فصلنامه مهندسی حمل و نقل، دوره ١٤، شماره ٤، تابستان ١٤٠٢، ص. ٣١٦٨-٣١٥١

مقاله علمی-پژوهشی

کنترل انعطاف‌پذیر سیگنال ترافیک مبتنی بر توسعه نمایش حالت در روش یادگیری تقویتی عمیق در هنگام وقوع تصادف در تقاطعات شهری

زهرا زینلی، دانشجوی دکترا، گروه مهندسی کنترل، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، ایران

مهدی سجودی (مسئول مکاتبات)، دانشیار، گروه مهندسی کنترل، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، ایران

E-mail: sojoodi@modares.ac.ir

دریافت: ١٤٠٠/١٢/١١ پذیرش: ١٤٠١/٠٣/٢١

چکیده

روش‌های یادگیری تقویتی عمیق نتایج امیدوارکننده‌ای را در توسعه کنترل‌کننده‌های سیگنال ترافیک نشان داده‌اند. در این مقاله، انعطاف‌پذیری یک کنترل‌کننده مبتنی بر یادگیری تقویتی عمیق را در شرایط ترافیک با حجم زیاد و تحت طیف وسیعی از اختلالات محیطی مانند تصادفات، بررسی کرده و یک کنترل‌کننده قابل اعتماد را در محیط با ترافیک پویا پیشنهاد می‌دهیم. در این روش ،با استفاده از رویکرد گسسته‌سازی هر یک از خیابان‌های چهارراه به سلول‌هایی تقسیم شده وتاثیر اندازه این سلول‌ها به لحاظ متفاوت بودن یا یکسان بودن با یکدیگردر کارآیی الگوریتم بررسی می‌گردد. با انتخاب یک فضای حالت توسعه یافته و متراکم، اطلاعاتی به عامل به عنوان ورودی داده می‌شودکه بتواند درک کاملی از محیط را در اختیار عامل قرار دهد. برای آموزش عامل از روش یادگیری عمیق Q و بازپخش تجربه استفاده شده و مدل پیشنهادی در شبیه‌ساز ترافیک SUMO ارزیابی شده است. نتایج شبیه‌سازی کارایی روش پیشنهادی را در کاهش طول صف حتی در صورت وجود اختلال تأیید می‌کند.

واژه‌های کلیدی: ایمنی ترافیک، تصادف، کنترل ترافیک، یادگیری تقویتی عمیق

۱. مقدمه

افزایش در سال‌های اخیر مسئله ترافیک به دلیل اثرات مخرب آن از جمله آلودگی هوا، طولانی شدن زمان سفر و اتلاف انرژی به عنوان یک چالش در اکثر شهرها مطرح شده است. با توجه به محدود بودن ظرفیت جاده ممکن است گسترش جاده راهی مناسب برای افزایش جریان وسایل نقلیه و کاهش ترافیک به نظر رسد، ولی این سرمایه‌گذاری در دراز مدت پر هزینه و بی‌اثر خواهد بود. بنابراین موثرترین راه‌حل، بهینه‌سازی سیگنال کنترل چراغ راهنمایی جهت افزایش کارایی زیرساخت‌های موجود است[Genders and Razavi, 2016] . دو استراتژی اصلی به نام‌های زمان‌بندی ثابت[۱] و کنترل تطبیقی سیگنال ترافیکی[۲] برای کنترل چراغ راهنمایی وجود دارد. در کنترل زمان‌بندی ثابت، زمان‌بندی چراغ‌های راهنمایی به جای در نظرگرفتن داده‌های ترافیک بلادرنگ[۳]، بر اساس داده‌های پیشین تعیین می‌شود [Casas 2017]. ولی جریان ترافیک ثابت نبوده و می‌تواند تحت تأثیر شرایط آب و هوایی، تصادفات و رویدادهای خاص، متغیر باشد. بنابراین در این موارد کارایی روش زمان‌بندی ثابت به خطر

می‌افتد. کنترل تطبیقی سیگنال ترافیک، وسایل نقلیه را با استفاده از حلقه‌های مغناطیسی تعبیه‌شده در کف خیابان شناسایی کرده و بر این اساس می‌تواند مدت زمان سبز بودن سیگنال ترافیکی را بر اساس تقاضای ترافیک در زمان واقعی بهینه کند [Jamil, Ganguly and Nower, 2021; Wang, Cao and Hussain, 2021]. به منظور کنترل کارآمدتر سیگنال‌های ترافیکی، اخیرا تکنیک‌های هوش مصنوعی به کار گرفته شده‌اند. در این زمینه، روش یادگیری تقویتی[۴] در مسائل کنترل سیگنال ترافیک، مسئله را به صورت یک فرآیند تصمیم‌گیری مارکوف[۵] در نظر می‌گیرد [Yoon et al, 2021]. رویکرد یادگیری تقویتی می‌تواند محیط‌های پیچیده و نامطمئن[۶] را از طریق تعامل با محیط و مشاهده تغییرات در رفتار آنها مدل کند. در مسئله کنترل ترافیک، عامل[۷] حالتی از چهارراه مانند جریان ترافیک را مشاهده کرده و بر اساس آن عملی را انجام می‌دهد. سپس عامل یک سیگنال پاداش دریافت کرده و بر اساس آن یاد می‌گیرد که پاداش تجمعی را به حداکثر برساند. با اینحال، محدودیت چنین الگوریتمی این است که با افزایش اندازه محیط، فضای حالت و عمل به صورت نمایی رشد خواهد کرد. این محدودیت، استفاده از یادگیری تقویتی در شرایط واقعی را دشوار می‌سازد. در سال‌های اخیر، رویکرد یادگیری تقویتی عمیق، که ترکیبی از یادگیری تقویتی سنتی و شبکه عصبی عمیق است، معرفی شده است [Bálinta, Tamás and Tamás, 2022; Li, Lv, and Wang, 2016; Chu et al, 2019; Liang et al, 2019; Chu, Lam and Li, 2021; Wei et al. 2021].

بدون شک در این رویکرد، فرمولبندی عامل که شامل تعیین حالت، عمل و پاداش است، تاثیر چشمگیری در عملکرد الگوریتم دارد. پاداش، یک مقدار اسکالر است که محیط، آن را هنگامی که یک عمل توسط عامل انجام می‌شود، ارسال می‌کند. این مقدار به عنوان یک تابع هدف در مسائل بهینه‌سازی تفسیر می‌شود، بنابراین پاداش کلی باید در کل فرآیند بهینه‌سازی به حداکثر برسد. در کنترل ترافیک، عوامل مختلفی به عنوان پاداش در نظر گرفته می‌شوند، مانند طول صف، زمان انتظار، تعداد توقف‌ها و سرعت متوسط [Zheng et al, 2019; Wei et al, 2018;Paul and Mitra, 2022; Chu et al, 2019].

برای عمل، مواردی مانند مدت زمان سبز بودن فاز فعلی، حفظ فاز فعلی یا انتقال به فاز بعدی و یا انتخاب فاز بعدی از یک مجموعه از پیش تعیین‌شده از فازها می‌تواند به‌کار رود. [Gao et al, 2017; Van der pol and Oliehoek, 2016].

حالت[۸] به عنوان درک عامل از محیطی که در آن قرار دارد، تعریف می‌شود. عناصری مانند موقعیت وسایل نقلیه، میانگین سرعت، طول صف، زمان انتظار، فاز و مدت زمان فازها به عنوان برخی از حالت‌های در نظر گرفته شده در مقالات هستند [Liang et al, 2019; Chu et al,2017; Gao et al, 2017; Liang et al, 2018; Vidali et al, 2019].

