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

الهام گلپایگانی[۱]، عباس بابازاده[۲*]، امید نیری[۱]

۱- کارشناسی ارشد حمل‌ونقل، دانشکده مهندسی عمران، دانشکدگان فنی، دانشگاه تهران

۲- دانشیار، دانشکده مهندسی عمران، دانشکدگان فنی، دانشگاه تهران

پست الکترونیکی نویسندگان:

۱- elham.golpayegani@ut.ac.ir

۲- ababazadeh@ut.ac.ir

چکیده:

این مقاله عملکرد دو روش یادگیری تقویتی، شبکه Q عمیق دوئل دوگانه و شبکه Q عمیق استاندارد، را در کنترل تطبیقی فاز گردش به چپ چراغ‌های راهنمایی در یک تقاطع شهری مقایسه می‌کند. این روش‌های مقدار- محور، با بهره‌گیری از بهینه‌سازی در یادگیری تقویتی، مدت زمان سبز هر فاز را تعیین و یکی از دو فاز گردش به چپ محافظت شده یا مجاز را برای سیکل بعدی انتخاب می‌کند. شبیه‌سازی‌ها برای حالات توزیع یکنواخت و متغیر جریان خودروها و با دو جریان ترافیک سبک و سنگین انجام می‌شوند. نتایج نشان می‌دهند که الگوریتم شبکه عمیق دوئل دوگانه در فرایند یادگیری موثرتر از الگوریتم شبکه Q استاندارد عمل می‌کند. همچنین، یادگیری با شبکه Q دوئل دوگانه می‌تواند طول صف تجمعی وسایل نقلیه را در تمام حالات شبیه‌سازی حداقل به میزان ۲۶ درصد کاهش داده و جریان ترافیک را بهبود بخشد. این کاهش در حالت جریان ترافیک سنگین و یکنواخت بیشتر از سایر حالات بوده و به ۶۷ درصد می‌رسد. این تحقیق می‌تواند نقش مهمی در توسعه سیستم‌های هوشمند کنترل ترافیک ایفا کند.

واژگان کلیدی:

کنترل تطبیقی چراغ ترافیکی، فاز گردش به چپ، یادگیری تقویتی، شبکه عمیق دوئل دوگانه.

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

E. Golpayegani[۱], A. Babazadeh[۲], O. Nayeri[۱]

۱- **MSc, School of Civil Engineering, College of Engineering, University of Tehran, Tehran, Iran**

۲- **Associate Professor, School of Civil Engineering, College of Engineering, University of Tehran,** Tehran, Iran

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 (۳DQN) 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 ۳DQN 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 ۳DQN algorithm outperform DQN algorithm. The results also reveal that the ۳DQN algorithm can reduce cumulative vehicle queue lengths by at least ۲۶٪ in all cases, and up to ۶۷٪ 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.

Keywords: adaptive traffic light control, left-turn phase, reinforcement learning, double dueling deep Q-network

۱ ـ مقدمه و مرور ادبیات

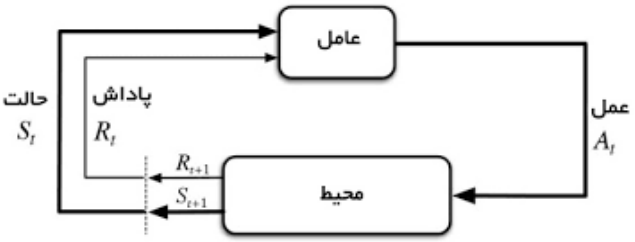
به علت محدودیت فضاهای شهری و نبود سرمایه کافی برای ایجاد زیرساخت‌های جدید، کنترل و استفاده بهینه از تسهیلات موجود ضروری می‌نماید (مانیون و همکاران، ۲۰۱۶). یکی از نقاط مهم برای کنترل جریان ترافیک، تقاطعات چراغدار هستند. در این نقاط به دلیل افزایش مصرف سوخت و تأخیرهای صورت گرفته پتانسیل مناسبی برای کنترل و بهبود جریان ترافیک وجود دارد و ابزار کنترل مناسبی که همان چراغ ترافیکی است نیز موجود است (اریکسن و همکاران، ۲۰۲۰). به طور کلی سه نوع کنترل چراغ ترافیکی وجود دارد. در نوع از پیش زمان‌بندی شده[۱] (زمان‌بندی ثابت) بر اساس 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 حلقه القایی[۳] در تقاطع دریافت می‌کند، تصمیم می‌گیرد که زمان سبز یک فاز را ادامه یا آن را به فاز بعدی اختصاص دهد (موسوی و همکاران، ۲۰۱۷). در واقع کنترل کننده فعال با یک برنامه پایه که برای همه فازها دارای حداقل و حداکثری از زمان سبز است، کار می‌کند. روش کار بر پایه یک قاعده است که با توجه به درخواست تقاضا، بر مدت زمان سبز در فاز فعلی می‌افزاید. درخواست‌ها برای فاز فعلی توسط وسایل نقلیه‌ای که به تقاطع نزدیک می‌شوند و برای فازهای مقابل با انتظار برای وسایل نقلیه در رویکردهای دیگر ایجاد می‌شوند (میلتیچ و همکاران، ۲۰۲۲). در نوع سوم نیز که کنترل تطبیقی[۴] نام دارد، زمان‌بندی به طور خودکار با توجه به وضعیت فعلی تقاطع، مدیریت و به‌روز می‌شود (موسوی و همکاران، ۲۰۱۷). ویژگی اصلی این نوع کنترل، بهینه‌سازی زمان‌بندی چراغ با توجه به تابع هدفِ تعیین شده است. در این جا نظارت و پایش جریان خودروها در رویکردهای تقاطع مانند کنترل فعال صورت می‌گیرد. تفاوت اصلی این دو نوع کنترل این است که کنترل تطبیقی، هم داده‌های ترافیکی در لحظه[۵] و هم پیش‌بینی تغییرات احتمالی در جریان ترافیک را در نظر می‌گیرد (میلتیچ و همکاران، ۲۰۲۲). روش کنترل تطبیقی با حل یک مسئله بهینه‌سازی به کمک هوش مصنوعی و الگوریتم‌های یادگیری ماشین، نتایج امیدوارکننده‌ای داشته است (موسوی، شوکت و هاولی، ۲۰۱۷). از جمله این روش‌ها الگوریتم ژنتیک[۶] یا الگوریتم تکاملی[۷]، منطق فازی[۸] و یادگیری تقویتی[۹] است (چین و همکاران، ۲۰۱۱). این مطالعه که بر کنترل تطبیقی چراغ ترافیکی متمرکز است، به کمک یادگیری تقویتی، مدت زمان سبز بهینه هر فاز را در یک تقاطع چهارراه تعیین می‌کند.

سیستم‌های هوشمند حمل‌ونقل در کنار هوش مصنوعی می‌تواند راه‌حل‌های مناسبی برای چالش‌های قرن حاضر ارائه دهد. 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 تقاطع‌ها است که پرطرفدارترین آن‌ها کنترل تطبیقی چراغ راهنمایی در تقاطع‌ها است (حیدری و یلماز، ۲۰۲۰). یادگیری تقویتی یک رویکرد محاسباتی برای یادگیری و تصمیم‌گیری بر پایه هدف است و بر خلاف دیگر رویکردهای یادگیری، به نظارت و یا مدل‌های کامل از محیط[۱۰] نیاز ندارد (ساتن و بارتو، ۲۰۱۸). در یادگیری تقویتی یک عامل[۱۱] می‌کوشد با تعامل با محیط پیرامون و به کمک سعی و خطا، در هر گام به بهترین اقدام[۱۲] ممکن دست یابد. به این صورت که عامل یک وضعیت[۱۳] از محیط را مشاهده می‌کند و بر اساس آن یک اقدام انجام می‌دهد که یک پاداش[۱۴] در بردارد. ساختار مدل یادگیری تقویتی در شکل ۱ قابل مشاهده است. در واقع عامل در تلاش است تا یک سیاست[۱۵] کنترل بهینه بیابد تا پاداش تجمعی از

---

۱ Pre-timed signal control
۲ Actuated signal control
۳ Inductive loop detector
۴ Adaptive signal control
۵ Real-time
۶ Evolutionary algorithm
۷ Evolutionary algorithm
۸ Fuzzy logic
۹ Reinforcement Learning
۱۰ Environment
۱۱ Agent
۱۲ Action
۱۳ State
۱۴ Reward
۱۵ Policy

طریق تعامل مکرر با محیط به حداکثر برسد. این پاداش مورد انتظار به ازای هر جفت وضعیت - اقدام[1]، مقدار Q[2] نامیده می‌شود (موسوی . همکاران ، ۲۰۱۷). 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[3] هستند. همچنین باید در نظر داشت در یادگیری تقویتی، وضعیت به عنوان ورودی برای سیاست و تابع مقدار در نظر گرفته می‌شود (ساتن و بارتو، ۲۰۱۸).



