Sharif Civil Engineering Journal (2025), **41**(3), 25-34

Sharif University of Technology

Sharif Civil Engineering Journal

https://sjce.journals.sharif.edu

Research Article

Performance of Deep Reinforcement Learning for Adaptive Left-Turn Phase Traffic Light Control

Elham Golpayegani, Abbas Babazadeh* and Omid Nayeri

School of Civil Engineering, College of Engineering, University of Tehran, Tehran, Iran
* corresponding author:(ababazadeh@ut.ac.ir)

Article Info

Article history:

Received: 21 July 2024

Revised: 23 January 2025

Accepted: 5 February 2025

Keywords:

Adaptive traffic light control, left-turn phase, reinforcement learning, double dueling deep Q-network.

Abstract

As traffic conditions become more complex and demanding, traditional methods of traffic signal control often fall short. The application of artificial intelligence and machine learning algorithms to traffic light timing has proven to be highly promising. This research uses reinforcement learning to manage traffic light phases automatically and efficiently, enhancing traffic flow and reducing intersection queue lengths. This paper examines the effectiveness of deep reinforcement learning techniques in optimizing the adaptive control of left-turn phases at urban intersections. The study introduces two deep reinforcement learning algorithms and compares the performance of the Double Dueling Deep Q-Network (3DQN) with the standard Deep Q-Network (DQN). These value-based methods in our proposed method use reinforcement learning optimization to determine the green duration for each phase and select either the protected or permitted left-turn phase for the next cycle. The adaptive control system adjusts traffic light timings in real-time without human intervention, ensuring smoother and more efficient traffic flow, significantly reducing queue lengths. The 3DQN algorithm uses a target network that updates target Q values at a slower rate to stabilize training and minimize errors. The dueling network splits the neural network into two parts: one to estimate the expected reward and the other to assess the relative importance of each action. Simulations were conducted with both uniform and variable car flow distributions, under light and heavy traffic volumes. They show that controllers using the 3DQN algorithm outperform the DQN algorithm. The results also reveal that the 3DQN algorithm can reduce cumulative vehicle queue lengths by at least 26% in all cases, and up to 67% in scenarios with heavy and uniform traffic flow. This research is crucial in developing intelligent traffic control systems and reducing traffic delays. The study highlights the potential of adaptive control systems using reinforcement learning to optimize traffic light timings and mitigate vehicle queue lengths, supporting the advancement of intelligent traffic control systems capable of adapting to dynamic urban conditions.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The authors declare no conflict of interest.

Author Contributions: All authors contributed to the reported work for conceptualization, methodology, validation, writing-review and editing, and supervision.

To Cite this article:
Golpayegani, E., Babazadeh, A. and Nayeri, O. 2025. Performance of deep reinforcement learning for adaptive left-turn phase traffic light control, Sharif Civil Engineering Journal, **41**(3), 25-34.
https://doi.org/10.24200/j30.2024.64476.3329

E-ISSN: 2676-4776

مجله علمی مهندسی عمران شریف

https://sjce.journals.sharif.edu

مقاله‌ی پژوهشی

عملکرد یادگیری تقویتی عمیق در کنترل تطبیقی فاز گردش به چپ چراغ راهنمایی

الهام گلپایگانی، عباس بابازاده* و امید نیری

دانشکده‌ی مهندسی عمران، دانشکدگان فنی، دانشگاه تهران، تهران.

*نویسنده مسئول (ababazadeh@ut.ac.ir)

اطلاعات مقاله

تاریخ دریافت : ۱۴۰۳/۰۴/۳۱

تاریخ اصلاحیه : ۱۴۰۳/۱۱/۰۴

تاریخ پذیرش : ۱۴۰۳/۱۱/۱۷

واژگان کلیدی:

کنترل تطبیقی چراغ ترافیکی،

فاز گردش به چپ،

یادگیری تقویتی،

شبکه‌ی عمیق دوئل دوگانه.

چکیده

در نوشتار حاضر، عملکرد دو روش یادگیری تقویتی، شبکه‌ی Q عمیق دوئل دوگانه و شبکه‌ی Q عمیق استاندارد در کنترل تطبیقی فاز گردش به چپ چراغ‌های راهنمایی در یک تقاطع شهری مقایسه شده‌اند. روش‌های مقدار- محور ذکرشده، با استفاده از بهینه‌سازی در یادگیری تقویتی، مدت زمان سبز هر فاز را تعیین و یکی از دو فاز گردش به چپ محافظت‌شده یا مجاز را برای سیکل بعدی انتخاب می‌کنند. شبیه‌سازی‌ها برای حالت‌های توزیع یکنواخت و متغیر جریان خودروها و با دو جریان ترافیک سبک و سنگین انجام شده و نتایج نشان داده‌اند که الگوریتم شبکه‌ی عمیق دوئل دوگانه در فرایند یادگیری مؤثرتر از الگوریتم شبکه‌ی Q استاندارد عمل کرده است. همچنین، یادگیری با شبکه‌ی Q دوئل دوگانه توانسته است طول صف تجمعی وسائط نقلیه را در تمام حالت‌های شبیه‌سازی، دست‌کم به میزان ۲۶٪ کاهش دهد و جریان ترافیک را بهبود بخشد. کاهش اخیر در حالت جریان ترافیک سنگین و یکنواخت بیشتر از سایر حالت‌ها بوده و به ۶۷٪ رسیده است. پژوهش حاضر می‌تواند نقش مهمی در توسعه‌ی سیستم‌های هوشمند کنترل ترافیک ایفا کند.

۱. مقدمه

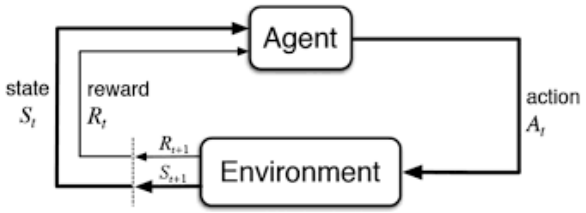
به‌علت محدودیت فضاهای شهری و نبود سرمایه‌ی کافی برای ایجاد زیرساخت‌های جدید، کنترل و استفاده‌ی بهینه از تسهیلات موجود ضروری به‌نظر می‌رسد.[۱] یکی از نقاط مهم برای کنترل جریان ترافیک، تقاطع‌های چراغ‌دار هستند؛ که به‌دلیل افزایش مصرف سوخت و تأخیرهای صورت‌گرفته، پتانسیل مناسبی برای کنترل و بهبود جریان ترافیک وجود دارد و ابزار کنترل مناسبی که همان چراغ ترافیکی است، نیز موجود است.[۲] به‌طورکلی، سه نوع کنترل چراغ ترافیکی وجود دارد. در نوع از پیش زمان‌بندی‌شده[۱] (زمان‌بندی ثابت) براساس داده‌های ترافیکی گذشته، یک زمان سبز ثابت برای هر فاز در نظر گرفته می‌شود. نوع کنترل فعال چراغ[۲] با استفاده از تقاضایی که از آشکارسازهای حلقه‌ی القایی[۳] در تقاطع دریافت می‌کند، تصمیم می‌گیرد که زمان سبز یک فاز را ادامه یا آن را به فاز بعدی اختصاص دهد.[۳] در واقع، کنترل‌کننده‌ی فعال با یک برنامه‌ی پایه، که برای همه‌ی فازها، کمینه و بیشینه‌یی از زمان سبز دارد، کار می‌کند. روش کار بر پایه‌ی یک قاعده است، که با توجه به درخواست تقاضا، بر مدت زمان سبز در فاز فعلی می‌افزاید. درخواست‌ها برای فاز فعلی توسط وسائط نقلیه‌ای که به تقاطع نزدیک می‌شوند و برای فازهای مقابل با انتظار برای وسائط نقلیه در رویکردهای دیگر ایجاد می‌شوند.[۴] در نوع سوم، نیز که کنترل تطبیقی[۴] نام دارد، زمان‌بندی به‌طور خودکار با توجه به وضعیت فعلی تقاطع، مدیریت و به‌روز می‌شود.[۳] ویژگی اصلی کنترل تطبیقی، بهینه‌سازی زمان‌بندی چراغ با توجه به تابع هدفِ تعیین‌شده است. در اینجا نظارت و پایش جریان خودروها در رویکردهای تقاطع، مانند کنترل فعال صورت می‌گیرد. تفاوت اصلی دو نوع کنترل ذکرشده این است که کنترل تطبیقی، هم داده‌های ترافیکی در لحظه[۵] و هم پیش‌بینی تغییرات احتمالی در جریان ترافیک را در نظر می‌گیرد.[۴] روش کنترل تطبیقی با حل یک مسئله‌ی بهینه‌سازی به کمک هوش مصنوعی و الگوریتم‌های یادگیری ماشین، نتایج امیدوارکننده‌ای داشته است؛[۳] که از جمله‌ی آن‌ها، می‌توان به روش‌ها الگوریتم ژنتیک[۶] یا الگوریتم تکاملی[۷]، منطق فازی[۸]، و یادگیری تقویتی[۹] اشاره کرد.[۵] مطالعه‌ی حاضر که بر کنترل تطبیقی چراغ ترافیکی متمرکز بوده است، به کمک یادگیری تقویتی، مدت زمان سبز بهینه‌ی هر فاز را در یک تقاطع چهارراه تعیین کرده است.