برخی محققان استفاده از نمایش تصویری چهارراه را به عنوان

حالت برای درک بهتر محیط پیشنهاد کرده‌اند [Liang et al, 2019; Mousavi, Schukat. and Howley, 2017]. با اینحال، مواردی وجود دارد که حالت‌های پیچیده مانند تصاویر منجر به عملکرد عالی و بی‌نقص نشده و بنابراین حالات ساده ترجیح داده می‌شوند [Zheng et al. 2019]. در [Genders and Razavi, 2016]، با استفاده از گسسته‌سازی، جاده به سلول‌هایی با طول مساوی تقسیم شده و از اطلاعات درون این سلول‌ها که شامل چگالی خودروها و سرعت متوسط خودروهای درون هر سلول است، به عنوان حالت استفاده می‌شود. بر اساس [Genders and Razavi, 2016]، در [Vidali et al, 2019] یک روش یادگیری تقویتی عمیق برای کنترل چراغ راهنمایی در یک چهارراه با گسسته‌سازی خیابان پیشنهاد شده است. در این رویکرد، حالت به صورت یک بردار در نظر گرفته شده و حضور خودرو در یک سلول با عدد یک و عدم حضور خودرو با عدد صفر در این بردار نشان داده می‌شود. نتایج مقاله مذکور حاکی از این است که این تعریف حالت منجر به عملکرد مطلوب در سناریوی ترافیک با حجم پایین شده ولی در حجم بالای ترافیک، الگوریتم عملکرد خوبی را نشان نداده است. بنابراین در سناریوی ترافیک بالا عامل نیاز به درک بیشتری از محیط دارد. رویکرد گسسته سازی خیابان به تعدادی سلول برای تعریف حالت می‌تواند به صورتی باشد که اندازه سلول‌ها با هم مساوی و یا متفاوت باشند. البته هر کدام که درک بیشتری از محیط به عامل بدهد مقدم است. از سوی دیگر، عموما انتخاب حالت ساده‌تر ترجیح داده می‌شود زیرا در صورت پیچیده بودن حالت انتخابی، ممکن است در عمل دستیابی به اطلاعات حالت امکان پذیر نبوده و یا بار محاسباتی زیادی را به سیستم تحمیل کند. در این مقاله تلاش شده تا با درک بیشتر محیط، در سناریوی ترافیک با حجم بالا و در ساعات اوج ترافیک نیز عملکرد مطلوب حاصل شود. همچنین در دو حالت یکسان بودن و متفاوت بودن اندازه سلول‌ها، الگوریتم مورد بررسی قرار گرفته است تا روش کارآمدتر تشخیص داده شود. اخیراً، تجزیه و تحلیل ایمنی و تصادف به طور فزاینده‌ای در کنترل ترافیک محبوب شده است [Roy, Hossain and Muromachi, 2022; Gong et al, 2020; Essa and Sayed, 2020; Li et al, 2020; Rodrigues and Azevedo, 2019; Mehmandar et al, 2020]. در [Gong et al, 2020]، یک الگوریتم کنترل تطبیقی سیگنال ترافیک برای به حداکثر رساندن کارایی ترافیک و ایمنی به طور همزمان پیشنهاد شده است. داده‌های ترافیکی بلادرنگ به عنوان ورودی به عامل اعمال شده تا در هر ثانیه، فاز مناسب جهت کاهش تاخیر وسیله نقلیه و خطر تصادف در تقاطع انتخاب شود. با مرور پژوهش‌های گذشته به این امر پی می‌بریم که با توجه به اینکه عامل آموزش دیده (چراغ راهنمایی)، محیط را از طریق حالت درک می‌کند، اختلال در محیط می‌تواند تأثیر منفی بر روند آموزش آن داشته باشد. شرایط بد آب و هوایی، تصادفات، توقف یک یا چند وسیله نقلیه و تغییر تراکم تردد در هنگام رویدادهای خاص از جمله عواملی هستند که محیط را غیرقابل پیش‌بینی کرده و منجر به اختلال می‌شوند. بنابراین، درک بهتر از محیط قطعا الگوریتم کنترل را انعطاف‌پذیر[9] و مقاوم خواهد کرد. با توجه به مباحث مطرح شده مواردی که در این مقاله به عنوان نوآوری به آن پرداخته شده به این شرح است: رویکرد ارائه شده در [Vidali et al, 2019] در سناریوی ترافیک پایین نتایج خوبی را به همراه دارد ولی در سناریوی ترافیک بالا عملکرد مطلوب نیست. بنابراین در این مقاله، مسئله کنترلی توسعه داده شده و کنترل چراغ راهنمایی در یک چهارراه در سناریوی ترافیک بالا با استفاده از یادگیری تقویتی عمیق در نظر گرفته‌شده‌است. نظر به اینکه در سناریوی ترافیک بالا به درک بیشتری از محیط نیاز است، حالت توسعه یافته و در عین سادگی ورودی بیشتری از محیط به عامل اعمال خواهد شد. همچنین تاثیر اندازه‌های سلول در سلول‌بندی خیابان بررسی شده است. دیگر نوآوری مقاله از این جهت است که انعطاف‌پذیری الگوریتم در شرایط وقوع تصادف نیز مطالعه می‌گردد. توسعه و گسترش حالت به این دلیل لحاظ شده که ارائه ورودی‌هایی که منعکس‌کننده تغییر

شرایط محیطی مرتبط با حادثه است، به عامل درک واقعی‌تری از محیط می‌دهد. نظر به اینکه تصادف باعث توقف خودروها می‌شود، عامل می‌تواند وضعیت فعلی تقاطع را با اضافه کردن تعداد خودروهای در صف به حالت، بهتر درک کند. در نتیجه، این مقاله علاوه برحضور یا عدم حضور خودروها، تعداد خودروهای در صف و همچنین فاز فعلی چراغ راهنمایی را نیز به عنوان حالت در نظر می‌گیرد. به این‌ترتیب الگوریتم پیشنهادی انعطاف‌پذیرتر شده است. در این مقاله، تصادف را به صورت توقف ناگهانی وسیله نقلیه در طی یک دوره زمانی شبیه‌سازی می‌کنیم. توقف یک وسیله نقلیه می‌تواند باعث بسته شدن یک لاین و در نتیجه تراکم ترافیک شود. تاثیر اندازه سلول‌ها در گسسته‌سازی محیط در این حالت نیز مطالعه می‌گردد. نوآوری قابل توجه در این تحقیق افزایش جریان عبوری از چهارراه و جلوگیری از کاهش صف طولانی در پشت چراغ قرمز در ترافیک با حجم بالا و در ساعات اوج ترافیک حتی در صورت وقوع اختلال در محیط به دلیل وقوع عواملی مانند تصادف است. الگوریتم پیشنهادی مستقل از زمان و مکان وقوع تصادف بوده، به طوریکه حتی اگر در ساعات اوج ترافیک در هر مکان از چهار راه تصادفی رخ دهد و یک یا چند ماشین در مسیر عبور خودروها متوقف گردند، با بهینه شدن سیگنال کنترلی صف طویلی در پشت چراغ راهنمایی ایجاد نمی شود و زمان انتظار خودروها جهت عبور از چهارراه به طور فزاینده‌ای افزایش نمی-یابد. بلکه طول صف تشکیل شده متشکل از خودروها در هنگام وقوع تصادف نیز نسبت به طول صف ایجاد شده با الگوریتم‌های موجود کاهش یافته است. این در حالیست که همگرا شدن به سیگنال کنترلی بهینه نیز با سرعت بیشتر نسبت به الگوریتم‌های ارائه شده در [Vidali et al, 2019] و همچنین برخی کارهای مرتبط قبلی است. افزایش سرعت به این معنی است که الگوریتم با تعداد اپیزودهای کمتر در طول آموزش به سیگنال کنترل بهینه همگرا می‌شود. همچنین در صورت وقوع دو تصادف در دو مکان مختلف به صورت همزمان نیز الگوریتم کارایی خود را حفظ کرده و منحنی پاداش در طول دوره آموزش به همگرایی می‌رسد. از سوی دیگر کاربرد عملیاتی این تحقیق نیز مورد توجه است. زیرا در سال های اخیر با افزایش جمعیت و افزایش مهاجرت به شهرهای بزرگ شاهد افزایش ترافیک در خیابان‌ها هستیم. همچنین افزایش تعداد خودروهای عبوری از چهارراه امکان بروز تصادف را افزایش می دهد. بنابراین در عمل استفاده از الگوریتم‌های زمان بندی ثابت برای چراغ راهنمایی منجر به زمان انتظار طولانی و افزایش تعداد خودروهای موجود در صف در چهارراه می شود. الگوریتم ارائه شده در این تحقیق با بهره-گیری از تکنیک هوش مصنوعی و با تعامل با محیط به صورت بلادرنگ، وضعیت محیط را در هر لحظه به صورت آنلاین نظاره کرده و بر اساس وضعیت فعلی چهارراه بهینه‌ترین سیگنال کنترلی را اعمال می‌کند. بنابراین در محیط واقعی که عواملی مانند تصادف یا شرایط آب و هوایی و یا برگزاری رویدادهای خاص منجر به تغییر ناگهانی در جریان ترافیک می‌شوند، چراغ راهنمایی با مشاهده محیط عمل مناسب را به صورت بلادرنگ انجام داده و از ازدحام خودروها جلوگیری می کند. لذا با توجه به اینکه عدم قطعیت در محیط واقعی وجود دارد، الگوریتم ارائه شده می تواند با انعطاف‌پذیری قابل توجه در کاربردهای عملیاتی مفید واقع شود. جهت آموزش عامل از الگوریتم یادگیری Q استفاده می‌شود و یک شبکه کاملا متصل[10] ، مقادیر Q را برای هر عمل تقریب می‌زند. علاوه بر این، روش بازپخش تجربه[11] و شبکه هدف برای جلوگیری از واگرایی الگوریتم به‌کار رفته است [Gao et al, 2017]. برای ارزیابی عملکرد از شبیه‌ساز SUMO استفاده شده است [Krajzewicz et al, 2012]. این شبیه‌ساز می‌تواند مشاهدات تقاطع را دریافت کرده و از طریق یک رابط برنامه‌نویسی کاربردی[12] (API) عمل انتخابی را روی عامل اعمال کند. به این ترتیب شرایط یک ترافیک در محیط واقعی ایجاد می‌شود. حتی در سناریوهای ترافیک بالا، نتایج ارزیابی ثابت می‌کند که رویکرد پیشنهادی می‌تواند نتیجه مطلوبی

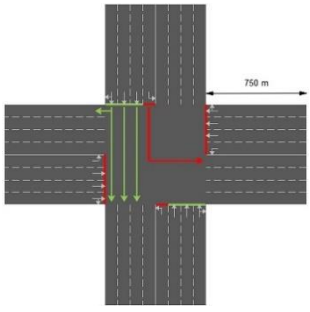
را ارائه دهد و در صورت وجود اختلالات محیطی به طور موثر با آن‌ها مقابله کند.