شکل ۱: ساختار مدل یادگیری تقویتی

پژوهشگران مسئله کنترل تطبیقی چراغ به کمک یادگیری تقویتی را از جهات مختلف مورد توجه قرار داده‌اند. هرچند تعریف مسئله در این مطالعات یکسان است، اما تعریف وضعیت‌ها، اقدامات، پاداش، سیاست‌ها، محیط، توابع مقدار و همچنین معیارهای بهینه‌سازی در هر یک متفاوت است. به علاوه از شبیه‌سازهای گوناگونی نیز برای ایجاد شرایط چهارراه با توزیع‌های متنوع بهره گرفته شده است.

موسوی و همکاران (۲۰۱۷) با توجه به نتایج امیدوارکننده‌ای که ترکیب ساختار شبکه عصبی عمیق و روش‌های یادگیری تقویتی داشته است، دو نوع الگوریتم یادگیری تقویتی ارائه کرده‌اند که یکی بر پایه گرادیان سیاست عمیق و دیگری مبتنی بر تابع مقدار می‌باشد. عامل، یک تصویر لحظه‌ای از وضعیت فعلی شبیه‌ساز ترافیکی دریافت می‌کند و بر پایه گرادیان سیاست عمیق، توسط مشاهدات خود کنترل بهینه چراغ را مستقیما می‌یابد؛ اما عامل مبتنی بر تابع مقدار ابتدا مقادیر اقدامات ممکن برای کنترل چراغ را تخمین زده و سپس اقدام با بیشترین مقدار را انتخاب می‌کند. روش‌های یادگیری عمیق یکی از بهترین راهکارها برای مسائلی است که داده‌ها دارای ابعاد زیادی بوده و استخراج ویژگی‌های مورد نظر دشوار است. پس از ورود به یک شبکه عمیق، خروجی داده‌ها با گذر از هر لایه به عنوان ورودی لایه بعد در نظر گرفته شده و ویژگی‌های غیرخطی آن‌ها هر بار تغییر شکل می‌دهد. یک شبکه عمیق یادگیری Q از مشاهدات عامل استفاده کرده و آن را برای یادگیری سیاست کنترل بهینه به کار می‌برد. در واقع روش شبکه یادگیری عمیق Q شبکه عصبی عمیق را با یادگیری Q (یک الگوریتم بدون مدل در یادگیری تقویتی) تلفیق می‌کند تا بتواند تابع اقدام - مقدار و در نتیجه سیاست مورد نظر عامل را بیابد. تفاوت دو رویکرد مبتنی بر تابع مقدار (مقدار-محور) و بر پایه سیاست (سیاست-محور) در آخرین لایه شبکه یادگیری عمیق Q قابل بررسی است. به این صورت که در رویکرد مبتنی بر تابع مقدار آخرین لایه شبکه یادگیری عمیق Q یک لایه خطی کاملا متصل با تعدادی نورون خروجی (مقدار Q) متناظر به هر اقدام است. در حالی که در مدل دوم، لایه آخر دو مجموعه خروجی را نشان می‌دهد، یکی که منجر به توزیع احتمال بر روی اقدامات (یعنی همان سیاست) شده و دیگری یک گره خطی منفرد که منجر به تخمین تابع وضعیت - مقدار می‌شود. روش‌های گرادیان سیاست می‌کوشد یک تابع سیاست پارامتری شده را با روش گرادیان نزولی بهینه کند. به این صورت که مستقیما به دنبال یادگیری سیاست بهینه در فضای سیاست‌ها است.

به علت شرایط دشوار برداشت داده در محیط دنیای واقعی، کاربردهای یادگیری تقویتی برای کنترل چراغ ترافیکی در شبیه‌ساز اجرا می‌شود. SUMO [4] محبوب‌ترین شبیه‌ساز ترافیکی منبع باز است. این شبیه‌ساز با کتابخانه TraCI[5] در پایتون امکان تعامل کاربر با محیط را فراهم می‌کند (حیدری و یلماز، ۲۰۲۰).

تعریف وضعیت از اهمیت بالایی برخوردار است و به دستگاه‌های تجهیزات ترافیکی گوناگون مانند دوربین و آشکارساز حلقه‌ای

[1] State - action
[2] Q-Value
[3] Value Function
[4] Simulation of Urban MObility
[5] Traffic Control Interface

بستگی زیادی دارد (حیدری و یلماز، ۲۰۲۰). در مطالعه موسوی و همکاران (۲۰۱۷) نمایش وضعیت‌ها به صورت تصویر، ورودی به یک شبکه عصبی پیچشی است. در این صورت با ذخیره و سپس تغذیه لایه‌های شبکه عمیق پیچشی با تصاویری از مشاهدات متوالی، نه تنها 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 اقدامات ممکن، اختصاص فاز سبز به معبر شمالی - جنوبی یا شرقی - غربی است. این در حالی است که در مقاله اریکسن و همکاران (اریکسن و همکاران، ۲۰۲۰) در هر گام زمانی یک ثانیه‌ای تصمیم‌گیری می‌شود که زمان سبز ادامه یابد یا به جهت دیگر داده شود. در پژوهش توحبی و همکاران (۲۰۱۷) تعریف جدیدی برای وضعیت‌ها معرفی شد. در بیشتر مطالعات حداکثر طول صف در هر فاز برای تعریف وضعیت به کار می‌رود؛ زیرا تعداد خودروها در یک صف نمایش قدرتمندتری از وضعیت ترافیکی نسبت به دیگر موارد دارد. اما اگر ابعاد وسایل نقلیه نیز در نظر گرفته نشود، یادگیری می‌تواند دچار مشکل شود. برای نمونه، در یک فاصله ۱۰ متری از تقاطع ممکن است دو خودروی سواری یا یک خودروی سنگین وجود داشته باشد که در صورت فرض بالا با وجود عملکرد متفاوت در محیط تقاطع، دارای طول صف یکسان خواهند بود. این مطالعه از مفهومی به نام صف باقی مانده[1] بهره برده است که بیانگر طول صف در یک خط عبور تقسیم بر طول خط عبور است. نتایج این تحقیق نشان می‌دهند که تعریف جدید برای وضعیت از جهت توان عملیاتی تقاطع و متوسط تأخیر، به ویژه در حجم بالای وسائل نقلیه و شرایط متغیر ترافیکی عملکرد خوبی داشته است. پژوهش ژنگ و همکاران (۲۰۱۹) به بررسی این نکته پرداخته است که آیا برای تعریف وضعیت‌ها لزومی به کاربرد نمایش‌های پیچیده مانند تصاویر وجود دارد یا خیر. چه بسا که این موضوع ممکن است باعث کند شدن فرایند آموزش در یادگیری تقویتی شود، در حالی که ثمره قابل توجهی نیز نداشته باشد. این مقاله با برقراری ارتباط بین یادگیری تقویتی و رویکردهای معمول حمل‌ونقلی نشان داد که با فرض جریان یکنواخت ترافیک، اولا طول صف به عنوان پاداش در یادگیری تقویتی معادل بهینه‌سازی زمان سفر در روش‌های حمل‌ونقلی است؛ دوم این که تعداد خودروها در هر خط عبور به همراه اطلاع از فاز فعلی چراغ پویایی سیستم را به طور کامل شرح می‌دهد. با داشتن این دو ویژگی، عامل می‌تواند به سیاست کنترل بهینه دست یابد.