[1] Pre-timed signal control
[2] Actuated signal control
[3] Inductive loop detector
[4] Adaptive signal control
[5] Real-time
[6] Evolutionary algorithm
[7] Evolutionary algorithm
[8] Fuzzy logic
[9] Reinforcement Learning

استناد به این مقاله:

گلپایگانی، الهام، بابازاده، عباس و نیری، امید، ۱۴۰۴. عملکرد یادگیری تقویتی عمیق در کنترل تطبیقی فاز گردش به چپ چراغ راهنمایی. مهندسی عمران شریف، ۴۱(۳)، صص.۲۵-۳۴. https://doi.org/ 10.24200/j30.2024.64476.3329

شاپا الکترونیکی: ۲۶۷۶-۴۷۷۶



شکل ۱. ساختار مدل یادگیری تقویتی.

سیستم‌های هوشمند حمل‌ونقل در کنار هوش مصنوعی می‌توانند راه‌حل‌های مناسبی برای چالش‌های قرن حاضر ارائه دهند. یادگیری تقویتی عمیق، یک از موفق‌ترین حوزه‌های هوش مصنوعی است، که به یادگیری انسان نزدیک است. از جمله کاربردهای آن در سیستم‌های هوشمند حمل‌ونقل خودروهای خودران، چراغ رمپ، تغییر خط عبور، شتاب، یا کاهش سرعت و مانور در تقاطع‌هاست، که پرطرفدارترین آن‌ها کنترل تطبیقی چراغ راهنمایی در تقاطع‌هاست.[۶] یادگیری تقویتی، یک رویکرد محاسباتی برای یادگیری و تصمیم‌گیری بر پایه‌ی هدف است و برخلاف دیگر رویکردهای یادگیری، به نظارت و یا مدل‌های کامل از محیط[۱] نیاز ندارد.[۷] در یادگیری تقویتی، یک عامل[۲] می‌کوشد با تعامل با محیط پیرامون و به کمک سعی و خطا، در هر گام به بهترین اقدام[۳] ممکن دست یابد. به این صورت که عامل یک وضعیت[۴] از محیط را مشاهده می‌کند و براساس آن یک اقدام انجام می‌دهد، که یک پاداش[۵] در بر دارد. ساختار مدل یادگیری تقویتی در شکل ۱ مشاهده می‌شود. در واقع، عامل در تلاش است تا یک سیاست[۶] کنترل بهینه بیابد، تا پاداش تجمعی از طریق تعامل مکرر با محیط به میزان بیشینه برسد. این پاداش موردانتظار به ازاء هر جفت وضعیت- اقدام[۷]، مقدار Q[۸] نامیده می‌شود.[۳] از آنجا که تعیین مقادیر از تعیین پاداش‌ها بسیار دشوارتر است، تقریباً تمام الگوریتم‌های یادگیری تقویتی به دنبال روش‌هایی برای تخمین مقادیر بهینه هستند. بسیاری از روش‌های یادگیری تقویتی پیرامون تخمین تابع مقدار[۹] هستند. همچنین باید در نظر داشت در یادگیری تقویتی، وضعیت به‌عنوان ورودی برای سیاست و تابع مقدار در نظر گرفته می‌شود.[۷]

پژوهشگران به مسئله‌ی کنترل تطبیقی چراغ به کمک یادگیری تقویتی از جهت‌های مختلف توجه کرده‌اند. هر چند تعریف مسئله در مطالعات مذکور یکسان بوده است، اما تعریف وضعیت‌ها، اقدام‌ها، پاداش‌ها، سیاست‌ها، محیط، توابع مقدار، و همچنین معیارهای بهینه‌سازی در هر یک متفاوت است. به‌علاوه از شبیه‌سازهای گوناگونی نیز برای ایجاد شرایط چهارراه با توزیع‌های متنوع استفاده شده است.[۳] با توجه به نتایج امیدوارکننده‌ای که ترکیب ساختار شبکه‌ی عصبی عمیق و روش‌های یادگیری تقویتی داشته است، دو نوع الگوریتم یادگیری تقویتی ارائه شده است، که یکی بر پایه‌ی گرادیان سیاست عمیق و دیگری مبتنی بر تابع مقدار است. عامل مبتنی بر پایه‌ی گرادیان سیاست عمیق، یک تصویر لحظه‌ای از وضعیت فعلی شبیه‌ساز ترافیکی دریافت می‌کند و بر پایه‌ی گرادیان سیاست عمیق، توسط مشاهده‌های خود، کنترل بهینه‌ی چراغ را مستقیما می‌یابد؛ اما عامل مبتنی بر تابع مقدار، ابتدا مقادیر اقدام‌های ممکن برای کنترل چراغ را تخمین می‌زند و سپس اقدام با بیشترین مقدار را انتخاب می‌کند. روش‌های یادگیری عمیق، یکی از بهترین راهکارها برای مسائلی هستند که داده‌ها، ابعاد زیادی دارند و استخراج ویژگی‌های موردنظر دشوار است. پس از ورود به یک شبکه‌ی عمیق، خروجی داده‌ها با گذر از هر لایه به‌عنوان ورودی لایه‌ی بعد در نظر گرفته می‌شوند و ویژگی‌های غیرخطی آن‌ها هر بار تغییرشکل می‌دهد. یک شبکه‌ی عمیق یادگیری Q از مشاهده‌های عامل استفاده می‌کند و آن را برای یادگیری سیاست کنترل بهینه به‌کار می‌برد. در واقع، روش شبکه‌ی یادگیری عمیق Q، شبکه‌ی عصبی عمیق را با یادگیری Q (یک الگوریتمِ بدون مدل در یادگیری تقویتی) تلفیق می‌کند تا بتواند تابع اقدام- مقدار و در نتیجه سیاست موردنظر عامل را بیابد. تفاوت دو رویکرد مبتنی بر تابع مقدار (مقدار- محور) و بر پایه‌ی سیاست (سیاست- محور) در آخرین لایه‌ی شبکه‌ی یادگیری عمیق Q قابل بررسی است. به این صورت که در رویکرد مبتنی بر تابع مقدار، آخرین لایه‌ی شبکه‌ی یادگیری عمیق Q، یک لایه‌ی خطی کاملاً متصل با تعدادی نورون خروجی (مقدار Q) متناظر به هر اقدام است. در حالی که در مدل دوم، لایه‌ی آخر، دو مجموعه‌ی خروجی را نشان می‌دهد، یکی که منجر به توزیع احتمال بر روی اقدام‌ها (یعنی همان سیاست) می‌شود و دیگری یک گره‌ی خطی منفرد است، که منجر به تخمین تابع وضعیت- مقدار می‌شود. روش‌های گرادیان سیاست می‌کوشند یک تابع سیاست پارامتری‌شده را با روش گرادیان نزولی بهینه سازند. به این صورت که مستقیماً به دنبال یادگیری سیاست بهینه در فضای سیاست‌ها هستند.

به‌علت شرایط دشوار برداشت داده در محیط دنیای واقعی، کاربردهای یادگیری تقویتی برای کنترل چراغ ترافیکی در شبیه‌ساز اجرا می‌شود. SUMO[۱۰]، محبوب‌ترین شبیه‌ساز ترافیکی منبع باز است. این شبیه‌ساز با کتابخانه‌ی TraCI[۱۱] در پایتون، امکان تعامل کاربر با محیط را فراهم می‌کند.[۶]

تعریف وضعیت، اهمیت بسیاری دارد و به دستگاه‌های تجهیزات ترافیکی گوناگون، مانند دوربین و آشکارساز حلقه‌ای بستگی زیادی دارد.[۶] در مطالعه‌ی موسوی و همکاران (۲۰۱۷)،[۳] نمایش وضعیت‌ها به‌صورت تصویر، ورودی به یک شبکه‌ی عصبی پیچشی است. در این صورت با ذخیره و سپس تغذیه‌ی لایه‌های شبکه‌ی عمیق پیچشی با تصاویری از مشاهده‌های متوالی، نه فقط می‌توان موقعیت و حضور خودروها را در هر خط عبور تشخیص داد، بلکه سرعت و جهت حرکت آن‌ها نیز اطلاعات ارزشمندی به عنوان وضعیت برای سیستم خواهد بود. در هر گام زمانی نیز اقدام‌های ممکن، اختصاص فاز سبز به معبر شمالی- جنوبی یا شرقی- غربی است. این در حالی است که در مطالعه‌ی اریکسن[۱۲] و همکاران (۲۰۲۰)،[۲] در هر گام زمانی یک ثانیه‌ای تصمیم‌گیری می‌شود که زمان سبز ادامه یابد یا به جهت دیگر، زمان سبز داده شود. در پژوهش طوبی[۱۳] و همکاران (۲۰۱۷)،[۸] تعریف جدیدی برای وضعیت‌ها معرفی شده است. در بیشتر مطالعات، بیشینه‌ی طول صف در هر فاز برای تعریف وضعیت به‌کار می‌رود؛ زیرا تعداد خودروها در یک صف، نمایش قدرتمندتری از وضعیت ترافیکی نسبت به دیگر موارد دارند. اما اگر ابعاد وسائط نقلیه نیز در