ساختار مقاله به این شرح است: در بخش ۲ مسئله به همراه فرضیات در نظر گرفته شده بیان می‌شود. بخش ۳ به تعریف حالت، عمل، پاداش و شبکه عصبی عمیق انتخاب شده می‌پردازد. در بخش ٤ سودمندی الگوریتم پیشنهادی با استفاده ازشبیه‌سازی ارزیابی شده و نتایج مقایسه با الگوریتم دیگر ارائه شده است. بخش ۵ به نتیجه‌گیری و ارائه پیشنهادات برای آینده پرداخته است.

۲. فرضیات و بیان مسئله

۲-۱ فرضیات و محدودیت‌ها

محیط در نظر گرفته شده در این مقاله یک چهارراه است (شکل ۱). چراغ راهنمایی به عنوان عامل جهت کنترل جریان ترافیک به کار می‌رود. هر یک از خیابان‌های منتهی به چهارراه به طول ۷۵۰ متر در نظر گرفته شده و دارای چهار لاین هستند. در لاین راست رانندگان می‌توانند حرکت مستقیم یا گردش به راست انجام دهند. در دو لاین میانی فقط حرکت مستقیم و در لاین چپ فقط گردش به چپ مجاز است. در این مسئله، برای سیگنال ترافیک مجموعا ۸ فاز تعریف شده که چهار فاز اصلی به نام فاز سبز و چهار فاز فرعی به نام فاز زرد هستند. این فازها در جدول یک نشان داده شده‌اند. E ، W، S و N به ترتیب مخفف بازوی شرقی، غربی، شمالی و جنوبی چهارراه هستند.



شکل ۱. یک چهارراه

برای فاز سبز، مدت زمان ۱۰ ثانیه و برای فاز زرد ٤ ثانیه در نظر گرفته شده است. در این مقاله لاین چپ توسط چراغ راهنمایی مختص به خود کنترل شده در حالی که سه لاین دیگر توسط یک چراغ راهنمایی مشترک کنترل می‌شوند.

۲-۲ بیان مسئله بهینه سازی

کنترل سیگنال به فرآیند انتخاب یک فاز مناسب از میان مجموعه از پیش تعریف‌شده فازها اشاره دارد. نظر به اینکه هدف اصلی این مقاله، افزایش جریان ترافیک عبوری از چهارراه و کاهش طول صف تشکیل شده در سناریوی ترافیک بالا و به هنگام وقوع تصادف و در ساعات اوج ترافیک است، بنابراین تابع هزینه بر اساس تعداد وسایل نقلیه موجود در صف در همه لاین‌ها به صورت مسئله (۱) تعریف می‌شود.

$q(t)$ تعداد تمام وسایل نقلیه در صف در تمام لاین‌ها در زمان t است.

$$Queue\ length = \int_t q(t)dt \qquad (۱)$$

اگر فرض کنیم $q_l(t)$ تعداد وسایل نقلیه در صف موجود در لاین l در زمان t باشد، مسئله بهینه‌سازی در این مقاله به صورت زیر تعریف شده است:

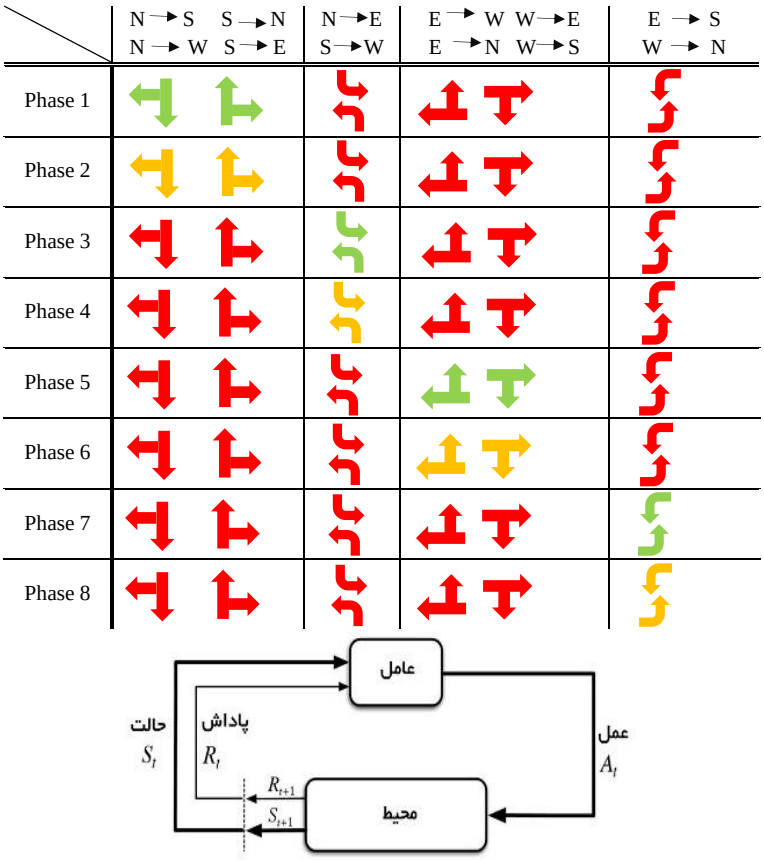
$$\underset{l\in L}{\text{minimize}}\, q_l(t) \qquad (۲)$$

L مجموعه همه لاین‌ها در چهارراه است. هدف این مسئله به حداقل رساندن حداکثر طول صف تشکیل شده در تمام لاین‌ها در زمان t می‌باشد.

۲-۳ روش حل مسئله: مروری بر روش یادگیری تقویتی عمیق

روش یادگیری تقویتی در واقع یادگیری یک سیاست بهینه است که پاداش تجمعی[۱۳] دریافت شده را بر اساس حالت فعلی سیستم به حداکثر می رساند. در واقع هدف عامل در این روش، بیشینه‌سازی پاداش تجمعی در طول زمان است. چرخه استاندارد یادگیری تقویتی در شکل ۲ نشان داده شده است.

جدول ۱. فازهای سیگنال ترافیک



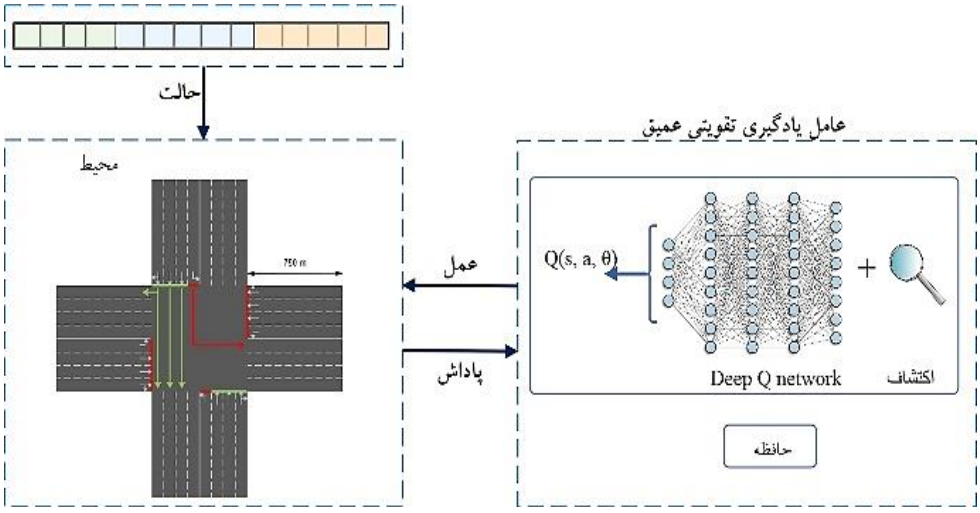
شکل ۲. چرخه روش یادگیری تقویتی

به این صورت است که در ابتدای گام زمانی t، عامل در تعامل با محیط، حالت فعلی محیط s_tرا مشاهده می‌کند. پس از مشاهده این حالت، عامل عمل a_t را انجام می دهد و سیگنال ترافیکی جهت کنترل چراغ راهنمایی فعال می‌شود. در اثر این سیگنال کنترلی و به علت حرکت وسایل نقلیه، حالت محیط به s_{t+1} تغییر می‌کند و در پایان گام زمانی، عامل سیگنال پاداش r_{t+1} را دریافت می‌کند. با استفاده از سیگنال پاداش دریافتی و بر اساس یک معیار عملکرد، عامل از مناسب بودن یا نامناسب بودن عمل انجام شده آگاه خواهد شد. در روش یادگیری تقویتی به هر عمل انجام شده توسط عامل، یک مقدار ارزش به نام مقدار Q اختصاص می‌یابد. اگر عامل مقادیر Q بهینه حالات متوالی را بداند، سیاست بهینه صرفاً انتخاب عملی است که بالاترین پاداش تجمعی را به همراه دارد. مقادیر Q بهینه به ازای انجام عمل a در حالت s که با $Q^*(s,a)$ نشان داده می‌شود، با استفاده از معادله بازگشتی معروف به معادله بهینه بلمن[۱۴] به‌دست می‌آید:

[Sutton and Barto, 2018]:

$$Q^*(s,a) = \mathbb{E}\{R_t + \gamma max_{\acute{a}} Q^*(S_{t+1}, \acute{a}) | S_t = s, A_t = a\} \quad (۳)$$
$$for\ all\ s \in \mathcal{S}, a \in \mathcal{A}$$