در مطالعه‌ای که توسط (لا و بهاتناگار، ۲۰۱۱) انجام شده است، برای اولین بار مسئله کنترل تطبیقی چراغ به کمک الگوریتم یادگیری تقویتی با تقریب تابع بررسی شده است. الگوریتم یادگیری Q یکی از مهم‌ترین الگوریتم‌های یادگیری تقویتی است که به سیاست بهینه همگرا می‌شود. همچنین ابعاد گسترده فضای وضعیت - اقدام در مسئله کنترل چراغ موجب به‌کارگیری روش‌های تقریب تابع برای کارایی محاسباتی شده است. جندرز و رضوی در (جندرز و رضوی، ۲۰۱۶) از یکی از انواع تقریب‌گرهای توابع استفاده کردند که در آن تابع اقدام - مقدار به عنوان یک شبکه 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؛ به خصوص که بازخورد به عامل برای یک یادگیری پایدار و سریع به همین تابع بستگی دارد. پاداش می‌تواند جمع وزنی از ویژگی‌های خطوط عبوری ورودی به تقاطع باشد. مشخصاتی شامل طول صف، تأخیر، زمان انتظار به‌روز شده، مقادیر صفر و یکی برای فاز فعلی چراغ، تعداد خودروهای عبوری از تقاطع و کل زمان سفر وسایل نقلیه از آن جمله‌اند (وی و همکاران، ۲۰۱۸). معمولاً هدف مسئله کنترل چراغ، به حداقل رساندن تأخیر خودروها در تقاطع است. به جای تأخیر، تابع پاداش در یک مطالعه، تعداد خودروهای خروجی از تقاطع بین دو گام زمانی متوالی تعریف شد که هم در زمان آموزش مدل بهره‌گیری از آن آسان است و هم برخلاف توابع پاداش استفاده شده در بیشتر مطالعات، می‌تواند در کوتاه‌مدت نتایج مناسبی برای تشویق یا تنبیه رفتار عامل ارائه دهد. جالب توجه است که این تابع بر اساس تأخیر نیست اما می‌تواند به تدریج موجب کاهش تأخیر متوسط خودروها گردد (ژنگ و همکاران، ۲۰۲۲). با تحلیل تعاریف مختلف پاداش در یادگیری

[1] Residual queue

تقویتی می‌توان دریافت که عملکرد توابع پاداش به حجم وسائل نقلیه در تقاطع، وسائل نقلیه مجهز به GPS و نیز تجهیزات مورد استفاده برای پایش تقاطع وابسته است. می‌توان تصور کرد که برداشت مواردی مانند تأخیر تجمعی از محیط به ادوات پیشرفته‌تری مانند نظارت تصویری یا وسایل نقلیه مجهز به GPS نیاز دارد (توحبی و همکاران، ۲۰۱۷).

تقاطع‌های چراغ‌دار می‌کوشند فضا و زمان را به گونه‌ای به حرکات مختلف ترافیکی تخصیص دهند که ایمنی و کارآمدی تأمین شود (استاماتیادیس و همکاران، ۲۰۱۶). پژوهش هان و همکاران (۲۰۲۲) با در نظر گرفتن مسائل مربوط به عابران پیاده در تقاطع‌ها، روش کنترل به کمک یادگیری تقویتی را با جنبه‌های ایمنی ترکیب کرده است. در این روش، زمان انتظار عابران پیاده و نیز خودروهای عبوری از تقاطع به طور هم‌زمان در نظر گرفته شده است. در حالی که گاهی اوقات، اندرکنش بین وسایل نقلیه گردش به چپ و جریان ترافیک مخالف می‌تواند خطرات ایمنی ایجاد کند. بنابراین، برای افزایش ایمنی و کارایی تقاطع‌های دارای چراغ راهنمایی، انتخاب فازهای مناسب برای به حداقل رساندن تأخیرهای تقاطع بسیار مهم است (استاماتیادیس و همکاران، ۲۰۱۶). در پژوهشی که به بررسی عوامل مؤثر بر ایمنی تقاطع‌ها پرداخته است، میزان اهمیت این عوامل با استفاده از روش تحلیل سلسله‌مراتبی تعیین شد. عوامل مورد بررسی شامل عوامل انسانی، زیرساختی، تسهیلات و تجهیزات، و عوامل ترافیکی بودند. بر اساس یافته‌های این پژوهش، عوامل تأثیرگذار تسهیلات و تجهیزات در زمینه تجهیزات کنترل ترافیک، نوع کنترل چراغ راهنمایی، فازبندی چراغ راهنمایی، نمایشگرهای عابران پیاده و وسایل نقلیه، خط‌کشی‌ها، تابلوها و روشنایی هستند. در این میان، نوع کنترل چراغ بیشترین اهمیت را دارد و در مرتبه اول قرار می‌گیرد، فازبندی چراغ در مرتبه دوم و نمایشگرهای چراغ در مرتبه آخر قرار دارند. همچنین، با بررسی میزان اهمیت عوامل ترافیکی مشخص شد که حجم گردش به چپ در مرتبه اول، ترکیب تردد در مرتبه دوم، جریان ترافیک در مرتبه سوم و حجم گردش به راست در مرتبه آخر قرار دارد؛ بنابراین، به‌کارگیری و طراحی مناسب چراغ‌های راهنمایی به‌عنوان یکی از عوامل اصلی در بهبود ایمنی و کارایی تقاطع‌های چراغ‌دار شناخته شد (افندی زاده و همکاران، ۱۳۹۳).

مسئله‌ای که در این مطالعه مورد بررسی قرار گرفته، عملکرد شبکه دوئل دوگانه در یادگیری تقویتی عمیق برای کنترل تطبیقی چراغ ترافیکی در یک تقاطع چهارراهی در در سیکل‌های متوالی چراغ است. هر ورودی تقاطع دارای سه خط عبوری است و فرض بر این است که جریان ورودی به تقاطع تحت تأثیر جریان بالادست قرار ندارد. عامل کنترل‌کننده چراغ در هر گام، وضعیت محیط (تقاطع چراغ‌دار) را مشاهده می‌کند. این وضعیت می‌تواند شامل پارامترهای مختلف ترافیکی باشد. با هدف کاهش تداخل حرکات گردشی خودروها با جریان‌های مخالف، یک روش فازبندی منعطف ارائه شده است که در هر سیکل توسط یک عامل دیگر یادگیری تقویتی تعیین می‌شود. این روش امکان تغییر تعداد فازهای هر سیکل را بسته به وضعیت ترافیک فراهم می‌کند و از حداقل دو تا حداکثر چهار فاز مختلف استفاده می‌کند. به این ترتیب، در هر سیکل، ترکیب بهینه فازها برای کاهش زمان انتظار خودروها انتخاب می‌شود. هدف از این مطالعه، دستیابی به یک الگوریتم مناسب در حوزه یادگیری تقویتی عمیق برای کنترل تطبیقی چراغ راهنمایی با بهره‌گیری از نوآوری پیشنهادی (منعطف‌سازی فازبندی گردش به چپ) است. این مقاله به شناسایی ترکیب‌های بهینه فازبندی چراغ‌های ترافیکی می‌پردازد. اگرچه هدف اصلی آن افزایش ایمنی نیست، اما توجه ویژه‌ای به تداخلات ناشی از گردش به چپ وسایل نقلیه و حرکات روبه‌رویی دارد. علاوه بر این، تحقیقات گذشته عمدتاً بر ارزیابی بهبودهای مربوط به حرکات گردش به چپ بر اساس داده‌های تاریخی تصادف تمرکز داشته‌اند که این داده‌ها اطلاعات پیش‌بینی‌کننده‌ای ارائه نمی‌دهند. بنابراین، ایجاد ابزارهایی که بتوانند تعادل مناسبی بین ایمنی و کارایی تقاطع برقرار کنند و تعیین بهترین فاز گردش به چپ را تسهیل نمایند، از اهمیت بالایی برخوردار است.