[1] Environment
[2] Agent
[3] Action
[4] State
[5] Reward
[6] Policy
[7] State - action
[8] Q-Value
[9] Value Function
[10] Simulation of Urban MObility
[11] Traffic Control Interface
[12] Eriksen
[13] Touhbi

نظر گرفته نشود، یادگیری می‌تواند دچار مشکل شود. برای نمونه، در یک فاصله‌ی ۱۰ متری از تقاطع ممکن است دو خودروی سواری یا یک خودروی سنگین وجود داشته باشند که در صورت فرض ذکرشده با وجود عملکرد متفاوت در محیط تقاطع، طول صف یکسان داشته باشند. مطالعه‌ی اخیر از مفهومی به نام صف باقی‌مانده[14] بهره برده است، که بیانگر طول صف در یک خط عبور تقسیم بر طول خط عبور است. نتایج پژوهش مذکور نشان می‌دهند که تعریف جدید برای وضعیت از جهت توان عملیاتی تقاطع و متوسط تأخیر، به‌ویژه در حجم بالای وسائط نقلیه و شرایط متغیر ترافیکی عملکرد خوبی داشته است. ژنگ[15] و همکاران (۲۰۱۹)،[۹] به بررسی این نکته پرداخته‌اند که آیا برای تعریف وضعیت‌ها لزومی به کاربرد نمایش‌های پیچیده، مانند تصاویر وجود دارد یا خیر. چه بسا که این موضوع ممکن است باعث کُندشدن فرایند آموزش در یادگیری تقویتی شود، در حالی که ثمره‌ی قابل‌توجهی نیز نداشته باشد. در نوشتار اخیر با برقراری ارتباط بین یادگیری تقویتی و رویکردهای معمول حمل‌ونقلی نشان داده شده است که با فرض جریان یکنواخت ترافیک، اولاً طول صف به‌عنوان پاداش در یادگیری تقویتی معادل بهینه‌سازی زمان سفر در روش‌های حمل‌ونقلی است؛ دوم اینکه تعداد خودروها در هر خط عبور به همراه اطلاع از فاز فعلی چراغ، پویایی سیستم را کاملاً شرح می‌دهد. با داشتن دو ویژگی اخیر، عامل می‌تواند به سیاست کنترل بهینه دست یابد.

لا[16] و همکارش (۲۰۱۱)،[۱۰] برای اولین بار مسئله‌ی کنترل تطبیقی چراغ به کمک الگوریتم یادگیری تقویتی را با تقریب تابع بررسی کرده‌اند. الگوریتم یادگیری Q، یکی از مهم‌ترین الگوریتم‌های یادگیری تقویتی است، که به سیاست بهینه همگرا می‌شود. همچنین ابعاد گسترده‌ی فضای وضعیت- اقدام در مسئله‌ی کنترل چراغ موجب به‌کارگیری روش‌های تقریب تابع برای کارایی محاسباتی شده است. جندرز[17] و رضوی (۲۰۱۶)،[۱۱] از یکی از انواع تقریب‌گرهای توابع استفاده کرده‌اند، که در آن تابع اقدام- مقدار به‌عنوان یک شبکه‌ی عصبی پیچشی عمیق مدل می‌شود. زیرا شبکه‌های عصبی مصنوعی، قابلیت خوبی برای تقریب تابع دارند.

یکی از مشکلات یادگیری تقویتی عمیق، انتخاب یک تابع مناسب برای پاداش است؛ به‌خصوص که بازخورد به عامل برای یک یادگیری پایدار و سریع به تابع مذکور بستگی دارد. پاداش می‌تواند جمع وزنی از ویژگی‌های خطوط عبوری ورودی به تقاطع باشد. مشخصاتی، شامل: طول صف، تأخیر، زمان انتظار به‌روزشده، مقادیر ۰ و ۱ برای فاز فعلی چراغ، تعداد خودروهای عبوری از تقاطع، و کل زمان سفر وسائط نقلیه از آن جمله‌اند.[۱۲] معمولاً هدف مسئله‌ی کنترل چراغ، کمینه‌سازی تأخیر خودروها در تقاطع است. به جای تأخیر، تابع پاداش در مطالعه‌ی ژنگ و همکاران (۲۰۲۲)، تعداد خودروهای خروجی از تقاطع بین دو گام زمانی متوالی تعریف شده است، که هم در زمان آموزش مدل، بهره‌گیری از آن آسان است و هم برخلاف توابع پاداش استفاده‌شده در بیشتر مطالعات می‌تواند در کوتاه‌مدت نتایج مناسبی برای تشویق یا تنبیه رفتار عامل ارائه دهد. جالب توجه است که تابع پاداش براساس تأخیر نیست، اما می‌تواند به‌تدریج موجب کاهش تأخیر متوسط خودروها شود.[۱۳] با تحلیل تعاریف مختلف پاداش در یادگیری تقویتی می‌توان دریافت که عملکرد توابع پاداش به حجم وسائط نقلیه در تقاطع، وسائط نقلیه‌ی مجهز به GPS، و نیز تجهیزات استفاده‌شده برای پایش تقاطع وابسته است. می‌توان تصور کرد که برداشت مواردی مانند تأخیر تجمعی از محیط به ادوات پیشرفته‌تری، مانند نظارت تصویری یا وسائط نقلیه‌ی مجهز به GPS نیاز دارد.[۸]

تقاطع‌های چراغ‌دار می‌کوشند فضا و زمان را به گونه‌ای به حرکت‌های مختلف ترافیکی تخصیص دهند که ایمنی و کارآمدی تأمین شود.[۱۴] هان[18] و همکاران (۲۰۲۲)،[۱۵] با درنظرگرفتن مسائل مربوط به عابران پیاده در تقاطع‌ها، روش کنترل به کمک یادگیری تقویتی را با جنبه‌های ایمنی ترکیب کرده‌اند. در روش مذکور، زمان انتظار عابران پیاده و نیز خودروهای عبوری از تقاطع به‌طورهم‌زمان در نظر گرفته شده است. در حالی که گاهی اوقات، اندرکنش بین وسائط نقلیه‌ی گردش به چپ و جریان ترافیک مخالف می‌تواند خطرهای ایمنی ایجاد کند. بنابراین، برای افزایش ایمنی و کارایی تقاطع‌های دارای چراغ راهنمایی، انتخاب فازهای مناسب برای کمینه‌سازی تأخیرهای تقاطع بسیار مهم است.[۱۴] افندی‌زاده و همکاران (۲۰۱۴)، در بررسی عوامل مؤثر در ایمنی تقاطع‌ها، میزان اهمیت عواملی مانند: عوامل انسانی، زیرساختی، تسهیلات و تجهیزات، و عوامل ترافیکی را با استفاده از روش تحلیل سلسله‌مراتبی تعیین کرده و دریافته‌اند که عوامل تأثیرگذار تسهیلات و تجهیزات در زمینه‌ی تجهیزات کنترل ترافیک، نوع کنترل چراغ راهنمایی، فازبندی چراغ راهنمایی، نمایشگرهای عابران پیاده و وسائط نقلیه، خط‌کشی‌ها، تابلوها، و روشنایی هستند. در این میان، نوع کنترل چراغ، بیشترین اهمیت را داشته و در مرتبه‌ی اول قرار گرفته است، فازبندی چراغ در مرتبه‌ی دوم و نمایشگرهای چراغ در مرتبه‌ی آخر قرار گرفته‌اند. همچنین، با بررسی میزان اهمیت عوامل ترافیکی مشخص شده است که حجم گردش به چپ در مرتبه‌ی اول، ترکیب تردد در مرتبه‌ی دوم، جریان ترافیک در مرتبه‌ی سوم، و حجم گردش به راست در مرتبه‌ی آخر قرار داشته‌اند. بنابراین، به‌کارگیری و طراحی مناسب چراغ‌های راهنمایی، به‌عنوان یکی از عوامل اصلی در بهبود ایمنی و کارایی تقاطع‌های چراغ‌دار شناخته شده است.[۱۶]