بر اساس این معادله، مقدار ارزش بهینه برای عمل a در حالت s شامل پاداش آنی (R_t) است که عامل پس از انجام عمل a در حالت s دریافت کرده به علاوه پاداش آینده بهینه‌ای که پس از آن دریافت می‌کند. عبارت $max_{\acute{a}}$ به معنی انتخاب با ارزشترین عمل از بین مجموعه اعمال ممکن در حالت S_{t+1} است. شایان ذکر است که حل معادله (۳) مستلزم محدود بودن حالت‌ها و مشخص بودن احتمالات گذار است. ولی در محیط‌های ترافیکی پیچیده، حالت‌های متعددی وجود دارد، به طوریکه محاسبه مقدار ارزش Q



شکل ۳. مدل یادگیری تقویتی عمیق برای کنترل چراغ راهنماییبرای هر جفت حالت–عمل بسیار دشوار است. بنابراین کنترل سیگنال ترافیک به عنوان یک مسئله یادگیری تقویتی عمیق فرموله شده که در شکل ۳ قابل مشاهده است. مکانیزم یادگیری در این مقاله، روش یادگیری عمیق Q[۱۵] است. در این رویکرد، معادله (۳) مستقیماً حل نمی‌شود، بلکه از یک شبکه عصبی عمیق پارامتری (DNN) برای تقریب مقادیر Q بهینه $Q^*(s,a)$ توسط تابع یادگیری Q[۱۶] استفاده می‌گردد. این تابع جهت به‌روزرسانی مقادیر Q به شرح زیر است:

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha((r_{t+1} + \gamma max_a Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)) \quad (٤)$$

در این معادله (s_t,a_t) و (s_{t+1},a_{t+1}) به ترتیب زوج‌های حالت–عمل فعلی و بعدی را نشان می‌دهند. α نرخ یادگیری و r_{t+1} پاداشی است که از انجام عمل a_t در حالت s_t به دست آمده است. γ ضریب تخفیف است که مقداری بین صفر و یک داشته و بیان کننده میزان اهمیت عامل به پاداش‌های آینده است. پس از اتمام دوره آموزش عامل، بهینه‌ترین عمل a_t برای انجام دادن وقتی در حالت s_tهستیم، عملی است که مقدار $Q(s_t,a_t)$ را بیشینه کند.

۳. ساخت و تحلیل مدل یادگیری تقویتی عمیق پیشنهادی

در توسعه یک سیستم کنترل چراغ راهنمایی با استفاده از رویکرد یادگیری تقویتی، حالت، عمل و تابع پاداش باید تعریف گردند. انتخاب هر یک از این عناصر نقش به‌سزایی در عملکرد سیستم کنترلی دارد. در ادامه این بخش به معرفی این عناصر در مدل پیشنهادی می‌پردازیم.

۳-۱ عمل

بر اساس حالت فعلی محیط، عامل باید عمل مناسب را انجام دهد. عمل تعریف شده در این مقاله، انتخاب فاز سبز از مجموعه از پیش تعیین‌شده‌ای از فازهاست. این مجموعه از پیش تعیین شده به صورت {NSA, NSLA, EWA, EWLA} است. NSA به معنی انتخاب فاز سبز برای وسایل نقلیه‌ای است که در بازوهای شمالی و جنوبی چهارراه بوده و قصد حرکت مستقیم یا چرخش به راست را دارند. به همین ترتیب EWA به انتخاب فاز سبز برای وسایل نقلیه موجود در بازوی غربی و شرقی چهارراه اشاره دارد که حرکت مستقیم یا گردش به راست دارند. NSLA به معنی فعال بودن فاز سبز برای خودروهایی است که در بازوهای شمال و جنوب به چپ می‌چرخند. همین امر در مورد وسایل نقلیه بازوهای راست و چپ که قصد چرخش به چپ را دارند برای عمل EWLA صادق است. همانطور که قبلا

ذکر شد، زمان فاز سبز و فاز زرد به ترتیب برابر با ۱۰ ثانیه و ٤ ثانیه است. اگر عمل انتخابی در زمان t با عمل انتخابی در زمان t-1 متفاوت باشد، یک فاز زرد ٤ ثانیه‌ای بین دو عمل فعال شده در غیر این صورت، فاز زرد وجود ندارد و فاز سبز فعلی ادامه خواهد یافت [Vidali et al, 2019].

۳-۲ پاداش

پاداش در مدل یادگیری تقویتی به منزله فیدبک برای ارزیابی اعمال انجام شده پیشین است. انتخاب پاداش مناسب برای اطمینان از موفقیت‌آمیز بودن فرآیند یادگیری و دستیابی به استراتژی بهینه بسیار مهم است. پاداش باید بر اساس معیاری از کارایی سیگنال ترافیک تعریف شود. نظر به اینکه هدف اصلی این مقاله کاهش طول صف تشکیل شده در ترافیک با حجم زیاد و همچنین در شرایط وقوع تصادف است، زمان انتظار وسایل نقلیه که معادل با طول صف بوده، به عنوان معیار عملکرد در نظر گرفته می‌شود. بنابراین، پاداش به صورت کاهش زمان انتظار تجمعی بین دو عمل متوالی انجام شده تعریف می‌گردد. تابع پاداش به صورت زیر است:

$$r_t = wt_{t-1} - wt_t \qquad (۵)$$

که در آن:

$$wt_t = \sum_{i=1}^{N} w_{i,t} \qquad (٦)$$

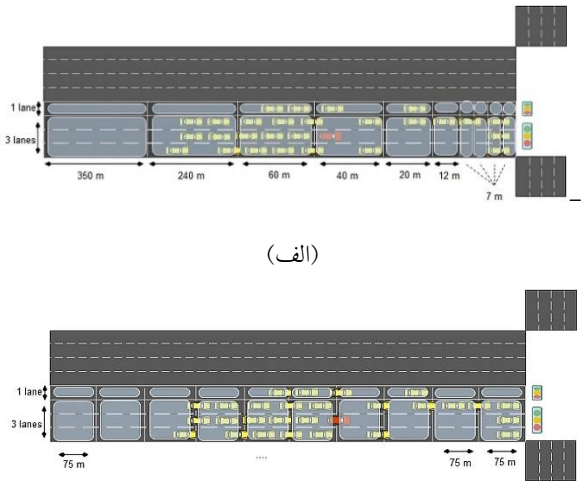
زمان انتظار وسیله نقلیه i در مرحله زمانی t با $w_{i,t}$ نشان داده می‌شود. در اینجا، زمان انتظار به زمانی بر حسب ثانیه اشاره دارد که یک وسیله نقلیه سرعت کمتر از ۰/۱ متر بر ثانیه دارد. N تعداد کل وسایل نقلیه در مرحله زمانی t است. انتخاب عمل مناسب باعث می شود در زمان t وسایل نقلیه کمتری نسبت به زمان قبلی t-1 درصف وجود داشته باشند و زمان انتظار برای وسایل نقلیه کاهش یابد. به این ترتیب، پاداش در طول زمان افزایش می‌یابد و مثبت بودن تابع پاداش نشاندهنده انجام یک عمل مناسب توسط عامل است.

۳-۳ حالت

به منظور کارآیی هرچه بیشتر الگوریتم، حالت انتخابی باید درک کافی از چهارراه به ویژه در مورد توزیع وسایل نقلیه در هر خیابان، در اختیار عامل قرار دهد. در این مقاله با الهام از رویکرد گسسته‌سازی معرفی شده در [Genders and Razavi, 2016]، ما خیابان‌ها را به دو صورت سلول‌بندی کرده و تاثیر هر دو را مورد ارزیابی قرار می‌دهیم. در روش اول طول سلول‌ها را مساوی در نظر گرفته و در روش دوم طول سلول‌ها با هم متفاوت هستند، به طوری که هر چه از چراغ راهنمایی فاصله بگیریم اندازه سلول بزرگتر می‌شود. این دو نحوه سلول‌بندی در شکل ٤ نشان داده شده است. توجه شود که در رویکرد گسسته‌سازی در [Vidali et al, 2019] اگرچه طول سلول‌ها با هم متفاوت است ولی در هر سلول فقط اطلاعات حضور یا عدم حضور خودرو به عنوان حالت در نظر گرفته شده است. این در حالیست که در شرایط حجم بالای ترافیک و یا در صورت بروز اختلالی مانند تصادف، این اطلاعات جهت درک محیط کافی نیست. از طرفی یکی از پارامترهای متاثر از وقوع تصادف و یا حجم بالای ترافیک، تعداد وسایل نقلیه موجود در صف است. برای انعطاف‌پذیری بیشتر الگوریتم، این اطلاعات باید به عنوان مشاهده‌ای از محیط در اختیار عامل قرار گیرد. در نتیجه، به طور مثال بردار $E = [1,1,0,1,0, \dots ,1]^T$ نشاندهنده حضور یا عدم حضور وسایل نقلیه است. بردار $Q = [5,3,0,2,2, \dots ,1]^T$ نیز برای نشان دادن تعداد وسایل نقلیه موجود در صف در هر سلول استفاده می‌شود. ابعاد این بردارها برابر است با تعداد سلول‌های تعریف شده برای چهارراه. علاوه بر این، فاز سبز فعلی چراغ راهنمایی به عنوان یک بردار چهار بعدی one-hot مانند $P = [1,0,0,0]^T$ کدگذاری می‌شود زیرا برای عامل، چهار فاز سبز تعریف شده است. هنگامی که چراغ سبز برای یک فاز روشن می‌شود، ورودی مربوطه در این بردار، مقدار یک می‌گیرد. در نتیجه، در این مقاله، بردار حضور یا عدم حضور وسایل نقلیه به همراه بردار تعداد خودروهای موجود در صف در هر سلول و فاز فعلی چراغ راهنمایی فعلی،

به عنوان بردار حالت به صورت $S = [E^T, Q^T, P^T]^T$ به هم پیوسته‌اند.