۲ ــ مدل یادگیری تقویتی در مسئله کنترل تطبیقی چراغ راهنمایی

در این مطالعه، نوعی فازبندی منعطف برای چراغ‌های ترافیکی در نظر گرفته شده است. گام زمانی ده ثانیه‌ای در نظر گرفته می‌شود و مدت زمان سبز هر فاز می‌تواند شامل چندین گام زمانی باشد. ترتیب فازها در هر سیکل ثابت است. حرکات گردش به چپ برای ورودی‌های مقابل معابر (شمالی-جنوبی یا شرقی-غربی) یا به

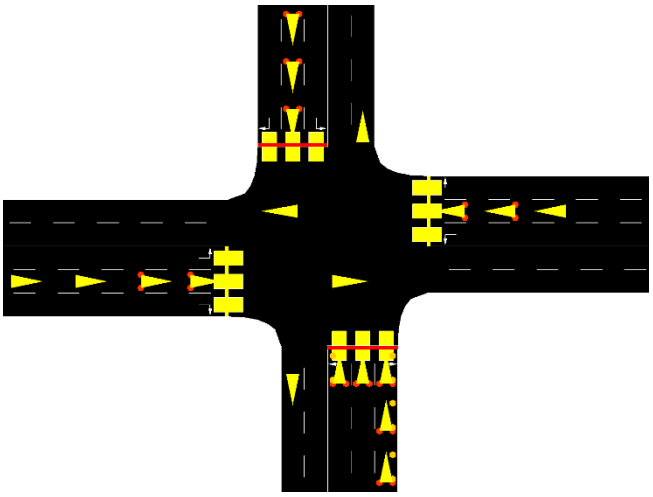
صورت مجاز در فاز مشترک با حرکات مستقیم و گردش به راست انجام می‌شوند، یا به صورت محافظت شده در یک فاز مجزا و از خط اختصاصی گردش به چپ صورت می‌گیرند. بنابراین، چهار انتخاب مختلف فازبندی گردش به چپ در هر سیکل وجود خواهد داشت.

- شمالی - جنوبی: مجاز/ شرقی - غربی: مجاز
- شمالی - جنوبی: محافظت شده/ شرقی - غربی: مجاز
- شمالی - جنوبی: مجاز/ شرقی - غربی: محافظت شده
- شمالی - جنوبی: محافظت شده/ شرقی - غربی: محافظت شده

مقادیر داده‌های استفاده‌شده برای کنترل تطبیقی چراغ راهنمایی در این تحقیق، بر اساس داده‌های واقعی برداشت شده از تقاطع بلوار کشاورز و وصال شیرازی در تهران، در شبیه‌ساز پیاده‌سازی شده‌اند. اطلاعات دریافتی از سازمان ترافیک شهرداری تهران نشان می‌دهد 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 این تقاطع می‌شوند و حرکات راست، مستقیم و چپ انجام می‌دهند. با توجه به موقعیت این تقاطع در مرکز شهر و حجم بالای تردد در بلوار کشاورز، تعداد خودروهای شبیه‌سازی‌شده تعدیل شده است. هدف این مطالعه ایجاد محیطی در شبیه‌ساز است که نمایانگر بسیاری از تقاطع‌های شهری باشد. بنابراین، تعداد خودروهای عبوری و هندسه تقاطع به صورت نسخه‌ای تعدیل شده در نظر گرفته شده‌اند. در محیط شبیه‌ساز، ورود خودروها به تقاطع با دو توزیع متغیر ویبل و یکنواخت انجام می‌شوند. نرخ جریان ورودی در حالت ترافیک کم ۱۰۰۰ و در حالت ترافیک زیاد ۲۰۰۰ خودرو در ساعت در نظر گرفته می‌شود. زمان شبیه‌سازی نیز یک ساعت لحاظ شده است.

۲ ـ ۱ ـ 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، یکی برای گردش به چپ و یکی برای حرکت مستقیم در صد متری تقاطع است. محیط تقاطع چهارراه در شبیه‌ساز SUMO پیاده‌سازی شده است و به همراه موقعیت آشکارسازهای حلقه‌ای در شکل ۲ نمایش داده شده است.



شکل ۲: محیط تقاطع چهارراه و موقعیت آشکارسازهای حلقه‌ای

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. این داده‌ها می‌توانند برای تحلیل وضعیت جریان ترافیک و استفاده در یادگیری تقویتی به کار گرفته شوند. فلوچارت روش کنترل چراغ راهنمایی در شکل ۳ نشان داده شده است.

همان گونه که اشاره شد، انتخاب نوع فازبندی سیکل بعد در پایان سیکل فعلی انجام می‌شود. در این راستا، باید به چندین فاکتور توجه کرد. ظرفیت تقاطع، ویژگی‌های هندسی و تاریخچه تصادف، فاکتورهای معمولی هستند که در انتخاب فازبندی مورد استفاده قرار می‌گیرند. اما یک فاکتور دیگر که نباید نادیده گرفته شود، حاصل ضرب حجم گردش به چپ و حجم مخالف است. این فاکتور نشان می‌دهد که چقدر تقاطع شلوغ و پیچیده است. بسیاری از محققان و مهندسان از این فاکتور به عنوان معیار اصلی برای انتخاب طرح‌های فازبندی استفاده می‌کنند (استاماتیادیس و همکاران، ۲۰۱۶). به صورت دقیق‌تر ابتدا حاصل ضرب نرخ جریان حرکات گردش به چپ در نرخ جریان خودروهای مقابل که حرکت مستقیم داشته و با حرکات گردش به چپ تداخل دارند، در هر چهار رویکرد محاسبه می‌شود. در این جا نیز یک بار دیگر به کمک یادگیری تقویتی نوع فازبندی انتخاب می‌گردد. به این صورت که حاصل ضرب‌های ذکر شده برای هر چهار رویکرد، به عنوان وضعیت در نظر گرفته می‌شود.

وضعیت مشاهده شده توسط عامل یادگیری زمان سبز و نوع فازبندی در واقع یک بردار با ۱۰ و ۵ مولفه است که در شکل ۴ نشان داده شده است.

شکل ۳: فلوچارت روش کنترل چراغ راهنمایی

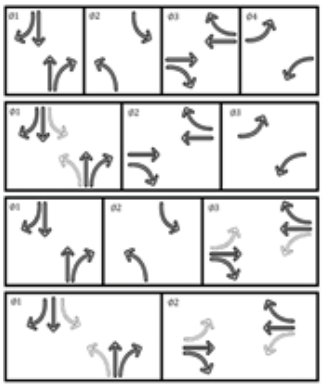
نمایش وضعیت برای عامل یادگیری زمان سبز	نمایش وضعیت برای عامل یادگیری نوع فازبندی
نرخ جریان خودروهای رویکرد غربی	حاصل ضرب نرخ جریان گردش به چپ رویکرد غربی و جریان مخاا
نرخ جریان خودروهای رویکرد شمالی	حاصل ضرب نرخ جریان گردش به چپ رویکرد شمالی و جریان مخا
نرخ جریان خودروهای رویکرد شرقی	حاصل ضرب نرخ جریان گردش به چپ رویکرد شرقی و جریان مخا
نرخ جریان خودروهای رویکرد جنوبی	حاصل ضرب نرخ جریان گردش به چپ رویکرد جنوبی و جریان مخا
نرخ جریان خودروهای گردش به چپ رویکرد غربی	نوع فازبندی سیکل
نرخ جریان خودروهای گردش به چپ رویکرد شمالی	
نرخ جریان خودروهای گردش به چپ رویکرد شرقی	
نرخ جریان خودروهای گردش به چپ رویکرد جنوبی	
نوع فازبندی سیکل کنونی	
شماره فاز در سیکل کنونی	