مسئله‌ای که در مطالعه‌ی حاضر بررسی شده است، عملکرد شبکه‌ی دوئل دوگانه در یادگیری تقویتی عمیق برای کنترل تطبیقی چراغ ترافیکی در یک تقاطع چهارراهی در در سیکل‌های متوالی چراغ بوده است. هر ورودی تقاطع، سه خط عبوری داشته و فرض بر این بوده است که جریان ورودی به تقاطع، تحت تأثیر جریان بالادست قرار ندارد. عامل کنترل‌کننده‌ی چراغ در هر گام، وضعیت محیط (تقاطع چراغ‌دار) را مشاهده می‌کند. این وضعیت می‌تواند شامل پارامترهای مختلف ترافیکی باشد. با هدف کاهش تداخل حرکت‌های گردشی خودروها با جریان‌های مخالف، یک روش فازبندی منعطف ارائه شده است، که هر سیکل توسط یک عامل دیگر یادگیری تقویتی تعیین می‌شود. روش مذکور امکان تغییر تعداد فازهای هر سیکل را بسته به وضعیت ترافیک فراهم و از کمینه‌ی ۲ تا بیشینه‌ی ۴ فاز مختلف استفاده می‌کند. به این ترتیب، در هر سیکل، ترکیب بهینه‌ی فازها برای کاهش زمان انتظار خودروها انتخاب می‌شود. هدف از مطالعه‌ی حاضر، دستیابی به یک الگوریتم مناسب در حوزه‌ی یادگیری تقویتی عمیق برای کنترل تطبیقی چراغ راهنمایی با استفاده از نوآوری پیشنهادی (منعطف‌سازی فازبندی گردش به چپ) بوده است. نوشتار حاضر به شناسایی ترکیب‌های بهینه‌ی فازبندی چراغ‌های ترافیکی پرداخته است. اگرچه

[14] Residual queue
[15] Zheng
[16] La
[17] Genders
[18] Han

هدف اصلی آن افزایش ایمنی نبوده است، اما توجه ویژه‌ای به تداخل‌های ناشی از گردش به چپ وسائط نقلیه و حرکت‌های روبرویی داشته است. علاوه‌بر این، مطالعات پیشین عمدتاً بر ارزیابی بهبودهای مربوط به حرکت‌های گردش به چپ براساس داده‌های تاریخی تصادف تمرکز داشته‌اند، اما این داده‌ها اطلاعات پیش‌بینی‌کننده‌ای ارائه نمی دهند بنابراین، ایجاد ابزارهایی که بتوانند تعادل مناسبی بین ایمنی و کارایی تقاطع برقرار کنند و تعیین بهترین فاز گردش به چپ را تسهیل سازند، اهمیت بالایی دارد.

۲. مدل یادگیری تقویتی در مسئله‌ی کنترل تطبیقی چراغ راهنمایی

در مطالعه‌ی حاضر، نوعی فازبندی منعطف برای چراغ‌های ترافیکی با گام زمانی ۱۰ ثانیه‌ای در نظر گرفته شده و مدت زمان سبز هر فاز شامل چند گام زمانی بوده است. ترتیب فازها در هر سیکل ثابت بوده است. حرکت‌های گردش به چپ برای ورودی‌های مقابل معابر (شمالی- جنوبی یا شرقی- غربی) یا به‌صورت مجاز در فاز مشترک با حرکت‌های مستقیم و گردش به راست انجام شده یا به‌صورت محافظت‌شده در یک فاز مجزا و از خط اختصاصی گردش به چپ صورت گرفته است. بنابراین، چهار انتخاب مختلف فازبندی گردش به چپ در هر سیکل وجود داشته است:

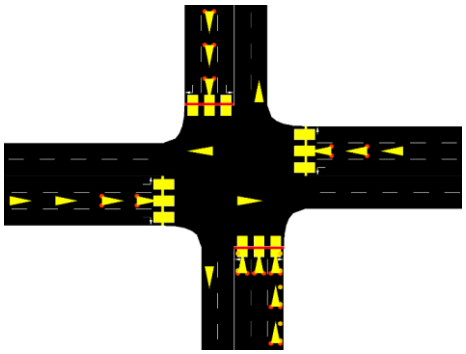
- شمالی- جنوبی: مجاز/ شرقی- غربی: مجاز؛
- شمالی- جنوبی: محافظت‌شده/ شرقی- غربی: مجاز؛
- شمالی- جنوبی: مجاز/ شرقی- غربی: محافظت‌شده؛
- شمالی- جنوبی: محافظت‌شده/ شرقی- غربی: محافظت‌شده.

در پژوهش حاضر، مقادیر داده‌های استفاده‌شده برای کنترل تطبیقی چراغ راهنمایی براساس داده‌های واقعی برداشت‌شده از تقاطع بلوار کشاورز و وصال شیرازی در تهران، در شبیه‌ساز پیاده‌سازی شده‌اند. اطلاعات دریافتی از سازمان ترافیک شهرداری تهران نشان داده است که در طول ساعت اوج صبح (۷:۳۰ تا ۸:۳۰) در یک روز معمولی پاییزی در سال ۱۴۰۲، ۵۵۰۶ واحد خودرو سواری (PCU) وارد تقاطع مذکور شده و حرکت‌های راست، مستقیم، و چپ انجام داده‌اند. با توجه به موقعیت تقاطع اخیر در مرکز شهر و حجم بالای تردد در بلوار کشاورز، تعداد خودروهای شبیه‌سازی‌شده تعدیل شده است. هدف مطالعه‌ی حاضر، ایجاد محیطی در شبیه‌ساز بوده است، که نمایانگر بسیاری از تقاطع‌های شهری باشد. بنابراین، تعداد خودروهای عبوری و هندسه‌ی تقاطع به‌صورت نسخه‌ای تعدیل‌شده در نظر گرفته شده‌اند. در محیط شبیه‌ساز، ورود خودروها به تقاطع با دو توزیع متغیر ویبل و یکنواخت انجام شده‌اند. نرخ جریان ورودی در حالت ترافیک کم، ۱۰۰۰ و در حالت ترافیک زیاد، ۲۰۰۰ خودرو در ساعت در نظر گرفته شده است. زمان شبیه‌سازی نیز ۱ ساعت لحاظ شده است.

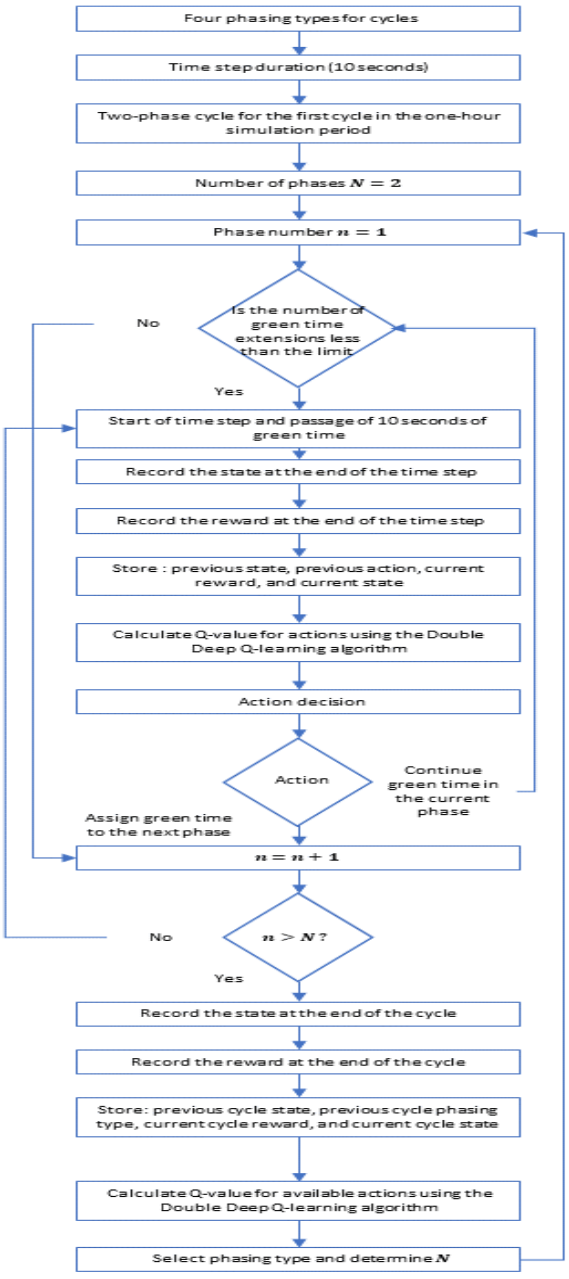
۱.۲. محیط و نمایش وضعیت‌ها

تقاطع شامل سه خط عبور برای ورود خودروها از هر جهت است، که شامل یک خط اختصاصی برای گردش به راست، یک خط برای گردش به چپ، و یک خط برای حرکت مستقیم در ۱۰۰ متری تقاطع است. محیط تقاطع چهارراه در شبیه‌ساز SUMO پیاده‌سازی و به همراه موقعیت آشکارسازهای حلقه‌ای در شکل ۲ نمایش داده شده است.

با استفاده از آشکارسازهای حلقه‌ای، که در هر یک از خطوط قرار داده شده‌اند، می‌توان تعداد کل خودروهای عبوری از هر رویکرد تقاطع را در گام‌های زمانی مشخص جمع‌آوری کرد. داده‌های اخیر می‌توانند برای تحلیل وضعیت جریان



شکل ۲. محیط تقاطع چهارراه و موقعیت آشکارسازهای حلقه‌ای.



شکل ۳. فلوچارت روش کنترل چراغ راهنمایی.

ترافیک و استفاده در یادگیری تقویتی استفاده شوند. فلوچارت روش کنترل چراغ راهنمایی در شکل ۳ مشاهده می‌شود.