۳-٤ ساختار شبکه عصبی عمیق

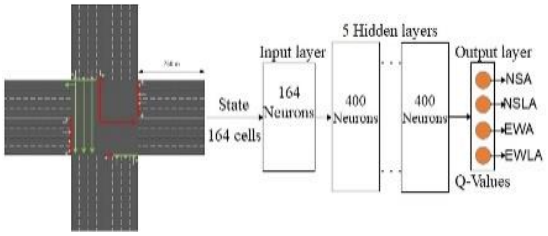


(الف)

(ب)

شکل ٤. گسسته سازی خیابان (الف) اندازه سلول ها متفاوت (ب) اندازه سلول ها یکسان

در این مقاله، از یک شبکه کاملاً متصل برای تخمین مقادیر Q برای همه عمل‌های قابل انجام $a \in A$ در حالت مشاهده شده S استفاده می‌شود. یک نمایش بصری برای این شبکه در شکل ۵ نشان داده شده است.

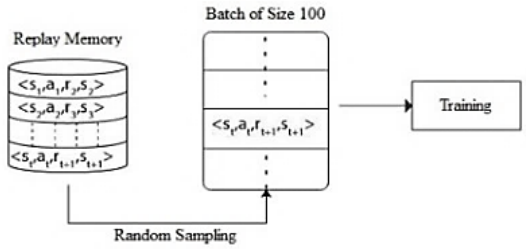


شکل ۵. ساختار شبکه عصبی عمیق

ورودی این شبکه، حالت مشاهده شده از چهارراه یعنی s_t بوده و به عنوان خروجی، یک بردار شامل مقادیر ارزش تخمین زده شده یعنی $Q(s, a, \theta)$ تولید می‌گردد. θ پارامتر شبکه است. پارامترها باید به صورتی آموزش داده شوند که $Q(s, a, \theta)$ به مقدار بهینه $Q^*(s_t, a_t)$ همگرا شود. در شکل ۵، ترکیبی از بردارهای حضور وسیله نقلیه، تعداد وسایل نقلیه در صف و بردار فاز چراغ راهنمایی، ورودی‌های لایه اول شبکه هستند. بنابراین، لایه ورودی از ١٦٤ نورون تشکیل شده است. شبکه شامل پنج لایه پنهان با ٤٠٠ نورون با تابع فعال‌سازی غیرخطی یکسوکننده (ReLU) و یک لایه خروجی متشکل از چهار نورون با عملکرد فعال‌سازی خطی است. نورون‌های لایه خروجی بیانگر مقدار Q یک عمل در یک حالت هستند.

۳-۵ آموزش مدل

برای آموزش شبکه عمیق، از تکنیک‌های بازپخش تجربه و شبکه Q هدف استفاده می‌شود. در این روش، اطلاعات در دسته‌هایی[۱۷] شامل چندین نمونه به صورت $e = \{s_t, a_t, r_{t+1}, s_{t+1}\}$ طبقه‌بندی شده و از آن‌ها برای آموزش عامل استفاده می‌گردد. از طریق تکنیک بازپخش تجربه، همبستگی بین نمونه‌ها حذف شده و پارامترهای شبکه Q فعلی به شبکه هدف کپی می‌گردد. در هر نمونه آموزشی، عامل برای یادگیری پارامترهای شبکه عصبی عمیق به داده‌های آموزشی شامل ورودی‌های $e = \{s_t, a_t, r_{t+1}, s_{t+1}\}$و خروجی‌های مقادیر Q شبکه هدف یعنی $Q^*(s_t, a_t)$ نیاز دارد. داده‌های ورودی برای آموزش به طور تصادفی از حافظه بازیابی شده و هر نمونه در هر دسته برای آموزش به کار گرفته می‌شود (شکل ٦).



شکل ٦. نحوه نمونه برداری از حافظه

مقادیر Q هدف یعنی $Q^*(s_t, a_t)$ با معادله (۷) تخمین زده می‌شوند:

$$Q^*(s_t, a_t) = r_{t+1} + \gamma max_{a_{t+1}} \acute{Q}(s_{t+1}, a_{t+1}, \theta) \quad (۷)$$

که در آن $\acute{Q}(s_{t+1}, a_{t+1}, \theta)$ خروجی یک شبکه هدف جداگانه با ساختاری مشابه با شبکه عصبی عمیق اصلی بوده و نشان‌دهنده مقدار Q مرتبط با یک عمل بعد از انجام عمل a_t در حالت s_t می‌باشد. برای به‌روزرسانی مقادیر Q تخمین‌زده شده، معادله (۷) از اطلاعات موجود در هر نمونه استفاده می‌کند. برای هر نمونه آموزشی، تعداد نمونه‌های بازیابی شده از حافظه، اندازه دسته را تشکیل می‌دهد. حافظه برای ذخیره این دسته‌ها استفاده شده و آنها را در مراحل زمانی مختلف بر روی عامل اعمال می‌کند. این حافظه دارای ظرفیت محدودی برای ذخیره نمونه‌هاست. نمونه های قدیمی با پر شدن حافظه کنار گذاشته خواهند شد. پارامترهای شبکه باید به گونه‌ای آموزش داده شوند که تابع میانگین مربعات خطای[۱۸] (MSE) زیر را به حداقل برسانند. در معادله (۸)، m سایز داده ورودی است. نظر به اینکه عامل فقط تعداد محدودی از حالت‌ها را تجربه کرده و نه کل فضای حالت را، مقادیر Q برای حالت‌های تجربه نشده، ممکن است.

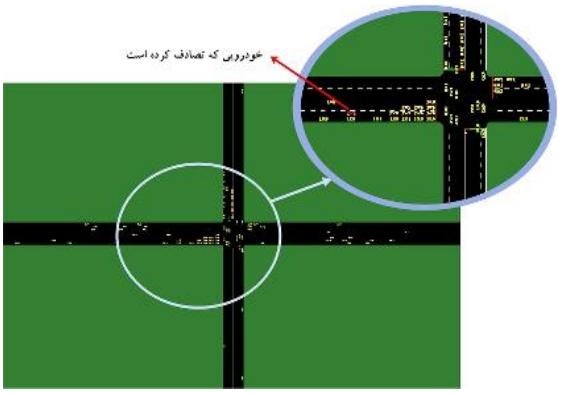
$$loss\ function = \frac{1}{m}\sum_{t=1}^{m}\{\left(r_{t+1} + \gamma max_{a_{t+1}} \acute{Q}(s_{t+1}, a_{t+1}, \theta)\right) - Q(s_t, a_t, \theta)\} \quad (۸)$$

دقیق نباشند. علاوه بر این، خود فضای حالت نیز ممکن است دائماً در حال تغییر باشد که منجر به ناکارآمدی مقادیر Q تخمین زده شده قبلی خواهد شد. بنابراین، عامل باید تصمیم بگیرد که آیا از مقادیر Q آموخته شده قبلی استفاده کند و عملی را با بالاترین مقدار Q استخراج[۱۹] کند یا سایر عمل‌های ممکن را نیز بررسی کند. در این مقاله از روش ε-greedy به عنوان سیاست انتخاب عمل، ε تعریف شده به این صورت که عامل با احتمال ε یک عمل اکتشافی[۲۰] را انجام داده و با احتمال ε –۱ یک عمل استخراجی را با بالاترین مقدار Q تخمین زده شده انتخاب می‌کند. در ابتدای آموزش ε=1 بوده یعنی عامل فقط اقدامات اکتشافی را انجام می‌دهد. پس از یک دوره آموزشی مشخص، عملیات استخراجی آغاز می‌شود. فرمول‌بندی روش ε-greedy به صورت زیر است:

$$\varepsilon = 1 - \frac{current\ episode}{total\ number\ of\ episodes} \quad (۹)$$

٤. اعتبارسنجی و تحلیل نتایج

در این بخش، نتایج شبیه‌سازی برای ارزیابی رویکرد پیشنهادی شرح داده شده است. ارزیابی‌ها با استفاده از شبیه‌ساز SUMO انجام شده که در آن ترافیک بلادرنگ قابل شبیه‌سازی است. شکل ۷ محیط مورد مطالعه برای ارزیابی را به همراه داده‌های شبیه‌سازی شده در SUMO نشان می‌دهد. در این مقاله، یک چهارراه مورد تحلیل قرار گرفته است. پیکربندی‌های مختلف شبکه و توزیع‌های دیگر برای ورودی ترافیکی در کارهای آینده مورد بررسی قرار می‌گیرد. ابتدا چگونگی تولید ترافیک و توزیع آن توضیح داده شده و سپس مدل پیشنهادی در دو بخش ارزیابی می‌گردد. همچنین عملکرد دو رویکرد گسسته‌سازی تعریف شده از لحاظ اندازه سلول‌ها، بر اساس معیارهای عملکرد مختلف در دو بخش مقایسه خواهند شد.