شکل ۴: مولفه‌های وضعیت‌های مشاهده شده توسط عامل یادگیری زمان سبز و نوع فازبندی

۲ـ۲ ـ مجموعه اقدامات

عامل یادگیری تقویتی پس از مشاهده وضعیت‌ها و دریافت پاداش در پایان هر گام زمانی ده ثانیه‌ای، تصمیم به تمدید زمان سبز برای ده ثانیه دیگر یا رفتن به فاز بعد می‌گیرد. انتخاب اقدام به کمک یادگیری تقویتی صورت می‌پذیرد. به این صورت که دو اقدام صفر و یک وجود دارد. عدد صفر برای وقتی که زمان سبز تمدید می‌شود و عدد یک 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 این ده ثانیه، تفاوت زمان انتظار در سیکل قبلی و فعلی به عنوان پاداش آنی (مربوط به گام زمانی ده ثانیه‌ای) به دست می‌آیند. در این جا تعداد صد اپیزود (شبیه‌سازی یک ساعته) برای یادگیری در نظر گرفته می‌شود. لازم به ذکر است که با داشتن پاداش آنی گام‌های زمانی در طول یک ساعت شبیه‌سازی، می‌توان مشاهده کرد که عملکرد عامل یادگیری تقویتی چگونه با گذشت اپیزودها در جهت بهینه‌سازی (ماکزیمم‌سازی پاداش‌ها) تغییر می‌کند.

همان گونه که اشاره شد، در پایان هر سیکل، انتخاب نوع فازبندی سیکل بعد به کمک یادگیری تقویتی عمیق انجام می‌پذیرد. چهار نوع فازبندی در نظر گرفته می‌شود که سعی می‌شود این بار نیز مناسب‌ترین نوع آن انتخاب شود. فازبندی منعطف پیشنهادی در شکل ۵ نمایش داده شده است. ابتدا حاصل ضرب نرخ جریان حرکات گردش به چپ در نرخ جریان خودروهای مقابل که حرکت مستقیم داشته و با حرکات گردش به چپ تداخل دارند، به همراه نوع فازبندی سیکل قبل به عنوان وضعیت در نظر گرفته می‌شود. سپس یک تعریف پاداش برای عامل انتخاب نوع فازبندی در نظر گرفته می‌شود که تفاوت زمان انتظار در سیکل قبلی و فعلی است.



شکل ۵: فازبندی منعطف پیشنهادی

۲ـ۳ ـ تابع پاداش

اگر $wait$ زمان انتظار خودروها باشد، یک تعریف پاداش برای عامل به صورت معادله (۳-۱) در نظر گرفته می‌شود. به این صورت که عامل می‌کوشد با به حداکثر رساندن آن، عملکرد خود را در انتخاب اقدام مناسب بهبود دهد.

$$r_t = wait_{t-1} - wait_t \qquad (۳-۱)$$

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

۲ـ۴ـ مدل یادگیری

همان گونه که اشاره شد، هدف عامل به حداکثر رساندن پاداش تجمعی‌ای است که در درازمدت دریافت می‌کند که در ساده‌ترین حالت، حاصل جمع پاداش‌ها است:

$$R_t \; \mathrm{B} r_t + r_{t+1} + r_{t+2} + \mathrm{L} \qquad (۳-۲)$$

برای جلوگیری از بی‌نهایت شدن مقدار پاداش تجمعی، از تخفیف استفاده می‌شود. بر اساس این رویکرد، عامل سعی می‌کند اقداماتی

را انتخاب کند که مجموع پاداش‌های تخفیفی که در آینده دریافت می‌کند به حداکثر برسد:

$$R_t \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau \quad (۳-۳)$$

که در آن γ پارامتری است به نام نرخ تنزیل که $0 \le \gamma \le 1$. همچنین:

$$R_t \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \cdots) = r_t + \gamma R_{t+1} \quad (۴-۳)$$

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 زیر تعریف می‌شوند:

$$Q^\pi(s,a) = E[R_t \mid s,a,\pi] \quad (۵-۳)$$

$$V^\pi(s) = E_{a:\,\pi(s)}[Q^\pi(s,a)] \quad (۶-۳)$$

تابع مقدار وضعیت-اقدام (تابع Q) روشی برای تعیین کمیت مزیت انجام یک اقدام خاص در یک وضعیت معین است. با تابع مقدار V متفاوت است، که فقط میزان خوب بودن یک وضعیت s را تخمین می‌زند. با توجه به معادله (۳-۴) تابع Q را می‌توان با استفاده از برنامه‌نویسی پویا به صورت بازگشتی محاسبه کرد:

$$Q^\pi(s,a) = E_{s'}[r + \gamma E_{a':\,\pi(s')}[Q^\pi(s',a')] \mid s,a,\pi] \quad (۷-۳)$$

که در آن مقادیر جفت وضعیت-اقدام (s,a) در گام فعلی و مقادیر (s',a') در گام بعدی است.

مفهوم اساسی شبکه عمیق Q یافتن تابع Q است که به عنوان ورودی وضعیت فعلی را می‌گیرد و بازده مورد انتظار هر اقدام را خروجی می‌دهد. تابع مقدار اقدام بهینه با توجه به یک محیط می‌تواند با استفاده از معادله بلمن به صورت

$$Q^*(s,a) = E_{s'}[r + \max_{a'} Q^*(s',a') \mid s,a] \quad (۸-۳)$$

بیان شود(وانگ و همکاران، ۲۰۱۶). در این‌جا، از شبکه عصبی به عنوان یک تقریب غیرخطی برای تخمین تابع مقدار - اقدام استفاده می‌شود. 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 این اقدامات یک مقدار Q وجود دارد که در هر بار تصمیم‌گیری عامل، اقدام با مقدار بیشتر انتخاب می‌گردد. در این مطالعه مانند بسیاری از پژوهش‌های گذشته، از یک شبکه عصبی عمیق برای تعیین مقدار Q استفاده شده است. برای آموزش عامل از یک مجموعه از نمونه‌ها استفاده می‌شود که شامل چهارتایی وضعیت قبلی، اقدام قبلی (انتخاب شده)، پاداش و وضعیت فعلی در هر گام است. این مجموعه از یک مجموعه بزرگ‌تر به نام حافظه به صورت تصادفی برداشته می‌شود. در واقع پس از تجربه عامل در هر گام، از مشاهده وضعیت تا انتخاب اقدام مناسب، وضعیت قبلی، اقدام قبلی (انتخاب شده)، پاداش و وضعیت فعلی به این حافظه افزوده می‌شود. پس از پر شدن اندازه آن نیز، تجربه‌های قدیمی‌تر به تدریج حذف شده و تجربه‌های جدید جایگزین می‌شوند. از یک شبکه عصبی عمیق برای تناظر وضعیت‌ها به مقدار Q اقدامات بهره برده شده است تا بتوان فرمول اصلی یادگیری تقویتی را اعمال نمود.

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 این مطالعه، شبکه عصبی با ترکیب‌های مختلفی از لایه‌ها، گره‌ها و توابع فعال‌سازی آزمایش شده است. بر اساس تلاش‌هایی که منجر به کسب پاداش‌های بهتر شدند، ساختاری با سه لایه پنهان پیاده‌سازی گردید که به ترتیب شامل ۱۶، ۳۲ و ۶۴ گره است. در لایه اول از تابع فعال‌سازی SELU استفاده شده و در لایه‌های بعدی (از جمله لایه نهایی) تابع ReLU به کار رفته است. لایه نهایی که مقادیر Q را تخمین می‌زند، به تعداد اقدامات ممکن گره دارد. با استفاده از این شبکه، پس از مشاهده وضعیت در شبیه‌ساز ترافیک، اقدام با بالاترین مقدار انتخاب می‌شود. ابر پارامترهای مورد استفاده در مدل یادگیری تقویتی عمیق در جدول ۱ ارائه شده است.