همان‌گونه که اشاره شد، انتخاب نوع فازبندی سیکل بعد در پایان سیکل فعلی انجام می‌شود. در این راستا، باید به چند عامل توجه کرد: ظرفیت تقاطع،

State representation for the green time learning agent	State representation for the phasing type learning agent
Flow rate of vehicles in the westbound approach	Product of the left-turn flow rate of the westbound approach and the corresponding opposing through flow
Flow rate of vehicles in the northbound approach	Product of the left-turn flow rate of the northbound approach and the corresponding opposing through flow
Flow rate of vehicles in the eastbound approach	Product of the left-turn flow rate of the eastbound approach and the corresponding opposing through flow
Flow rate of vehicles in the southbound approach	Product of the left-turn flow rate of the southbound approach and the corresponding opposing through flow
Left-turn flow rate of the westbound approach	Cycle's phasing type
Left-turn flow rate of the northbound approach	
Left-turn flow rate of the eastbound approach	
Left-turn flow rate of the southbound approach	
Current cycle's phasing type	
Phase number in the current cycle	

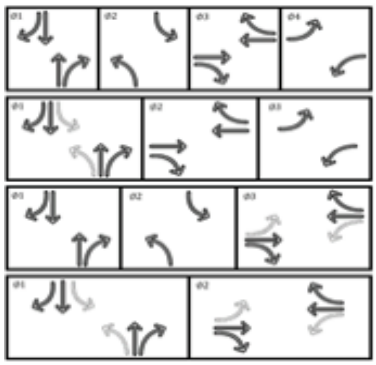
شکل ۴. مؤلفه‌های وضعیت‌های مشاهده‌شده توسط عامل یادگیری زمان سبز و نوع فازبندی.

ویژگی‌های هندسی، و تاریخچه‌ی تصادف؛ که عوامل معمولی هستند و در انتخاب فازبندی استفاده می‌شوند. اما یک عامل دیگر که نباید نادیده گرفته شود، حاصل‌ضرب حجم گردش به چپ و حجم مخالف است؛ که نشان می‌دهد چقدر تقاطع شلوغ و پیچیده است. بسیاری از پژوهشگران و مهندسان از عامل اخیر به‌عنوان معیار اصلی برای انتخاب طرح‌های فازبندی استفاده می‌کنند.[۱۴] به‌صورت دقیق‌تر، ابتدا حاصل‌ضرب نرخ جریان حرکت‌های گردش به چپ در نرخ جریان خودروهای مقابل، که در حرکت مستقیم با حرکت‌های گردش به چپ تداخل داشته‌اند، در هر چهار رویکرد محاسبه شده‌اند. لذا یک بار دیگر به کمک یادگیری تقویتی، نوع فازبندی انتخاب شده است. به این صورت که حاصل‌ضرب‌های ذکرشده برای هر ۴ رویکرد، به‌عنوان وضعیت در نظر گرفته شده است. وضعیت مشاهده‌شده توسط عامل یادگیری زمان سبز و نوع فازبندی در واقع یک بردار با ۱۰ و ۵ مؤلفه است، که در شکل ۴ مشاهده می‌شود.

۲.۲. مجموعه‌ی اقدام‌ها

عامل یادگیری تقویتی پس از مشاهده‌ی وضعیت‌ها و دریافت پاداش در پایان هر گام زمانی ۱۰ ثانیه‌ای، تصمیم به تمدید زمان سبز برای ۱۰ ثانیه‌ی دیگر یا رفتن به فاز بعد می‌گیرد. انتخاب اقدام به کمک یادگیری تقویتی صورت می‌پذیرد. به این صورت که دو اقدام ۰ و ۱ وجود دارد. عدد ۰ برای وقتی که زمان سبز تمدید می‌شود و عدد ۱ برای اختصاص آن به فاز بعدی. اگر عامل تصمیم به اختصاص زمان سبز به فاز بعدی بگیرد، ابتدا زمان زرد و سپس تمام‌قرمز طی خواهند شد. در پایان ۱۰ ثانیه‌ی ذکرشده، تفاوت زمان انتظار در سیکل قبلی و فعلی به‌عنوان پاداش آنی (مربوط به گام زمانی ۱۰ ثانیه‌ای) به‌دست می‌آیند؛ که در پژوهش حاضر، تعداد ۱۰۰ اپیزود (شبیه‌سازی ۱ ساعته) برای یادگیری در نظر گرفته شده است. این تذکر لازم است که با داشتن پاداش آنی گام‌های زمانی در طول ۱ ساعت شبیه‌سازی، می‌توان مشاهده کرد که عملکرد عامل یادگیری تقویتی چگونه با گذشت اپیزودها در جهت بهینه‌سازی (بیشینه‌سازی پاداش‌ها) تغییر می‌کند.

همان‌گونه که اشاره شد، در پایان هر سیکل، انتخاب نوع فازبندی سیکل بعد به کمک یادگیری تقویتی عمیق انجام پذیرفته است. چهار نوع فازبندی در نظر گرفته شده و سعی بر این بوده است که این بار نیز مناسب‌ترین نوع آن انتخاب شود. فازبندی منعطف پیشنهادی در شکل ۵ مشاهده می‌شود. ابتدا حاصل‌ضرب نرخ جریان حرکت‌های گردش به چپ در نرخ جریان خودروهای مقابل با حرکت مستقیم، که با حرکت‌های گردش به چپ تداخل داشته‌اند، به



شکل ۵. فازبندی منعطف پیشنهادی.

همراه نوع فازبندی سیکل قبل به عنوان وضعیت در نظر گرفته شده است. سپس یک تعریف پاداش برای عامل انتخاب نوع فازبندی در نظر گرفته شده است، که تفاوت زمان انتظار در سیکل قبلی و فعلی بوده است.

۳.۲. تابع پاداش

اگر $wait$ زمان انتظار خودروها باشد، یک تعریف پاداش برای عامل به‌صورت رابطه‌ی ۱ در نظر گرفته شده است. به این صورت که عامل می‌کوشد با بیشینه‌سازی آن، عملکرد خود را در انتخاب اقدام مناسب بهبود بخشد.

$$r_t = wait_{t-1} - wait_t \quad (1)$$

با توجه به تعریف پاداش تجمعی، با افزایش تعداد اپیزودها هر چه مقدار پاداش بیشتر شود، عامل مذکور عملکرد بهتری در یادگیری تقویتی از خود نشان می‌دهد و برعکس.

۴.۲. مدل یادگیری

همان‌گونه که اشاره شد، هدف عامل بیشینه‌سازی پاداش تجمعی‌ای است که در درازمدت دریافت می‌کند، که در ساده‌ترین حالت، حاصل جمع پاداش‌هاست:

$$R_t \doteq r_t + r_{t+1} + r_{t+2} + \cdots \quad (2)$$

برای جلوگیری از بی‌نهایت‌شدن مقدار پاداش تجمعی، از تخفیف استفاده می‌شود؛ که براساس آن، عامل سعی می‌کند عملیاتی را انتخاب کند که مجموع پاداش‌های تخفیفی که در آینده دریافت می‌کند به میزان بیشینه برسد:

$$\begin{aligned} R_t &\doteq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \\ &= \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau \end{aligned} \quad (3)$$

که در آن، γ پارامتری است به نام نرخ تنزیل که: $0 \le \gamma \le 1$. همچنین، R_t مطابق رابطه‌ی ۴ محاسبه می‌شود:

$$\begin{aligned} R_t &\doteq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \\ &= r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \cdots) \\ &= r_t + \gamma R_{t+1} \end{aligned} \quad (4)$$

بنابراین، R_t در گام‌های زمانی متوالی به گونه‌ای با یکدیگر مرتبط هستند.

برای عاملی که طبق یک سیاست تصادفی π رفتار می‌کند، مقادیر جفت وضعیت- اقدام (s,a) و وضعیت s به صورت روابط ۵ و ۶ تعریف می‌شوند:

$$Q^\pi(s,a) = E[R_t \mid s,a,\pi] \quad (5)$$

$$V^\pi(s) = E_{a\sim\pi(s)}[Q^\pi(s,a)] \quad (6)$$

تابع مقدار وضعیت- اقدام (تابع Q) روشی برای تعیین کمیت مزیت انجام یک اقدام خاص در یک وضعیت معین است و با تابع مقدار V، که فقط میزان

جدول ۱. ابرپارامترهای استفاده‌شده در مدل یادگیری تقویتی عمیق.

