

استراتژی تصمیم‌گیری در بزرگراه برای خودروی خودران جهت انجام مانور سبقت‌گیری با استفاده از روش یادگیری تقویتی عمیق

علی ریزه‌وندی، شهرام آزادی*

۱- دانشکده مهندسی مکانیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران.

*Azadi@kntu.ac.ir

چکیده

رانندگی خودکار یک فناوری جدید برای کاهش تصادفات رانندگی و بهبود رانندگی می‌باشد. در این پژوهش، یک سیاست تصمیم‌گیری مبتنی بر یادگیری تقویتی عمیق برای خودروهای خودران جهت سناریو سبقت‌گیری در بزرگراه ارائه شده است. برای این منظور ابتدا یک محیط ترافیکی بزرگراهی ایجاد می‌شود که هدف در آن عبور عامل از وسایل نقلیه اطراف با یک مانور کارآمد و ایمن می‌باشد. همچنین یک چارچوب کنترل سلسله مراتبی برای کنترل این وسایل نقلیه ارائه شده است که دستورات سطح بالا تصمیمات رانندگی را مدیریت می‌کند و دستورات سطح پایین به نظارت بر سرعت و شتاب وسیله نقلیه می‌پردازد. سپس، روش خاص مبتنی بر یادگیری تقویتی عمیق به نام الگوریتم گرادبان سیاست قطعی عمیق برای استخراج سیاست تصمیم‌گیری در بزرگراه استفاده می‌شود. سپس عملکرد الگوریتم گرادبان سیاست قطعی عمیق با الگوریتم شبکه عمیق کیو مورد مقایسه قرار گرفته است و نتایج استخراج شده از دو الگوریتم مورد ارزیابی و بررسی قرار خواهند گرفت. همچنین در این پژوهش برای شبیه‌سازی مسئله ذکر شده یعنی سبقت‌گیری در محیط بزرگراه از نرم افزار متلب نسخه ۲۰۲۲ استفاده شده است. نتایج شبیه‌سازی نشان می‌دهد که سیاست سبقت‌گیری مبتنی بر الگوریتم گرادبان سیاست قطعی عمیق می‌تواند وظایف رانندگی در بزرگراه را به طور اثربخش و ایمن انجام دهد.

کلمات کلیدی

خودروهای خودران، تصمیم‌گیری، یادگیری تقویتی عمیق، سبقت‌گیری، الگوریتم گرادبان سیاست قطعی عمیق.

رانندگی خودمختار وسیله نقلیه را قادر می‌سازد تا بدون دخالت انسان در سناریوهای مختلف رانندگی شرکت کند [۱]، [۲]. با عنایت به پتانسیل‌های فراوان هوش مصنوعی^۱، وسایل نقلیه خودران به یکی از کانون‌های تحقیقاتی در سراسر جهان تبدیل شده‌اند [۳]. بسیاری از خودروسازان مانند تویوتا، تسلا، فورد، آئودی، وایمو، مرسدس بنز، جنرال موتورز و غیره در حال توسعه خودروهای خودران خود هستند و به پیشرفت‌های قابل توجه‌ای در این زمینه دست یافته‌اند. در این میان، محققان حوزه خودرو مترصد ساخت خودروهای خودران در سطوح بالا می‌باشند [۴]. چهار ماژول مهم در خودروهای خودران وجود دارد که عبارتند از ادراک، تصمیم‌گیری، برنامه‌ریزی و کنترل [۵]. ادراک شامل درک خودروی خودران از محیط اطراف می‌باشد که توسط سنسورهای مانند لیدار، رادار، دوربین، سیستم موقعیت یاب جهانی^۲ و غیره این فرایند انجام می‌شود. ماژول تصمیم‌گیری رفتارهای رانندگی وسایل نقلیه را مدیریت می‌کند و این رفتارها شامل شتاب، ترمز، تغییر خط، حفظ خط و غیره است [۷]. ماژول برنامه‌ریزی به خودروهای خودران کمک می‌کند تا مسیرهای حرکتی معقول را از نقطه‌ای به نقطه دیگر طی کنند. در نهایت، ماژول کنترل به اجزای سیستم انتقال قدرت فرمان می‌دهد تا به طور دقیق عمل کنند تا مانورهای رانندگی را به پایان برسانند و مسیر برنامه‌ریزی را دنبال کنند. با توجه به درجات هوشمندی ماژول‌های ذکر شده، خودروهای خودران در شش سطح از سطح صفر تا سطح پنج طبقه‌بندی می‌شود [۸].