شایان توجه است که این الگوریتم برای هر دو انتخاب زمان سبز و نوع فازبندی مورد استفاده قرار می‌گیرد.

به منظور ارزیابی عملکرد الگوریتم دوئل دوگانه در کنترل تطبیقی چراغ راهنمایی، برداشت مقادیر Q علاوه بر شبکه DQN۳، از یک شبکه DQN[1] نیز انجام گرفته است. شبکه دوئل دوگانه ترکیبی از الگوریتم شبکه عمیق Q دوگانه و دوئل است :

۱.مدل شبکه عمیق Q دوگانه برای رسیدن به مقدار Q هدف، مقدار بهینه اقدام را از طریق شبکه هدف به دست می‌آورد. به بیان دقیق‌تر، شبکه هدف روشی است که در شبکه‌های یادگیری عمیق Q برای تثبیت آموزش استفاده می‌شود. همان گونه که پیش‌تر اشاره شد، در این شبکه‌ها، تابع مقدار Q با یک شبکه عصبی تقریب زده می‌شود. در حین آموزش، وزن‌های شبکه به گونه‌ای به‌روزرسانی 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، این ممکن است باعث ناپایداری در آموزش شود. برای حل این مسئله، یک شبکه هدف متفاوت از شبکه اصلی در نظر گرفته می‌شود. شبکه هدف ساختار مشابه با شبکه اصلی دارد ولی وزن آن کندتر به‌روزرسانی می‌شود. در این مطالعه وزن شبکه هدف در هر صد بار آموزش به روزرسانی می‌گردد. به این ترتیب، مقادیر Q هدف که برای به‌روزرسانی شبکه اصلی استفاده می‌شوند، برای چندین دور ثابت باقی می‌مانند که می‌تواند به تثبیت آموزش کمک کند (ون هاسلت، گئز و سیلور، ۲۰۱۶).

۲.در یادگیری تقویت عمیق، معمولا از یک شبکه عصبی برای تقریب تابع مقدار Q استفاده می‌شود که پاداش مورد انتظار را برای انجام یک اقدام در یک وضعیت مشخص تخمین می‌زند. شبکه دوئل، شبکه عصبی را به دو بخش تقسیم می‌کند: یک بخش برای تخمین تابع مقدار وضعیت و بخش دیگر برای تخمین تابع مزیت اقدام وابسته به وضعیت. تابع مقدار وضعیت، پاداش مورد انتظار در آینده را برای قرار گرفتن در یک وضعیت معین تخمین می‌زند، در حالی که تابع مزیت اقدام، اهمیت نسبی هر اقدام را در آن وضعیت تخمین می‌زند .

ارتباط تابع مزی A با مقدار و توابع Q به صورت زیر تعریف می‌شود:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \qquad (۹-۳)$$

شایان ذکر است که:

$$E_{a:\,\pi(s)}[A^{\pi}(s,a)] = 0 \qquad (۱۰-۳)$$

فایده این روش این است که شبکه می‌تواند بفهمد که کدام وضعیت‌ها ارزشمند هستند (دارای مقادیر Q بیشتر)، بدون آن که نیاز باشد تأثیر هر اقدام را در هر وضعیت یاد بگیرد. این باعث می‌شود که سیاست بهتری در شرایطی که بسیاری از اقدامات مقدار مشابه دارند، داشته باشد. این شبکه، عملکرد بهتری نسبت به یک شبکه عصبی معمولی دارد (وانگ و همکاران، ۲۰۱۶).

بنابراین، به کمک الگوریتم یادگیری تقویتی عمیق دوئل دوگانه می‌توان ارزش اقدامات ممکن را در پایان گام‌های زمانی ده ثانیه‌ای و نیز در انتهای هر سیکل به دست آورد و اقدام با بیشترین ارزش را انتخاب کرد. این تصمیمات به محیط شبیه‌سازی انتقال داده خواهند شد و بر وضعیت تقاطع چراغدار در گام‌های زمانی آینده اثرگذار خواهند بود.

۳ ـ نتایج عددی

کنترل چراغ ترافیکی در یک تقاطع چراغدار به شیوه تطبیقی، به کمک یادگیری تقویتی عمیق و با دو رویکرد بررسی می‌شود. عملکرد کنترل‌کننده با شبکه دوئل دوگانه در با مقایسه نتایج شبیه‌سازی برای کنترل چراغ ترافیکی با یادگیری تقویتی عمیق ارزیابی می‌شود. در کنترل تطبیقی و هر دو رویکرد، حداقل زمان سبز ۱۰ و حداکثر این زمان ۵۰ ثانیه لحاظ شده است. زمان زرد ۳ و مدت زمان تمام قرمز هم ۲ ثانیه لحاظ می‌شوند. معیار مورد استفاده نیز طول صف خودروها است که تعداد کل وسایل نقلیه متوقف شده در پایان هر شبیه‌سازی یک ساعته برای همه رویکردها است. سرعت کمتر از ۰.۱ متر بر ثانیه توقف در نظر گرفته می‌شود. همان گونه که در جدول ۲ آورده شده است، طول صف میان دو نوع الگوریتم یادگیری تقویتی در حالت‌های مختلف توزیع و میزان تقاضا مقایسه شده‌است. جدول ۳ نیز میزان کاهش طول صف تجمعی در الگوریتم دوئل دوگانه نسبت به Q عمیق را درحالت‌های مختلف توزیع و میزان تقاضا مقایسه کرده است. در شکل ۶، طول صف تجمعی خودروها در پایان هر ساعت

[1] Deep Q-Network

شبیه‌سازی نمایش داده شده است. هر اپیزود نمایانگر یک ساعت از اجرای شبیه‌سازی و عبور خودروها است. با افزایش تعداد اپیزودها، فرایند یادگیری بهبود می‌یابد و مشاهده می‌شود که با حرکت به سمت انتخاب‌های مناسب‌تر، خروجی‌های بهتری از شبیه‌ساز حاصل می‌شود. این روند نشان‌دهنده تأثیر مثبت تجربه بر عملکرد مدل در بهینه‌سازی ترافیک است. آن گونه که از نتایج جداول و نمودارهای طول صف خودروها در شکل ۶ مشخص است، 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 می‌دهد. دیگر معیارهای ارزیابی نیز روند مشابهی را نشان داده‌اند. لازم به ذکر است که در بررسی نمودارها و استخراج نتایج، میانگین اعداد پنج درصد پایانی مراحل به عنوان نتیجه یادگیری عامل تقویتی فرض می‌شوند. علاوه بر این شکل ۷ طول صف خودروها را در صورتی که از الگوریتم دوئل دوگانه و یادگیری تقویتی عمیق ساده استفاده شود، در حالت‌های مختلف توزیع و میزان تقاضا مقایسه کرده است. در شکل ۷، تأثیر کنترل‌کننده چراغ ترافیکی با استفاده از الگوریتم یادگیری تقویتی دوئل دوگانه در مقایسه با نتایج شبیه‌سازی کنترل‌کننده یادگیری تقویتی عمیق Q نشان داده شده است. پس از تحلیل طول صف تجمعی وسایل نقلیه در این تقاطع، مشاهده می‌شود که مقدار این پارامتر ترافیکی با استفاده از کنترل‌کننده پیشنهادی نسبت به یادگیری تقویتی عمیق Q کاهش یافته است.