Hyper parameters	Max speed (m/s)	Demand (vph)		Discount rate	Batch size	Memory size	Replace number	Batch size (phasing type)	Memory size (phasing type)
		Low	**High**						
Value	15	1000	2000	0.8	252	10000	100	64	2000

خوب‌بودن یک وضعیت s را تخمین می‌زند، متفاوت است. با توجه به معادله‌ی ۴، تابع Q را می‌توان با استفاده از برنامه‌نویسی پویا به صورت بازگشتی (مطابق رابطه‌ی ۷) محاسبه کرد:

$$Q^{\pi}(s,a) = E_{s'}[r + \gamma E_{a' \sim \pi(s')}\ [Q^{\pi}(s',a')] \mid s,a,\pi] \quad (7)$$

که در آن، مقادیر جفت وضعیت- اقدام (s,a) در گام فعلی و مقادیر (s',a') در گام بعدی هستند. مفهوم اساسی شبکه‌ی عمیق Q، یافتن تابع Q است، که به‌عنوان ورودی وضعیت فعلی را می‌گیرد و بازده موردانتظار هر اقدام را خروجی می‌دهد. تابع مقدار اقدام بهینه با توجه به یک محیط می‌تواند با استفاده از معادله‌ی بلمن به صورت به‌صورت رابطه‌ی ۸ بیان شود:[۱۸]

$$Q^{*}(s,a) = E_{s'}[r + \max_{a'} Q^{*}(s',a') \mid s,a] \quad (8)$$

در پژوهش حاضر، از شبکه‌ی عصبی به‌عنوان یک تقریب غیرخطی برای تخمین تابع مقدار- اقدام استفاده شده است. شبکه، یک نمایش از وضعیت را به عنوان ورودی می‌گیرد و برای هر اقدام ممکن یک بردار از مقادیر Q را خروجی می‌دهد. بنابراین، برای هر یک از اقدام‌های ذکرشده، یک مقدار Q وجود دارد، که در هر بار تصمیم‌گیری عامل، اقدام با مقدار بیشتر انتخاب می‌شود. در مطالعه‌ی حاضر، مانند بسیاری از پژوهش‌های پیشین، از یک شبکه‌ی عصبی عمیق برای تعیین مقدار Q استفاده شده است. برای آموزش عامل از یک مجموعه از نمونه‌ها استفاده شده است، که شامل چهارتایی: وضعیت قبلی، اقدام قبلی (انتخاب‌شده)، پاداش، و وضعیت فعلی در هر گام است. این مجموعه از یک مجموعه‌ی بزرگ‌تر به نام حافظه به‌صورت تصادفی برداشته می‌شود. در واقع، پس از تجربه‌ی عامل در هر گام، از مشاهده‌ی وضعیت تا انتخاب اقدام مناسب، وضعیت قبلی، اقدام قبلی (انتخاب‌شده)، پاداش، و وضعیت فعلی به این حافظه افزوده می‌شود. پس از پُرشدن اندازه‌ی آن نیز تجربه‌های قدیمی‌تر به‌تدریج حذف و تجربه‌های جدید جایگزین می‌شوند. از یک شبکه‌ی عصبی عمیق برای تناظر وضعیت‌ها به مقدار Q اقدام‌ها استفاده شده است، تا بتوان فرمول اصلی یادگیری تقویتی را اعمال کرد.

مدل به گونه‌ای طراحی شده است که در طول فرایند آموزش، نمونه‌های تصادفی از وضعیت‌ها را که از حافظه استخراج شده‌اند، به عنوان ورودی دریافت کند. در مطالعه‌ی حاضر، شبکه‌ی عصبی با ترکیب‌های مختلفی از لایه‌ها، گره‌ها، و توابع فعال‌سازی آزمایش شده است. براساس تلاش‌هایی که منجر به کسب پاداش‌های بهتر شده‌اند، ساختاری با ۳ لایه‌ی پنهان پیاده‌سازی شده است، که به ترتیب شامل ۱۶، ۳۲، و ۶۴ گره است. در لایه‌ی اول از تابع فعال‌سازی SELU استفاده شده و در لایه‌های بعدی (از جمله لایه‌ی نهایی)، تابع ReLU به‌کار رفته است. لایه‌ی نهایی که مقادیر Q را تخمین می‌زند، به تعداد اقدام‌های ممکن گره دارد. با استفاده از شبکه‌ی اخیر، پس از مشاهده‌ی وضعیت در شبیه‌ساز ترافیک، اقدام با بالاترین مقدار انتخاب شده است. ابرپارامترهای استفاده‌شده در مدل یادگیری تقویتی عمیق در جدول ۱ ارائه شده‌اند. شایان توجه است که الگوریتم مذکور برای هر دو انتخاب زمان سبز و نوع فازبندی استفاده شده است.

به‌منظور ارزیابی عملکرد الگوریتم دوئل دوگانه در کنترل تطبیقی چراغ راهنمایی، برداشت مقادیر Q علاوه‌بر شبکه‌ی ۳DQN از یک شبکه‌ی DQN[۱۹] نیز انجام گرفته است. شبکه‌ی دوئل دوگانه، ترکیبی از الگوریتم شبکه‌ی عمیق Q دوگانه و دوئل است:

۱. مدل شبکه‌ی عمیق Q دوگانه برای رسیدن به مقدار Q هدف، مقدار بهینه‌ی اقدام را از طریق شبکه‌ی هدف به‌دست می‌آورد. به‌بیان دقیق‌تر، شبکه‌ی هدف روشی است که در شبکه‌های یادگیری عمیق Q برای تثبیت آموزش استفاده می‌شود. همان‌گونه که پیشتر اشاره شده است، در شبکه‌های ذکرشده، تابع مقدار Q با یک شبکه‌ی عصبی تقریب زده می‌شود. در حین آموزش، وزن‌های شبکه به گونه‌ای بروزرسانی می‌شوند که خطای بین مقدار Q پیش‌بینی‌شده و مقدار Q هدف، که با فرمول بلمن محاسبه می‌شود، کاهش یابد. اما از آنجا که برای تخمین مقادیر Q پیش‌بینی‌شده و هدف از یک شبکه استفاده می‌شود، لذا ممکن است باعث ناپایداری در آموزش شود؛ که برای حل مسئله‌ی مذکور، یک شبکه‌ی هدف متفاوت از شبکه‌ی اصلی در نظر گرفته می‌شود. شبکه‌ی هدف ساختار مشابه با شبکه‌ی اصلی دارد، ولی وزن آن کُندتر بروزرسانی می‌شود. در مطالعه‌ی حاضر، وزن شبکه‌ی هدف در هر ۱۰۰ بار آموزش بروزرسانی شده است. به این ترتیب، مقادیر Q هدف، که برای بروزرسانی شبکه‌ی اصلی استفاده می‌شوند، برای چندین دور، ثابت باقی می‌مانند، که می‌توانند به تثبیت آموزش کمک کنند.[۱۷]

۲. در یادگیری تقویت عمیق، معمولاً از یک شبکه‌ی عصبی برای تقریب تابع مقدار Q استفاده می‌شود، که پاداش موردانتظار را برای انجام یک اقدام در یک وضعیت مشخص تخمین می‌زند. شبکه‌ی دوئل، شبکه‌ی عصبی را به دو بخش تقسیم می‌کند: ۱) تخمین تابع مقدار وضعیت؛ ۲) تخمین تابع مزیت اقدام وابسته به وضعیت. تابع مقدار وضعیت، پاداش موردانتظار در آینده را برای قرارگرفتن در یک وضعیت معین تخمین می‌زند، در حالی که تابع مزیت اقدام، اهمیت نسبی هر اقدام را در آن وضعیت تخمین می‌زند. ارتباط تابع مزیت A با مقدار و توابع Q به صورت رابطه‌ی ۹ تعریف می‌شود:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \quad (9)$$

شایان ذکر است که رابطه‌ی ۱۰ نیز برقرار است:

$$E_{a \sim \pi(s)}[A^{\pi}(s,a)] = 0 \quad (10)$$

[19] Deep Q-Network

فایده‌ی روش اخیر آن است که شبکه می‌تواند بفهمد که کدام وضعیت‌ها ارزشمند هستند (دارای مقادیر Q بیشتر)، بدون آنکه نیاز باشد تأثیر هر اقدام را در هر وضعیت یاد بگیرد. لذا سبب می‌شود که سیاست بهتری در شرایطی که بسیاری از اقدام‌ها مقدار مشابه دارند، داشته باشد. شبکه‌ی مذکور، عملکرد بهتری نسبت به یک شبکه‌ی عصبی معمولی دارد.[۱۸]

بنابراین، به کمک الگوریتم یادگیری تقویتی عمیق دوئل دوگانه می‌توان ارزش اقدام‌های ممکن را در پایان گام‌های زمانی ۱۰ ثانیه‌ای و نیز در انتهای هر سیکل به‌دست آورد و اقدام با بیشترین ارزش را انتخاب کرد. تصمیم‌های اخیر به محیط شبیه‌سازی انتقال داده می‌شوند و بر وضعیت تقاطع چراغدار در گام‌های زمانی آینده اثرگذار خواهند بود.