استراتژی تصمیم‌گیری در خودروهای خودران همچون مغز انسان در نظر گرفته می‌شود و بسیار حائز اهمیت می‌باشد [۹]. این سیاست اغلب توسط قوانین مبتنی بر تجربیات رانندگی انسان یا الگوبرداری از رویکردهای یادگیری نظارت شده، ایجاد می‌شود [۱۰]. به عنوان مثال، سونگ و همکاران از یک زنجیره مارکوف پیوسته برای پیش‌بینی حرکت وسایل نقلیه اطراف استفاده کردند. سپس، یک فرآیند تصمیم‌گیری مارکوف با قابلیت مشاهده نسبی^۳ برای ساخت چارچوب کلی تصمیم‌گیری استفاده شد [۱۲]. همچنین در [۱۳] قابلیت تصمیم‌گیری در موقعیت‌های ترافیک جاده‌های شهری توسعه داده شد. سیاست تصمیم‌گیری ارائه شده شامل معیارهای چندگانه ای است که به خودروهای شهری جهت انتخاب‌های عملی و معقول در شرایط ترافیکی مختلف کمک می‌کنند. در مرجع [۱۴]، نی و همکاران در مورد استراتژی تصمیم‌گیری تغییر خط برای اتومبیل‌های خودکار متصل مطالعاتی را انجام دادند. علاوه بر این، نویسندگان در [۱۵] ایده سیستم رانندگی شبیه انسان را ذکر کردند، این سیستم می‌تواند تصمیمات رانندگی را با در نظر گرفتن تقاضای رانندگی برای رانندگان انسانی تنظیم کند.

روش یادگیری تقویتی عمیق^۴ به عنوان یک ابزار قدرتمند برای مقابله با مشکلات تصمیم‌گیری متوالی طولانی در نظر گرفته می‌شود [۱۶]. در سال‌های اخیر، مطالعات زیادی در زمینه رانندگی خودکار مبتنی بر روش یادگیری تقویتی عمیق انجام شده است. به عنوان نمونه، دوان و همکاران یک ساختار سلسله مراتبی برای یادگیری سیاست تصمیم‌گیری از طریق روش یادگیری تقویتی^۵ ارائه دادند [۱۷]. در [۱۹] محققان از رویکردهای یادگیری تقویتی عمیق برای حل چالش‌های اجتناب از برخورد و دنباله‌روی مسیر برای خودروهای خودران استفاده کردند. نتایج گویای عملکرد بهتر رویکرد یادگیری تقویتی عمیق نسبت به روش یادگیری تقویتی برای دو چالش ذکر شده بود.

علاوه بر این، نویسندگان در [۲۱] نه تنها برنامه‌ریزی مسیر، بلکه مصرف سوخت برای وسایل نقلیه خودران را نیز در نظر گرفتند. الگوریتم مرتبط، یادگیری عمیق کیو^۶ است و ثابت شده است ماموریت‌های رانندگی را به طور مناسب انجام می‌دهد. هان و همکاران از این الگوریتم برای تصمیم‌گیری برای مانور تغییر خط یا حفظ خط برای خودروهای خودران متصل استفاده کردند، که در آن اطلاعات وسایل نقلیه نزدیک به عنوان دانش بازخورد از شبکه در نظر گرفته شد [۲۲]. سیاست به دست آمده جریان ترافیک و راحتی رانندگی را ارتقا بخشید. با این حال، روش‌های رایج یادگیری تقویتی عمیق به دلیل فضای عمل پیوسته و فضای حالت بزرگ، قادر به حل چالش سبقت‌گیری در بزرگراه نیستند [۲۳].