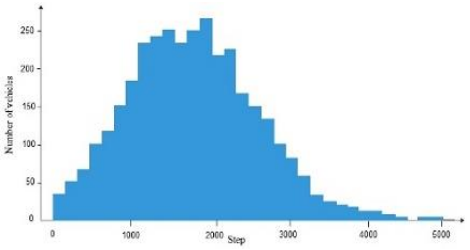


شکل ۷. شبیه‌سازی در سیمولاتور SUMO

در بخش اول یک سناریوی ترافیک با حجم بالا را بررسی می‌کنیم. در بخش دوم با شبیه‌سازی یک تصادف در یکی از لاین‌ها، انعطاف‌پذیری کنترلگر پیشنهادی را ارزیابی خواهیم کرد.

٤-١ تولید ترافیک

در مطالعات مختلف برای مدل‌سازی جریان واقعی ترافیک چندین توزیع احتمال پیشنهاد شده است [Maurya, Dey, and Das, 2015]. برای نزدیک شدن به شرایط واقعی ترافیک، در این مقاله از توزیع ترافیک موجود در [Vidali et al. 2019] استفاده می‌کنیم که تقاضای سفر بر اساس توزیع Weibull است. شکل ۸ ترافیک ایجاد شده در یک اپیزود یادگیری را با تعداد وسایل نقلیه در هر گام شبیه‌سازی نشان می‌دهد. این توزیع، جریان ترافیک را در طول یک روز کامل شبیه‌سازی کرده است. زیرا در ابتدا تعداد وسایل نقلیه افزایش می‌یابد که نشان دهنده رسیدن به ساعات اوج ترافیک است.



شکل ۸. تولید جریان ترافیک

سپس با گذشت زمان، تعداد خودروهای وارد شده کاهش می‌یابد، که این امر تضعیف تدریجی تراکم ترافیک را توصیف می‌کند. در هر اپیزود، مسیر مبدا و مقصد خودروهای تولید شده به صورت تصادفی است. لازم به ذکر است که جریان ترافیک در چهارراه‌های مختلف ممکن است در واقعیت به طور قابل توجهی متفاوت باشد. بنابراین، نتایج ارائه شده در این مقاله نشان دهنده عملکرد الگوریتم در ترافیک یک روز معمولی است.

٤-٢ تنظیمات سیمولاتور و پارامترها

در این مقاله عامل در ۵۰ اپیزود آموزش داده می‌شود. هر اپیزود شامل ٥٤٠٠ گام بوده و فرکانس زمانی ارائه شده توسط SUMO یک ثانیه در هر گام است. لذا مدت زمان هر اپیزود ۹۰ دقیقه است. مقادیر پارامترهای مورد استفاده در جدول ۲ بیان شده‌اند.

٤-٣ اعتبارسنجی مدل در سناریوی ترافیک با حجم بالا

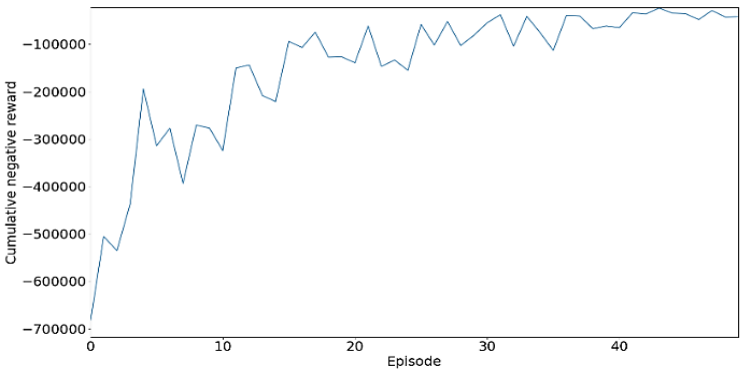
در این سناریو، ۳۰۰۰ وسیله نقلیه به چهارراه نزدیک می‌شوند. از این تعداد، ۷۵ درصد از وسایل نقلیه مستقیماً از تقاطع عبور کرده و ۲۵ درصد به چپ یا راست می‌پیچند [Vidali et al, 2019]. در انتخاب حالت محیط به هنگام گسسته‌سازی خیابان، هر دو شرایط یکسان بودن و متفاوت بودن اندازه سلول‌ها بررسی می‌گردد.

جدول ۲. پارامترهای روش یادگیری تقویتی

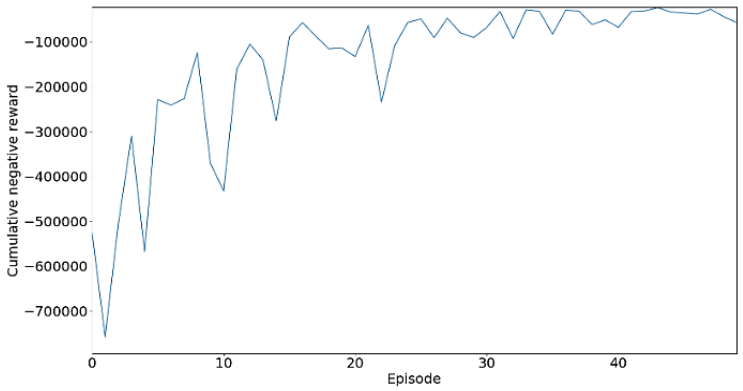
پارامتر	مقدار
ضریب تخفیف γ	۰/۷۵
نرخ یادگیری	۰/۰۰۱
اندازه حافظه	۵۰۰۰
اندازه دسته	۱۰۰
مقدار ε برای شروع	۱

همچنین، رویکرد پیشنهادی با عملکرد [Vidali et al, 2019] در سناریوی ترافیک بالا مقایسه می‌شود. شکل ۹ منحنی پاداش منفی تجمعی را در طول آموزش برای شرایطی که اندازه سلول‌ها یکسان باشد نشان داده و شکل ۱۰ همین منحنی را برای شرایط متفاوت بودن اندازه سلول‌ها نشان می‌دهد. می‌توان مشاهده کرد که عامل در ترافیک با حجم بالا در هر دو شرایط یکسان بودن

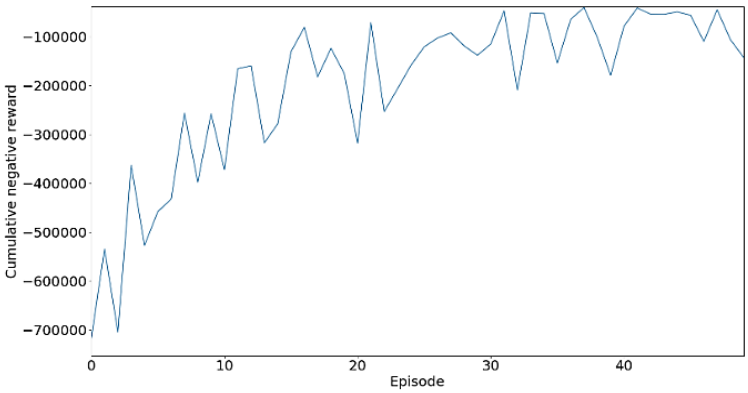
و متفاوت بودن اندازه سلول‌ها به خوبی آموزش دیده و با پیشرفت آموزش، پاداش تجمعی افزایش یافته است. علاوه بر این، منحنی پاداش نشان می‌دهد که یک سیاست پایدار آموخته شده است. یعنی نوسانی بین انتخاب اعمال بد و خوب وجود ندارد و الگوریتم به انتخاب تصمیمات نادرست منحرف نمی-شود. از سوی دیگر، رویکرد پیشنهادی با سرعت مطلوبی به سیاست بهینه همگرا می‌شود. شکل ۱۱ نیز منحنی پاداش را در رویکرد [Vidali et al, 2019] در سناریوی ترافیک با حجم بالا ترسیم کرده است که این منحنی نشان می‌دهد که حالت در نظر گرفته‌شده برای عامل در رویکرد [Vidali et al, 2019] برای درک محیط در ترافیک بالا کافی نیست. زیرا اگرچه پاداش تجمعی افزایش یافته ولی سیاست حاصل شده به دلیل وجود نوسانات، روند ناپایداری را از خود به نمایش گذاشته است و حتی با گذشت ۵۰ اپیزود از دوره آموزش، همچنان منحنی پاداش همگرا نشده است که نشان از عملکرد نامطلوب الگوریتم ارائه شده در مقاله [Vidali et al, 2019] در سناریوی ترافیک بالا دارد. این امر نشاندهنده لزوم توسعه حالت در ترافیک بالا بوده که در این مقاله به آن پرداخته شد. زیرا اگرچه اندازه سلول-ها در [Vidali et al, 2019] نیز متفاوت است ولی درک کم از محیط سبب تنزل کارآیی الگوریتم در ترافیک بالا شده است. برای مقایسه عملکرد الگوریتم پیشنهادی در دو وضعیت اندازه سلول‌ها، ماکزیمم طول صف تشکیل شده در نتایج تست در هر وضعیت و همچنین ماکزیمم طول صف در رویکرد [Vidali et al, 2019] در شکل ۱۲ نشان داده شده است. نتایج به دست آمده برتری الگوریتم پیشنهادی را نسبت به رویکرد [Vidali et al, 2019] در سناریوی ترافیک با حجم بالا نشان می‌دهد. در شرایط یکسان بودن اندازه سلول ها بیشترین طول صف تشکیل شده نسبت به [Vidali et al, 2019] ۱۱ درصد کاهش یافته و در شرایط متفاوت بودن اندازه سلول این مقداربه میزان حدودا ۲۸ درصد کاهش یافته است. این نتیجه اهمیت ورودی اعمال شده به الگوریتم را اثبات می‌کند. همچنین مشاهده می‌شود که وقتی اندازه سلول‌ها با هم متفاوت باشد، بیشترین طول صف تشکیل شده تقریبا به میزان ۱۹ درصد نسبت به یکسان بودن اندازه سلول‌ها کاهش یافته است. بنابراین اگر اندازه سلول‌ها متفاوت انتخاب شود به طوریکه سلول‌های نزدیک به چراغ راهنمایی کوچکتر از سلول‌های دورتر از چراغ باشند، نتیجه بهتری حاصل خواهد شد. این امر به این دلیل است که هر چه به چراغ راهنمایی نزدیکتر می‌شویم دریافت دقیق اطلاعات محیط اهمیت بیشتری می‌یابد. جدول ۳ مقایسه عددی بین عامل را در وضعیت‌های متفاوت بودن و یکسان بودن اندازه سلول و همچنین مقایسه با رویکرد [Vidali et al, 2019] در طی دوره آموزش را نشان می‌دهد. مقادیر میانگین پاداش‌های منفی تجمعی، تاخیرهای تجمعی و طول صف به عنوان معیارهای عملکرد استفاده شد. تاخیر به عنوان زمانی تعریف می‌شود که وسیله نقلیه بین دو عمل متوالی ثابت مانده است. مقایسه معیارهای عملکرد در جدول ۳ موثر بودن روش پیشنهادی را اثبات می‌کند. وقتی اندازه سلول‌ها متفاوت باشد، در مقایسه با روش [Vidali et al, 2019]، پاداش به طور متوسط به میزان تقریبا ۲۹ درصد افزایش یافته در حالی که طول صف و تاخیر به ترتیب تا حدود ١٤ و ۱۳ درصد کاهش یافته است.