۳. نتایج عددی

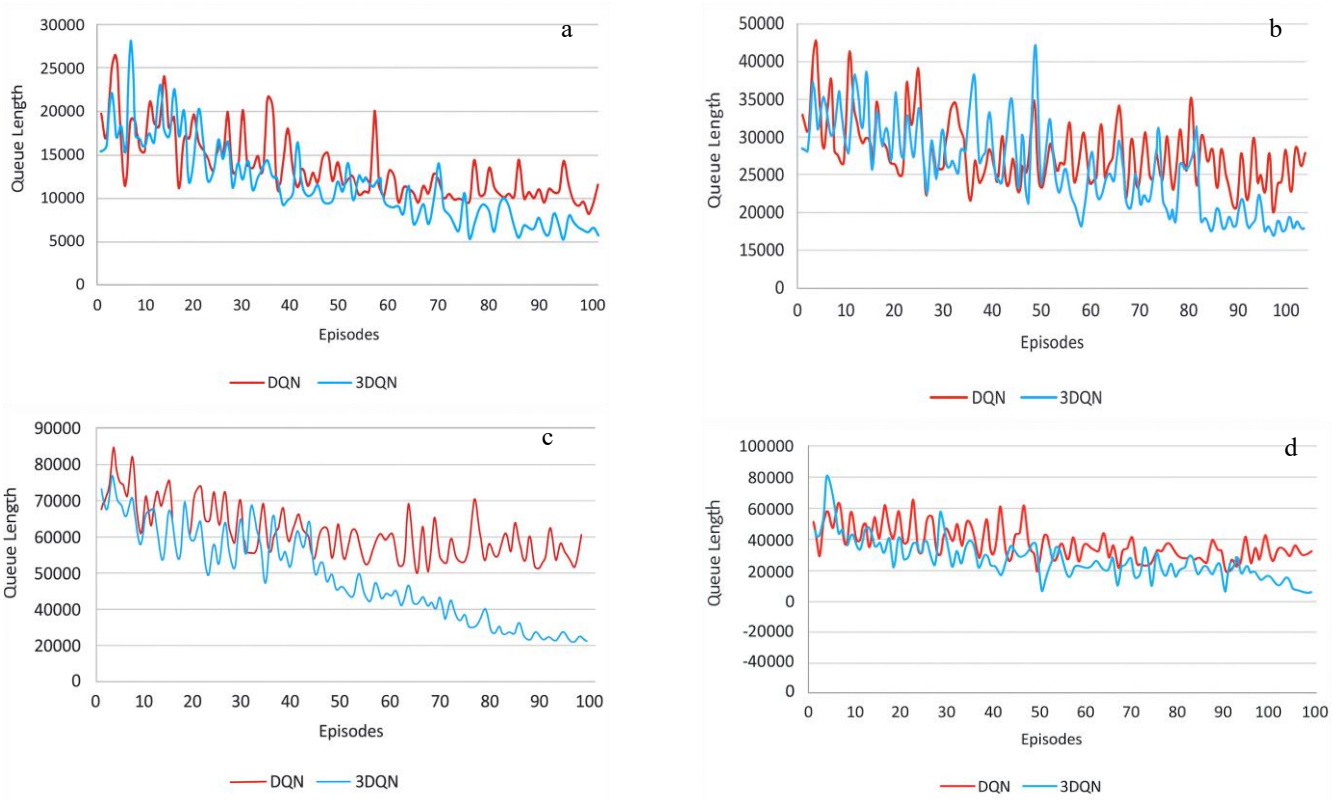
کنترل چراغ ترافیکی در یک تقاطع چراغدار به شیوه‌ی تطبیقی، به کمک یادگیری تقویتی عمیق و با دو رویکرد بررسی شده است. عملکرد کنترل‌کننده با شبکه‌ی دوئل دوگانه در با مقایسه نتایج شبیه‌سازی برای کنترل چراغ ترافیکی با یادگیری تقویتی عمیق ارزیابی شده است. در کنترل تطبیقی و هر دو رویکرد، کمینه‌ی زمان سبز ۱۰ و بیشینه‌ی آن، ۵۰ ثانیه لحاظ شده است. زمان زرد ۳ و مدت زمان تمام قرمز هم ۲ ثانیه لحاظ شده است. معیار استفاده‌شده نیز طول صف خودروها بوده است، که تعداد کل وسائط نقلیه‌ی متوقف‌شده در پایان هر شبیه‌سازی ۱ ساعته برای همه‌ی رویکردها بوده است. سرعت کمتر از ۰/۱ متر بر ثانیه، توقف در نظر گرفته شده است. همان گونه که در جدول ۲ ارائه شده است، طول صف میان دو نوع الگوریتم یادگیری تقویتی در حالت‌های مختلف توزیع و میزان تقاضا مقایسه شده است. در جدول ۳ نیز مقایسه‌ی میزان کاهش طول صف تجمعی در الگوریتم دوئل دوگانه نسبت به Q عمیق درحالت‌های مختلف توزیع و میزان تقاضا ارائه شده است. در شکل ۶، طول صف تجمعی خودروها در پایان هر ساعت شبیه‌سازی مشاهده می‌شود. هر اپیزود، نمایانگر یک ساعت از اجرای شبیه‌سازی و عبور خودروهاست. با افزایش تعداد اپیزودها، فرایند یادگیری بهبود یافته و مشاهده شده است که با حرکت به سمت انتخاب‌های مناسب‌تر، خروجی‌های بهتری از شبیه‌ساز حاصل می‌شود. این روند نشانگر تأثیر مثبت تجربه در عملکرد مدل در بهینه‌سازی ترافیک است. مطابق نتایج جداول و نمودارهای طول صف

جدول ۲. مقایسه‌ی طول صف تجمعی میان دو نوع الگوریتم در حالت‌های مختلف توزیع و میزان تقاضا.

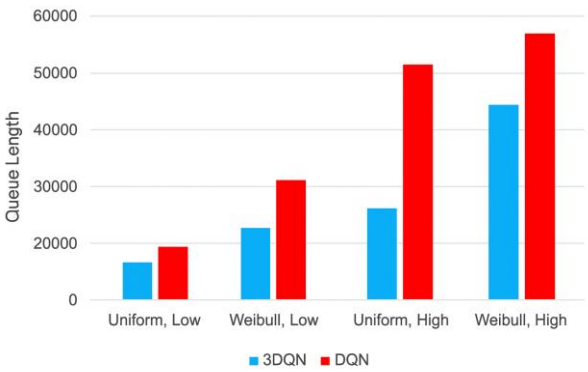
Cumulative Queue Length (Vehicles)				
Demand volume	Low		High	
Demand Distribution	Uniform	Weibull	Uniform	Weibull
DQN	9939	22940	49465	56459
3DQN	6597	12053	16001	41616

جدول ۳. مقایسه‌ی میزان کاهش طول صف تجمعی در الگوریتم دوئل دوگانه نسبت به Q عمیق درحالت‌های مختلف توزیع و میزان تقاضا.

Reduction in the cumulative queue length using the 3DQN				
Demand volume	Low		High	
Demand Distribution	Uniform	Weibull	Uniform	Weibull
Reduction (٪)	33.63	47.46	67.65	26.29



شکل ۶. مقایسه‌ی طول صف خودروها میان دو الگوریتم یادگیری تقویتی: (a) تقاضای یکنواخت و کم؛ (b) تقاضای متغیر و کم؛ (c) تقاضای یکنواخت و زیاد؛ (d) تقاضای متغیر و زیاد.



شکل ۷. مقایسه‌ی طول صف تجمعی خودروها میان دو الگوریتم یادگیری تقویتی عمیق در حالت‌های مختلف توزیع و میزان تقاضا.

خودروها در شکل ۶ مشخص است که کنترل تطبیقی به روش فازبندی منعطف به کمک الگوریتم دوئل دوگانه نتایج بهتری نسبت به کنترل تطبیقی با فازبندی منعطف در حالت یادگیری تقویتی DQN نشان داده است. دیگر معیارهای ارزیابی نیز روند مشابهی را نشان داده‌اند. این تذکر لازم است که در بررسی نمودارها و استخراج نتایج، میانگین اعداد ۵٪ پایانی مراحل به‌عنوان نتیجه‌ی یادگیری عامل تقویتی فرض شده‌اند. علاوه‌بر این، در شکل ۷، طول صف خودروها در صورتی که از الگوریتم دوئل دوگانه و یادگیری تقویتی عمیق ساده استفاده شود، در حالت‌های مختلف توزیع و میزان تقاضا مقایسه شده است. در شکل ۷، تأثیر کنترل‌کننده‌ی چراغ ترافیکی با استفاده از الگوریتم یادگیری تقویتی دوئل دوگانه در مقایسه با نتایج شبیه‌سازی کنترل‌کننده‌ی یادگیری تقویتی عمیق Q نشان داده شده است. پس از تحلیل طول صف تجمعی وسائط نقلیه در تقاطع ذکرشده، مشاهده شده است که مقدار پارامتر ترافیکی مذکور با استفاده از کنترل‌کننده‌ی پیشنهادی نسبت به یادگیری تقویتی عمیق Q کاهش یافته است.