جدول ۱ : ابر پارامترهای مورد استفاده در مدل یادگیری تقویتی عمیق

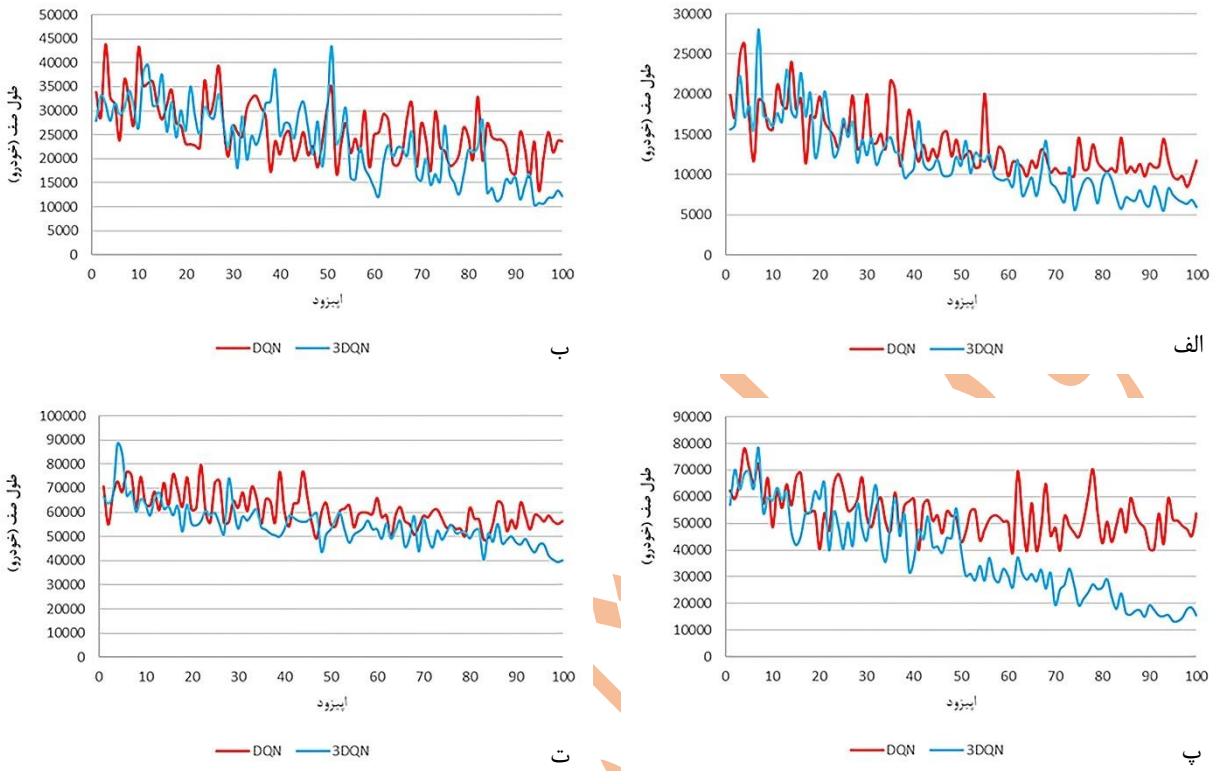
ابر پارامترها	حداکثر سرعت (m/s)	تقاضای خودروها (vph)		ضریب تخفیف	تعداد نمونه‌ها	اندازه حافظه	تعداد جایگزینی شبکه هدف	تعداد نمونه برای انتخاب نوع فازبندی	اندازه حافظه برای انتخاب نوع فازبندی
		کم	زیاد						
مقدار	۱۵	۱۰۰۰	۲۰۰۰	۰/۸	۲۵۲	۱۰۰۰۰	۱۰۰	۶۴	۲۰۰۰

جدول ۲ : مقایسه طول صف تجمعی میان دو نوع الگوریتم درحالت‌های مختلف توزیع و میزان تقاضا

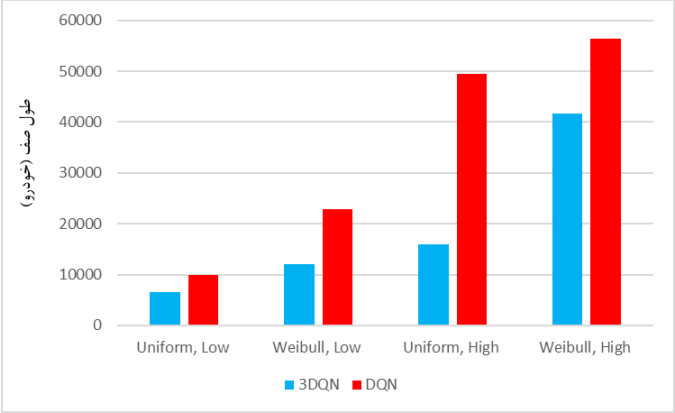
طول صف تجمعی (خودرو)					
حجم خودروها		کم		زیاد	
توزیع تقاضا		یکنواخت	متغیر	یکنواخت	متغیر
نوع الگوریتم	DQN	۹۹۳۹	۲۲۹۴۰	۴۹۴۶۵	۵۶۴۵۹
	3DQN	۶۵۹۷	۱۲۰۵۳	۱۶۰۰۱	۴۱۶۱۶

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

کاهش طول صف تجمعی در الگوریتم دوئل دوگانه				
حجم خودروها	کم		زیاد	
توزیع تقاضا	یکنواخت	متغیر	یکنواخت	متغیر
میزان کاهش (٪)	۳۳.۶۳	۴۷.۴۶	۶۷.۶۵	۲۶.۲۹



شکل ۶: مقایسه طول صف خودروها میان دو الگوریتم یادگیری تقویتی: (الف) تقاضای یکنواخت و کم؛ (ب) تقاضای متغیر و کم؛ (پ) تقاضای یکنواخت و زیاد؛ (ب) تقاضای متغیر و زیاد



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

۴ ـ نتیجه‌گیری

با استفاده از کنترل تطبیقی چراغ راهنمایی، که یکی از روش‌های پیشرفته در مدیریت ترافیک شهری به شمار می‌رود، می‌توان به طور قابل ملاحظه‌ای زمان انتظار خودروها در تقاطع‌های چراغ‌دار را کاهش داد. این روش با بهره‌گیری از الگوریتم‌های بهینه‌سازی، زمان‌های چراغ سبز را بر اساس شرایط ترافیکی لحظه‌ای بهینه می‌کند. این مطالعه بر نرخ جریان حرکات گردشی تمرکز دارد و با استفاده از یادگیری تقویتی، زمان سبز بهینه برای هر فاز در یک تقاطع چهارراه را تعیین می‌کند. علاوه بر این، برای کاهش تداخل حرکات گردشی خودروها با جریان‌های مخالف، نوع فازبندی منعطفی 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 حداقل دو تا حداکثر چهار فاز متفاوت باشد. در نتیجه، ترکیب مناسبی از فازها در هر سیکل به دست می‌آید. با این روش، کنترل تقاطع بسته به زمان روز و سطح فعالیت ترافیک در هر تقاطع، تغییر می‌کند. هدف از این روش، سازگار شدن با شرایط متغیر ترافیک است.

در این مقاله، دو نوع الگوریتم یادگیری تقویتی عمیق پیشنهاد شد. شبکه Q عمیق دوئل دوگانه، ترکیبی از الگوریتم‌های شبکه عمیق Q دوگانه و دوئل است. این شبکه از یک شبکه هدف برای تثبیت آموزش و کاهش خطای بین مقادیر Q پیش‌بینی‌شده و هدف استفاده می‌کند. شبکه هدف که وزن آن کندتر به‌روزرسانی می‌شود، مقادیر Q هدف را برای چندین دور ثابت نگه می‌دارد. از سوی دیگر، شبکه دوئل شبکه عصبی را به دو بخش تقسیم می‌کند: یکی برای تخمین تابع مقدار وضعیت و دیگری برای تخمین تابع مزیت اقدام. این دو تابع، پاداش مورد انتظار در آینده و اهمیت نسبی هر اقدام را تخمین می‌زنند. به منظور ارزیابی عملکرد الگوریتم دوئل دوگانه، برای برداشت مقادیر Q از یک شبکه DQN نیز استفاده شد. الگوریتم‌های پیشنهادی برای رسیدن به اهداف این پژوهش در محیط برنامه‌نویسی پایتون کد شدند. از طریق ریزشبیه‌سازی SUMO که با محیط برنامه‌نویسی در تعامل مکرر است، نشان داده شد که کنترل‌کننده پیشنهادی که از شبکه دوئل دوگانه بهره می‌برد، کارایی ترافیک را از نظر معیار طول صف تجمعی، بیشتر از کنترل‌کننده با شبکه عصبی رایج بهبود می‌دهد.