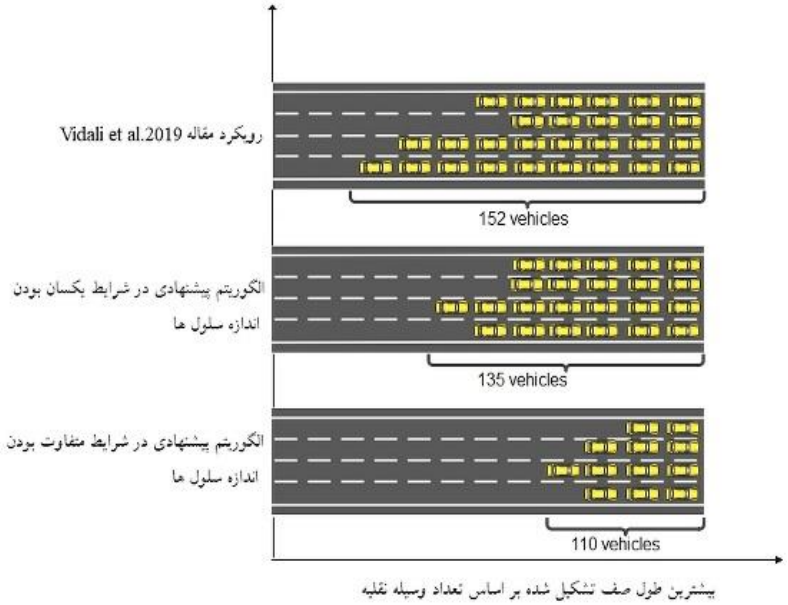
شکل ۹. منحنی پاداش در حین آموزش برای شرایط یکسان بودن اندازه سلول‌ها



شکل ۱۰. منحنی پاداش در حین آموزش برای شرایط متفاوت بودن اندازه سلول‌ها



شکل ۱۱. منحنی پاداش در حین آموزش برای [Vidali et al, 2019] در ترافیک با حجم بالا



شکل ۱۲. مقایسه بیشترین طول صف تشکیل شده در طول نتایج تست

جدول ۳. مقادیر عددی معیارهای عملکرد برای عامل در دوره آموزش در سناریوی ترافیک بالا

	روش پیشنهادی با سلول متفاوت	روش پیشنهادی با سلول یکسان	[Vidali et.al 2019]
متوسط پاداش منفی تجمعی	-۱٤٥۸۳۰	-۱٥۰٤٤۰	-۱۹۹۳۱۰
متوسط تاخیر تجمعی(ثانیه)	۳۹۷۲۹۰	٤۱۱۹٦۰	٤٥٦٤٥۰
متوسط طول صف (وسیله نقلیه)	۷۳	۷۷	۸٥

٤-٤ اعتبارسنجی انعطاف‌پذیری مدل پیشنهادی در حضور تصادف

برای بررسی اثرات یک تصادف، در این مقاله سعی می‌کنیم در ترافیک با حجم بالا بسته شدن مسیر ناشی از تصادف را در SUMO شبیه‌سازی کنیم. بدین منظور، یکی از خودروهایی که به چراغ راهنمایی نزدیک می‌شود، قبل از تقاطع در یکی از لاین-ها متوقف می‌شود تا توقف خودرو ناشی از تصادف را شبیه-سازی کرده باشد. این تصادف می‌تواند در هر مکانی از محیط و در هر زمانی رخ دهد. با انجام این کار، عامل در شرایط وقوع تصادف آموزش دیده و یاد می‌گیرد که در صورت بروز اختلال در محیط، سیگنال کنترلی را بهینه کند. ابتدا یکی از خودروهایی که در یکی از لاین‌های بازوی غربی به چهارراه نزدیک می‌شود را در فاصله ۵۰ متری از چراغ متوقف می‌کنیم. شکل ۱۳ منحنی پاداش را در طول آموزش برای شرایطی که اندازه سلول‌های تعریف شده یکسان باشد نشان داده و شکل ١٤ همین منحنی را برای شرایط متفاوت بودن اندازه سلول‌ها نشان می‌دهد. مشاهده می‌شود حتی در صورت وقوع تصادف عامل عملکرد خوبی را از خود به نمایش گذاشته و پاداش در طول دوره آموزش افزایش یافته و همگرا شده است. این نتایج نشان‌دهنده انعطاف‌پذیری الگوریتم پیشنهادی در برابر تصادفات احتمالی و در نتیجه کاهش تراکم ترافیک است. همچنین با مقایسه منحنی پاداش در دو شکل ۱۳ و ١٤، دیده می‌شود که در حالت متفاوت بودن اندازه سلول‌ها نوسانات موجود در منحنی کمتر بوده که نشان از دستیابی به یک سیاست پایدارتر دارد. بنابراین متفاوت بودن اندازه سلول‌ها در رویکرد گسسته‌سازی عملکرد مطلوب‌تری را به همراه دارد. برای نشان دادن کارایی مدل پیشنهادی در حضور تصادف، مقایسه‌ای بین رویکرد پیشنهادی در دو وضعیت یکسان بودن و متفاوت بودن اندازه سلول، با [Vidali et al, 2019] انجام

شده است. در [Vidali et al, 2019]، کنترل ترافیک بدون ایجاد اختلال در محیط انجام شده و چراغ راهنمایی بدون وقوع حادثه آموزش دیده است. بنابراین با توقف یک خودرو، تصادفی را در محیط آن در ترافیک با حجم بالا ایجاد کرده و کنترل کننده را در شرایط جدید آموزش می‌دهیم. شکل ۱۵ ماکزیمم طول صف ایجاد شده را در طول دوره تست نشان می‌دهد. مشاهده می‌شود الگوریتم پیشنهادی با اندازه متفاوت برای سلول‌ها بیشترین طول صف ایجاد شده را به میزان ۲۳ درصد نسبت به وضعیت یکسان بودن اندازه سلول‌ها و ۲۸ درصد نسبت به رویکرد [Vidali et al, 2019] کاهش داده است. این نتیجه، برتری الگوریتم پیشنهادی با وضعیت متفاوت بودن اندازه سلول-ها را در توسعه یک سیاست پایدار و انعطاف‌پذیر اثبات می‌کند. برای ارزیابی بیشتر مدل پیشنهادی دو تصادف را در دو مکان مختلف، یکی را در فاصله ۷۰۰ متری چراغ راهنمایی در خیابان جنوبی و تصادف دیگر را در ۷۳۰ متری چراغ در خیابان شمالی چهارراه به صورت شبیه‌سازی کرده به طوریکه زمان این دو تصادف با هم هپوشانی دارد و کارآیی الگوریتم را بررسی می-کنیم. شکل ١٦ منحنی پاداش چراغ راهنمایی در طول دوره آموزش را در صورت وقوع همزمان دو تصادف در حالت متفاوت بودن اندازه سلول‌ها نشان می‌دهد. مشاهده می شود که الگوریتم کارایی خود را حفظ کرده و منحنی پاداش در طول دوره آموزش در کمتر از ۵۰ اپیزود، به همگرایی رسیده است.