۴. نتیجه‌گیری

با استفاده از کنترل تطبیقی چراغ راهنمایی، که یکی از روش‌های پیشرفته در مدیریت ترافیک شهری به شمار می‌رود، می‌توان به‌طور قابل ملاحظه‌ای زمان انتظار خودروها در تقاطع‌های چراغ‌دار را کاهش داد. روش مذکور با بهره‌گیری از الگوریتم‌های بهینه‌سازی، زمان‌های چراغ سبز را براساس شرایط ترافیکی لحظه‌ای بهینه می‌سازد. مطالعه‌ی حاضر، بر نرخ جریان حرکت‌های گردشی تمرکز داشته و با استفاده از یادگیری تقویتی، زمان سبز بهینه برای هر فاز در یک تقاطع چهارراه را تعیین کرده است. علاوه‌بر این، برای کاهش تداخل حرکت‌های گردشی خودروها با جریان‌های مخالف، نوعی فازبندی منعطف در نظر گرفته شده است، که به یک عامل دیگر یادگیری تقویتی اجازه می‌دهد مناسب‌ترین نوع فازبندی را در هر سیکل انتخاب کند. بنابراین، برای کاهش بیشتر زمان انتظار خودروها، تعداد فازهای هر سیکل می‌تواند از کمینه‌ی ۲ تا بیشینه‌ی ۴ فاز متفاوت باشد. در نتیجه، ترکیب مناسبی از فازها در هر سیکل به‌دست می‌آید. با روش اخیر، کنترل تقاطع، بسته به زمان روز و سطح فعالیت ترافیک در هر تقاطع، تغییر می‌کند و هدف آن، سازگارشدن با شرایط متغیر ترافیک است.

در نوشتار حاضر، دو نوع الگوریتم یادگیری تقویتی عمیق پیشنهاد شده است. شبکه‌ی Q عمیق استاندارد و دوئل دوگانه، که ترکیبی از الگوریتم‌های شبکه‌ی عمیق Q دوگانه و دوئل است. شبکه‌ی ذکرشده از یک شبکه‌ی هدف برای تثبیت آموزش و کاهش خطای بین مقادیر Q پیش‌بینی‌شده و هدف استفاده می‌کند. شبکه‌ی هدف، که وزن آن کُندتر بروزرسانی می‌شود، مقادیر Q هدف را برای چند دور ثابت نگه می‌دارد. از سوی دیگر، شبکه‌ی دوئل شبکه‌ی عصبی را به دو بخش تقسیم می‌کند: یکی برای تخمین تابع مقدار وضعیت و دیگری برای تخمین تابع مزیت اقدام. دو تابع اخیر، پاداش موردانتظار در آینده و اهمیت نسبی هر اقدام را تخمین می‌زنند. به‌منظور ارزیابی عملکرد الگوریتم دوئل دوگانه، برای برداشت مقادیر Q از یک شبکه‌ی DQN نیز استفاده شده است. الگوریتم‌های پیشنهادی برای رسیدن به اهداف پژوهش حاضر در محیط برنامه‌نویسی پایتون کد شده‌اند. از طریق ریزشبیه‌سازی SUMO که با محیط برنامه‌نویسی در تعامل مکرر است، نشان داده شده است که کنترل‌کننده‌ی پیشنهادی که از شبکه‌ی دوئل دوگانه بهره می‌برد، کارایی ترافیک را از نظر معیار طول صف تجمعی، بیشتر از کنترل‌کننده با شبکه‌ی عصبی رایج بهبود بخشیده است.

با توجه به نتایج به‌دست‌آمده، اگر حجم خودروها کم باشد و جریان به‌صورت متغیر عبور کند، استفاده از شبکه‌ی دوئل دوگانه مؤثرتر از توزیع یکنواخت است. در حجم بالای خودروها، الگوریتم مذکور در حالت جریان یکنواخت عملکرد بهتری دارد. همچنین بررسی میزان کاهش طول صف تجمعی در دو نوع توزیع یکنواخت و متغیر نشان می‌دهد که در حالت جریان یکنواخت، وقتی حجم خودروها زیاد باشد، کاهش بیشتری وجود دارد. در حالی که در توزیع متغیر، هر چه حجم عبوری کمتر باشد، عملکرد الگوریتم دوئل دوگانه بهتر می‌شود.

در مطالعه‌ی حاضر، بهینه‌سازی فقط برای یک معیار، یعنی طول صف تجمعی، انجام شده است. با این حال، امکان گنجاندن هر معیار قابل‌اندازه‌گیری دیگری در فرآیند بهینه‌سازی با استفاده از روش‌های ذکرشده وجود دارد. به‌عبارت دیگر، می‌توان پاداش را به‌صورت جمع وزنی چند معیار تعریف کرد، تا عامل یادگیری تقویتی بتواند تصمیم‌گیری‌های بهتری انجام دهد. یکی دیگر از بهبودها می‌تواند درنظرگرفتن منطقه‌ی پارکینگ نزدیک به تقاطع باشد. همچنین، تأثیر عبور عابران پیاده نیز می‌تواند در انتخاب نوع فازبندی لحاظ شود. در مطالعات آتی، می‌توان به جای تمرکز بر یک تقاطع منفرد، شبکه‌ای از تقاطع‌ها را بررسی کرد و آثار متقابل آن‌ها را نیز در نظر گرفت.

References- منابع

1. Mannion, P., Duggan, J., and Howley, E., 2016. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. *Autonomic Road Transport Support Systems*, pp.47-66, doi: 10.1007/978-3-319-25808-9_4.

2. Eriksen, A. B., Lahrmann, H., Larsen, K. G., and Taankvist, J. H., 2020. Controlling signalized intersections using machine learning. *Transportation Research Procedia*, *48*, pp.987-997, doi: 10.1016/j.trpro.2020.08.127.

3. Mousavi, S. S., Schukat, M., and Howley, E., 2017. Traffic light control using deep policy-gradient and value-function based reinforcement learning. *IET Intelligent Transportation Systems*, *11*, pp.417-423. doi: 10.1049/iet-its.2017.0153.

4. Miletić, M., Ivanjko, E., Gregurić, M., and Kušić, K., 2022. A review of reinforcement learning applications in adaptive traffic signal control. *IET Intelligent Transportation Systems*, *16*, pp.1269-1285. doi: 10.1049/itr2.12208.

5. Chin, Y.K., Lee, L.K., Bolong, N., Yang, S.S. and Teo, K.T.K., 2011. Exploring Q-learning optimization in traffic signal timing plan management. In *Proceedings of the 3rd International Conference on Computational Intelligence, Communication Systems and Networks*, Bali, Indonesia, pp.269-274, doi: 10.1109/CICSyN.2011.64.

6. Haydari, A., and Yilmaz, Y., 2020. Deep reinforcement learning for intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), pp.11-32. doi: 10.1109/TITS.2020.3008612.

7. Sutton, R. S., and Barto, A. G., 2018. Introduction. In *Reinforcement learning: an introduction* (2nd ed., ch. 1, sec. 1.1, pp. 1-2). Cambridge, Massachusetts (London, England): The MIT Press.

8. Touhbi, S., Babram, M.A., Nguyen-Huu, T., Marilleaub, N., Hbid, M.L., Cambier, C. and Stinckwich, S., 2017. Adaptive traffic signal control exploring reward definition for reinforcement learning. In *Proceedings of the 8th International Conference on Ambient Systems, Networks and Technologies*, Madeira, Portugal, pp.513-520, doi: 10.1016/j.procs.2017.05.327.

9. Zheng, G., Zang, X., Xu, N., Wei, H., Yu, Z., Gayah, V., Xu, K. and Li, Z., 2019. Diagnosing reinforcement learning for traffic signal control, arXiv preprint arXiv: 1905.04716.

10. La, P. and S. Bhatnagar, 2011. Reinforcement learning with function approximation for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2): pp.412-421. doi: 10.1109/TITS.2010.2076327.

11. Genders, W. and Razavi, S., 2016. Using a deep reinforcement learning agent for traffic signal control. arXiv preprint arXiv: arXiv:1611.01142.

12. Wei, H., Zheng, G., Yao, H. and Li, Z., 2018. IntelliLight: a reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, United Kingdom, pp.2496-2505, doi: 10.1145/3219819.3220096.

13. Zheng, Q., Xu, H., Chen, J., Zhang, D., Zhang, K., & Tang, G., 2022. Double deep Q-network with dynamic bootstrapping for real-time isolated signal control: a traffic engineering perspective. *Applied Sciences*, *12*(17), p.8641. doi: 10.3390/app1217864.

14. Stamatiadis, N., Tate, S., and Kirk, A., 2016. Left-turn phasing decisions based on conflict analysis. *Transportation Research Procedia*, *14*, 3390-3398. doi: 10.1016/j.trpro.2016.05.291.

15. Han, G., Zheng, Q., Liao, L., Tang, P., Li, Z., and Zhu, Y., 2022. Deep reinforcement learning for intersection signal control considering pedestrian behavior. *Journal of Electronics*, *11*(21), p.3519. doi: 10.3390/electronics11213519.

16. Afandizadeh, Sh., Tavakoli Kashani, A. & Hasanpour, Sh., 2014. A model for safety prioritization of at-grade intersections. *Rahvar Research Studies*, *9*(3), pp.111–138, doi: noo.rs/PBTM2. [in Persian]

17. van Hasselt, H., Guez, A. and Silver, D. 2016., Deep reinforcement learning with double Q-learning, In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, doi: 10.1609/aaai.v30i1.10295.

18. Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M. and de Freitas, N., 2016. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, pp.1995-2003, arXiv preprint arXiv: 10.48550/arXiv.1511.06581