با توجه به نتایج به دست آمده، اگر حجم خودروها کم باشد و جریان به صورت متغیر عبور کند، استفاده از شبکه دوئل دوگانه مؤثرتر از توزیع یکنواخت است. در حجم بالای خودروها، این الگوریتم 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، هرچه حجم عبوری کمتر باشد، عملکرد الگوریتم دوئل دوگانه بهتر می‌شود.

در این مطالعه، بهینه‌سازی تنها برای یک معیار، یعنی طول صف تجمعی، انجام شده است. با این حال، امکان گنجاندن هر معیار قابل اندازه‌گیری دیگری در فرآیند بهینه‌سازی با استفاده از این روش‌ها وجود دارد. به عبارت دیگر، می‌توان پاداش را به صورت جمع وزنی چندین معیار تعریف کرد تا عامل یادگیری تقویتی بتواند تصمیم‌گیری‌های بهتری انجام دهد. یکی دیگر از بهبودها می‌تواند در نظر گرفتن منطقه پارکینگ نزدیک به تقاطع باشد. همچنین، تأثیر عبور عابران پیاده نیز می‌تواند در انتخاب نوع فازبندی لحاظ شود. در مراحل بعدی، می‌توان به جای تمرکز بر یک تقاطع منفرد، شبکه‌ای از تقاطع‌ها را مورد بررسی قرار داد و تأثیرات متقابل آن‌ها را نیز در نظر گرفت.

۵-منابع

افندی زاده، شهریار، توکلی کاشانی، علی، حسن پور، شهاب (۱۳۹۳، بهار). ارائه مدل اولویت‌بندی ایمنی تقاطع‌های همسطح. مطالعات پژوهشی راهور. (۳) ۹، ۱۳۸-۱۱۱.

Chin, Y.K., Lee, L.K., Bolong, N., Yang, S.S. and Teo, K.T.K. (۲۰۱۱). Exploring Q-learning optimization in traffic signal timing plan management. In Proceedings of the ۳rd International Conference on Computational Intelligence, Communication Systems and Networks, Bali, Indonesia, ۲۶۹-۲۷۴, doi: ۱۰,۱۱۰۹/CICSyN.۲۰۱۱,۶۴

Eriksen, A. B., Lahrmann, H., Larsen, K. G., and Taankvist, J. H. (۲۰۲۰). Controlling Signalized Intersections using Machine Learning. Transportation Research Procedia, ۴۸, ۹۸۷-۹۹۷, doi:۱۰,۱۰۱۶/j.trpro.۲۰۲۰,۰۸,۱۲۷.

Genders, W. and Razavi, S. (۲۰۱۶). Using a deep reinforcement learning agent for traffic signal control. arXiv preprint arXiv:۱۶۱۱,۰۱۱۴۲.

Han, G., Zheng, Q., Liao, L., Tang, P., Li, Z., and Zhu, Y. (۲۰۲۲). Deep reinforcement learning for intersection signal control considering pedestrian behavior. Journal of Electronics, ۱۱(۲۱), ۳۵۱۹. doi: ۱۰,۳۳۹۰/electronics۱۱۲۱۳۵۱۹

Haydari, A., and Yilmaz, Y. (۲۰۲۰). Deep Reinforcement Learning for Intelligent Transportation Systems: A Survey. IEEE Transactions on Intelligent Transportation Systems, ۲۳(۱), ۱۱-۳۲. doi: ۱۰,۱۱۰۹/TITS.۲۰۲۰,۳۰۰۸۶۱۲

La, P. and S. Bhatnagar, Reinforcement Learning With Function Approximation for Traffic Signal Control. IEEE Transactions on Intelligent Transportation Systems, (۲۰۱۱). ۱۲(۲): p. ۴۱۲-۴۲۱.

Mannion, P., Duggan, J., and Howley, E. (۲۰۱۶). An experimental review of reinforcement learning algorithms for adaptive traffic signal control. Autonomic road transport support systems, ۴۷-۶۶, doi:۱۰,۱۰۰۷/۹۷۸-۳-۳۱۹-۲۵۸۰۸-۹_۴.

Miletić, M., Ivanjko, E., Gregurić, M., and Kušić, K. (۲۰۲۲). A review of reinforcement learning applications in adaptive traffic signal control. IET Intelligent Transportation Systems, ۱۶, ۱۲۶۹-۱۲۸۵. doi: ۱۰,۱۰۴۹/itr۲,۱۲۲۰۸.

Mousavi, S. S., Schukat, M., and Howley, E. (۲۰۱۷). Traffic light control using deep policy-gradient and value-function based reinforcement learning. IET Intelligent Transportation Systems, ۱۱, ۴۱۷-۴۲۳. doi: ۱۰,۱۰۴۹/iet-its.۲۰۱۷,۰۱۵۳.

Stamatiadis, N., Tate, S., and Kirk, A. ۲۰۱۶. Left-turn phasing decisions based on conflict analysis. Transportation Research Procedia, ۱۴, ۳۳۹۰-۳۳۹۸. doi: ۱۰,۱۰۱۶/j.trpro.۲۰۱۶,۰۵,۲۹۱.

Sutton, R. S., and Barto, A. G. (۲۰۱۸). Introduction. In Reinforcement learning: an introduction (۲nd ed., ch. ۱, sec. ۱,۱, pp. ۱-۲). Cambridge, Massachusetts (London, England): The MIT Press.

Touhbi, S., Babram, M.A., Nguyen-Huu, T., Marilleaub, N., Hbid, M.L., Cambier, C. and Stinckwich, S. (۲۰۱۷). Adaptive traffic signal control exploring reward definition for reinforcement learning. In Proceedings of the ۸th International Conference on Ambient Systems, Networks and Technologies, Madeira, Portugal, ۵۱۳-۵۲۰, doi: ۱۰,۱۰۱۶/j.procs.۲۰۱۷,۰۵,۳۲۷.

van Hasselt, H., Guez, A. and Silver, D. (۲۰۱۶). Deep reinforcement learning with double Q-learning, In Proceedings of the ۳۰th AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, doi: ۱۰,۱۶۰۹/aaai.v۳۰i۱,۱۰۲۹۵.

Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M. and de Freitas, N. (۲۰۱۶). Dueling network architectures for deep reinforcement learning. In Proceedings of the ۳۳rd International Conference on Machine Learning, New York, NY, USA, ۱۹۹۵-۲۰۰۳, doi: ۱۰,۴۸۵۵۰/arXiv.۱۵۱۱,۰۶۵۸۱

Wei, H., Zheng, G., Yao, H. and Li, Z. (۲۰۱۸). IntelliLight: A reinforcement learning approach for intelligent traffic light control. In Proceedings of the ۲۴th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, United Kingdom, ۲۴۹۶-۲۵۰۵, doi: ۱۰,۱۱۴۵/۳۲۱۹۸۱۹,۳۲۲۰۰۹۶.

Zheng, G., Zang, X., Xu, N., Wei, H., Yu, Z., Gayah, V., Xu, K. and Li, Z. (۲۰۱۹). Diagnosing reinforcement learning for traffic signal control, arXiv preprint arXiv:۱۹۰۵,۰۴۷۱۶.

Zheng, Q., Xu, H., Chen, J., Zhang, D., Zhang, K., & Tang, G. (۲۰۲۲). Double deep Q-network with dynamic bootstrapping for real-time isolated signal control: a traffic engineering perspective. Applied Sciences, ۱۲(۱۷), ۸۶۴۱. doi: ۱۰,۳۳۹۰/app۱۲۱۷۸۶۴