. Е., Холодов Я. А. Использование сверточных нейронных сетей для прогнозирования скоростей транспортного потока на дорожном графе // Компьютерные исследования и моделирование. — 2018. — Т. 10, № 3. — С. 359–367.

Prokoptsev N. G., Alekseenko A. E., Kholodov Ya. A. Ispol'zovanie svertochnyh nejronnyh setej dlya prognozirovaniya skorostej transportnogo potoka na dorozhnom grafe [Traffic flow speed prediction on transportation graph with convolutional neural networks] // Computer Research and Modeling. — 2018. — Vol. 10, No. 3. — P. 359–367 (in Russian).

Agafonov A., Yumaganov A., Myasnikov V. Cooperative control for signalized intersections in intelligent connected vehicle environments // Mathematics. — 2023. — Vol. 11, No. 6. — P. 1540.

Ault J., Hanna J. P., Sharon G. Learning an interpretable traffic signal control policy. — 2013. — http://arxiv.org/abs/1912.11023

Guo Q., Li L., Ban X. Urban traffic signal control with connected and automated vehicles: A survey // Transportation Research Part C: Emerging Technologies. — 2019. — Vol. 101. — P. 313–334.

Han Y., Wang M., Leclercq L. Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation // Communications in Transportation Research. — 2023. — Vol. 3. — P. 100104.

Hester T., Vecerik M., Pietquin O., Lanctot M., Schaul T., Piot B., Horgan D., Quan J., Sendonaris A., Osband I., Dulac-Arnold G., Agapiou J., Leibo J., Gruslys A. Deep Q-learning from demonstrations // Proceedings of the AAAI Conference on Artificial Intelligence. — 2018. — Vol. 32, No. 1. — P. 3223–3230.

Li J., Yu C., Shen Z., Su Z., Ma W. A survey on urban traffic control under mixed traffic environment with connected automated vehicles // Transportation Research Part C: Emerging Technologies. — 2023. — Vol. 154. — P. 104258.

Lillicrap T. P., Hunt J. J., Pritzel A., Heess N., Erez T., Tassa Y., Silver D., Wierstra D. Continuous control with deep reinforcement learning. — 2019. — http://arxiv.org/abs/1509.02971

Lopez P. A., Wiessner E., Behrisch M., Bieker-Walz L., Erdmann J., Flotterod Y.-P., Hilbrich R., Lucken L., Rummel J., Wagner P. Microscopic traffic simulation using SUMO // 2018 21st International Conference on Intelligent Transportation Systems (ITSC). — 2018. — P. 2575–2582.

Mo Z., Li W., Fu Y., Ruan K., Di X. CVLight: decentralized learning for adaptive traffic signal control with connected vehicles // Transportation Research Part C: Emerging Technologies. — 2022. — Vol. 141. — P. 103728.

Ren F., Dong W., Zhao X., Zhang F., Kong Y., Yang Q. Two-layer coordinated reinforcement learning for traffic signal control in traffic network // Expert Systems with Applications. — 2024. — Vol. 235. — P. 121111.

Schaul T., Quan J., Antonoglou I., Silver D. Prioritized experience replay. — 2016. — http://arxiv.org/abs/1511.05952

TAPASCologne — SUMO Documentation. — [Electronic resource]. — https://sumo.dlr.de/docs/Data/Scenarios/TAPASCologne.html (accessed: 19.01.2024).

Traffic Scorecard – 2023 — INRIX. — [Electronic resource]. — https://inrix.com/traffic-scorecard-2023/ (accessed: 19.01.2024).

Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with Double Q-learning // Proceedings of the AAAI Conference on Artificial Intelligence. — 2016. — Vol. 30, No. 1. — P. 2094–2100.

Varaiya P. The max-pressure controller for arbitrary networks of signalized intersections // Advances in Dynamic Network Modeling in Complex Transportation Systems. — 2013. — P. 27–66.

Wang H., Yuan Y., Yang H. T., Zhao T., Liu Y. Traffic signal settings // Journal of Intelligent Transportation Systems. — 2023. — Vol. 27, No. 3. — P. 314–334.

Webster F. V. Traffic signal settings. — H. M. Stationery Office, 1958. — 56 p.

Wei H., Xu N., Zhang H., Zheng G., Zang X., Chen C., Zhang W., Zhu Y., Xu K., Li Z. CoLight: learning network-level cooperation for traffic signal control // Proceedings of the 28th ACM International Conference on Information and Knowledge Management. — 2023. — Vol. 2019. — P. 1913–1922.

Wei H., Zheng G., Gayah V., Li Z. A survey on traffic signal control methods. — 2020. — http://arxiv.org/abs/1904.08117

Wu C., Kim I., Ma Z. Deep reinforcement learning based traffic signal control: A comparative analysis // Procedia Computer Science. — 2023. — Vol. 220. — P. 275–282.

Zhang Y., Zhou Y., Lu H., Fujita H. Cooperative multi-agent actor–critic control of traffic network flow based on edge computing // Future Generation Computer Systems. — 2021. — Vol. 123. — P. 128–141.