۵. نتیجه‌گیری

در این مقاله، یک چارچوب یادگیری تقویتی عمیق برای بهبود کارایی جریان ترافیک در حضور تصادف و ترافیک با حجم بالا ارائه شده است. برای کارآیی الگوریتم در هنگام نوسانات تقاضای ترافیک، از رویکرد گسسته‌سازی جاده استفاده شده و حالت مشاهده شده توسط عامل توسعه داده شد. برای انعطاف‌پذیری کنترل‌کننده در برابر اختلالات احتمالی محیط، عامل در حضور تصادفات مورد آموزش قرار گرفت. نتایج شبیه-سازی نشان می‌دهد که الگوریتم یک استراتژی پایدار را بدون توجه به مکان و زمان تصادف آموخته است. وضعیت متفاوت بودن اندازه سلول‌ها برتری خود را در کاهش طول صف نسبت به وضعیت یکسان بودن اندازه سلول اثبات کرده و در مقایسه با [Vidali et al, 2019] نیز طول صف حتی در حضور تصادف کاهش یافت. همچنین همگرایی با مدل پیشنهادی نسبت به رویکرد [Vidali et al, 2019] با سرعت بیشتری حاصل می-شود. توسعه الگوریتم به شبکه چند عاملی شامل چندین چهارراه و در نظر گرفتن حالت همسایگان از طریق شبکه‌های عصبی مبتنی بر گراف از کارهای آینده نویسندگان می‌باشد.

٦. پی‌نوشت‌ها

1. Fixed timing
2. Adaptive traffic signal control
3. Real time
4. Reinforcement learning
5. Markov decision process
6. Uncertain
7. Agent
8. State
9. Resilient
10. Fully connected
11. Experience replay
12. Application programming interface
13. Cumulative reward
14. Bellman optimality equation
15. Deep Q-learning
16. Q learning function
17. Batch
18. Mean Square Error
19. Exploitation
20. Exploration

۷. مراجع

– مهماندار، م.، آریانا، م.، مبادری، ت. و خلیلی، ا. (۱۳۹۹) " ارزیابی مولفه های موثر بر ارتقای فرهنگ ایمنی ترافیک و کاهش تلفات با موتورسیکلت"، فصلنامه مهندسی حمل و نقل، سال یازدهم، شماره سوم، ص.٦٤٩-٦٦٣.

- Bálint, K., Tamás, T. and Tamás, B. (2022) "Deep Reinforcement Learning based approach for Traffic Signal Control". Transportation Research Procedia. Vol. 62, pp. 278–285.

- Casas N. (2017) "Deep deterministic policy gradient for urban traffic light control". arXiv preprint arXiv:170309035.

- Chu, K.F., Lam, A.Y. and Li, V.O. (2021) "Traffic Signal Control Using End-to-End Off-Policy Deep Reinforcement Learning". IEEE Transactions on Intelligent Transportation Systems.

- Chu, T., Wang, J., Codecà, L. and Li, Z. (2019) "Multi-agent deep reinforcement learning for large-scale traffic signal control". IEEE Transactions on Intelligent Transportation Systems. Vol. 21, No. 3, pp. 1086–1095.

- Essa, M. and Sayed, T. (2020) "Self-learning adaptive traffic signal control for real-time safety optimization". Accident Analysis & Prevention. Vol. 146:105713.

- Gao, J., Shen, Y., Liu, J., Ito, M. and Shiratori, N. (2017) "Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network". arXiv preprint arXiv:170502755.

- Genders, W. and Razavi., S. (2016) "Using a deep reinforcement learning agent for traffic signal control". arXiv preprint arXiv:161101142.

- Gong, Y., Abdel-Aty, M., Yuan, J. and Cai, Q. (2020) "Multi-objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control". Accident Analysis & Prevention. Vol. 144:105655.

- Jamil. ARM., Ganguly. KK. and Nower. N (2021) "Adaptive traffic signal control system using composite reward architecture based deep reinforcement learning". IET Intelligent Transport Systems. Vol. 14, No. 14, pp. 2030–2041.

- Krajzewicz, D., Erdmann, J., Behrisch, M. and Bieker, L. (2012) "Recent development and applications of SUMO-Simulation of Urban MObility". International journal on advances in systems and measurements. Vol. 5(3 & 4).

- Li, L., Lv, Y. and Wang, F.Y. (2016) "Traffic signal timing via deep reinforcement learning". IEEE/CAA Journal of Automatica Sinica. Vol. 3, No.3, pp. 247–254.

- Li, M., Li, Z., Xu, C. and Liu, T. (2020) "Deep reinforcement learning-based vehicle driving strategy to reduce crash risks in traffic oscillations. Transportation research record". Vol. 2674, No. 10, pp. 42–54.

- Liang, X., Du, X., Wang, G. and Han, Z. (2018) "Deep reinforcement learning for traffic light control in vehicular networks". arXiv preprint arXiv:180311115.

- Liang, X., Du, X., Wang, G and Han, Z. (2019) "A deep q learning network for traffic lights' cycle control in vehicular networks". IEEE Transactions on Vehicular Technology. Vol. 68, No. 2, pp. 1243–1253.

- Maurya, A.K., Dey, S. and Das, S. (2015) "Speed and time headway distribution under mixed traffic condition". Journal of the Eastern Asia Society for Transportation Studies. Vol, 11. pp. 1774–1792.

- Mousavi, S.S., Schukat M. and Howley, E. (2017) "Traffic light control using deep policy-gradient and value-function-based reinforcement learning". IET Intelligent Transport Systems. Vol. 11, No. 7, pp. 417–423.

- Paul, A. and Mitra, S. (2022) "Exploring reward efficacy in traffic management using deep reinforcement learning in intelligent transportation system". ETRI Journal. Vol. 44, No. 2, pp. 194–207.

- Rodrigues, F. and Azevedo, C.L. (2019) "Towards robust deep reinforcement learning for traffic signal control: Demand surges, incidents and sensor failures". IEEE intelligent transportation systems conference (ITSC). Auckland, New Zealand: 27-30 October 2019.

- Sutton, R.S.and Barto, A.G. (2018) "Reinforcement learning: An introduction". MIT press.

- Van der Pol, E., Oliehoek, F.A. (2016) "Coordinated deep reinforcement learners for traffic light control". Proceedings of Learning, Inference and Control of Multi-Agent Systems (NIPS).

- Vidali, A., Crociani, L., Vizzari, G. and Bandini, S. (2019) "A Deep Reinforcement Learning Approach to Adaptive Traffic Lights Management". In: WOA. Parma, Italy, 26-28 June 2019.

- Wang. T., Cao. J. and. Hussain A (2021) "Adaptive Traffic Signal Control for large-scale scenario with Cooperative Group-based Multi-agent reinforcement learning". Transportation research part C: emerging technologies. Vol. 125:103046.

- Wei. H., Zheng, G., Gayah, V. and Li, Z. (2021) "Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation". ACM SIGKDD Explorations Newsletter. Vol. 22, No. 2, pp. 12–18.

- 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 & Data Mining. pp. 2496–2505.

- Yoon, J., Ahn, K., Park, J and Yeo, H. (2021) "Transferable traffic signal control: Reinforcement learning with graph centric state representation". Transportation Research Part C: Emerging Technologies. Vol. 130:103321.

- Roy, A., Hossain, M. and Muromachi, Y. (2022) "A deep reinforcement learning-based intelligent intervention framework for real-time proactive road safety management". Accident Analysis & Prevention. Vol. 165, pp. 106512.

- 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:190504716.

زهرا زینلی مدرک کارشناسی در رشته مهندسی برق با گرایش الکترونیک را از دانشگاه شهید چمران و مدرک کارشناسی ارشد در رشته مهندسی برق گرایش کنترل را از دانشگاه تربیت مدرس اخذ نمود و در حال حاضر دانشجوی دکترای مهندسی برق گرایش کنترل در دانشگاه تربیت مدرس می‌باشد. زمینه‌های پژوهشی مورد علاقه ایشان، کنترل و بهینه‌سازی با روش‌های نوین، سیستم‌های حمل و نقل هوشمند ،تئوری یادگیری ماشین و سیسنم‌های چندعاملی می‌باشد.

مهدی سجودی مدرک کارشناسی مهندسی برق با گرایش الکترونیک خود را در سال ۱۳۸۱ از دانشگاه ارومیه دریافت کرد. همچنین مدارک کارشناسی ارشد و دکتری مهندسی برق با گرایش کنترل را در سالهای ۱۳۸٤ و ۱۳۹۰ از دانشگاه تربیت مدرس، دریافت نمود. از سال ۱۳۹۱ بعنوان هیات علمی به گروه کنترل دانشکده مهندسی برق و کامپیوتر دانشگاه تربیت مدرس پیوست و در حال تدریس دروس نظریه کنترل و بهینه سازی می باشند. علایق تحقیق وی در زمینه کنترل و بهینه سازی در سیستم های چند عامل و سیستمهای پیچیده، زیست شناسی سامانه ای، ، سیستم های سایبر–فیزیکی و حوزه های چند تخصصی نو آورانه می باشد.