¹ Artificial Intelligence

² Global Positioning System

³ Partially Observable Markov Decision Process

⁴ Deep Reinforcement Learning

⁵ Reinforcement Learning

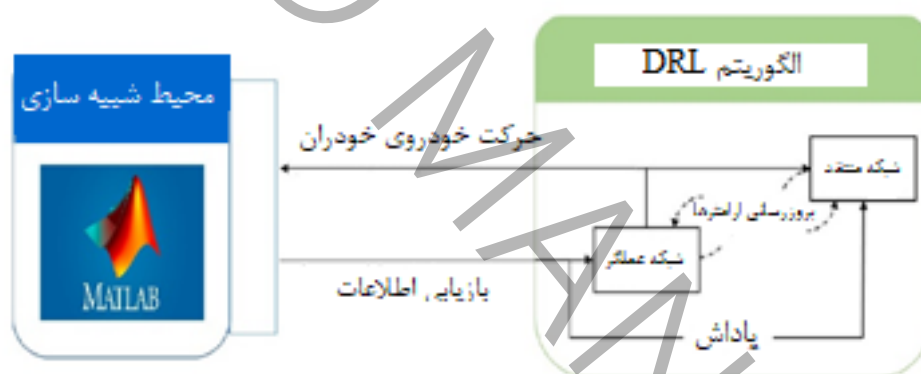
⁶ Deep Q Network

همچنین در [۲۴]، از روش یادگیری تقویتی عمیق با در نظر گرفتن تابع ریسک برای تصمیم‌گیری خودروی خودران در چند سناریو مختلف استفاده شد. در [۲۵] از روش رانندگی انتها به انتها برای پیش‌بینی تصمیم‌گیری مبتنی بر رفتار انسان در سیستم‌های حمل و نقل هوشمند استفاده شد. با توجه به عدم بازدهی روش‌های مذکور برای فضای عملکردی پیوسته، لذا در سال‌های اخیر استفاده از الگوریتم‌های یادگیری تقویتی عمیق که دارای فضای عملکردی پیوسته باشند برای فرایند تصمیم‌گیری در خودروهای خودران بسیار رواج پیدا کرده است. از اینرو در [۲۶] از الگوریتم عملگر-منتقد مبتنی بر یادگیری تقویتی برای یادگیری فرایند تصمیم‌گیری خودروی خودران در بزرگراه استفاده شد. همچنین در [۲۷] الگوریتم گرادیان سیاست قطعی عمیق جهت تصمیم‌گیری پیوسته خودروی خودران در محیط تقاطع شهری، مورد مطالعه قرار گرفت.

در این پژوهش، یک سیاست رانندگی مبتنی بر رویکرد یادگیری تقویتی عمیق جهت مانور سبقت‌گیری در محیط ترافیکی بزرگراه برای وسایل نقلیه خودران ارائه شده است. استراتژی تصمیم‌گیری پیشنهادی در سناریوی پیچیده دارای ایمنی و بازدهی مناسبی می‌باشد.

در این پژوهش، ابتدا سناریوی رانندگی مورد مطالعه در محیط بزرگراه طراحی می‌شود، که در آن هدف عبور عامل از یک سناریوی رانندگی خاص به طور موثر و ایمن می‌باشد. سپس، یک ساختار کنترل سلسله‌مراتبی برای کنترل حرکات جانبی و طولی عامل و وسایل نقلیه اطراف نشان داده می‌شود.

علاوه بر این، الگوریتم ویژه یادگیری تقویتی عمیق به نام الگوریتم گرادیان سیاست قطعی عمیق برای به‌دست آوردن استراتژی تصمیم‌گیری در محیط ترافیکی بزرگراه مورد استفاده قرار خواهد گرفت. در نهایت، عملکرد چارچوب کنترلی پیشنهادی شبیه‌سازی شده مورد بحث قرار خواهد گرفت. نحوه آموزش و یادگیری داده‌ها در پژوهش حاضر مطابق شکل ۱ می‌باشد.



شکل ۱: نحوه آموزش و یادگیری داده‌ها

Fig.1 Training process

مهمترین نقاط قوت و نوآوری‌های اصلی این پژوهش نسبت به [۲۷] عبارتند از: (۱) ارائه یک استراتژی بهینه سبقت‌گیری در محیط بزرگراهی مبتنی بر رویکرد یادگیری تقویتی عمیق برای خودروهای خودران (۲) استفاده از الگوریتم گرادیان سیاست قطعی عمیق برای فاز تصمیم‌گیری در خودروی خودران در محیط ترافیکی بزرگراهی (۳) بهبود هوشمندی رفتارهای خودروی خودران در محیط پویای ترافیکی.

در ادامه، ابتدا محیط رانندگی بزرگراه و ماژول‌های کنترل عامل و وسایل نقلیه اطراف در بخش ۲ شرح داده می‌شوند. سپس در بخش ۳ الگوریتم گرادیان سیاست قطعی عمیق مورد بحث و بررسی قرار می‌گیرد، به‌طوری‌که در آن پارامترهای چارچوب یادگیری تقویتی به تفصیل مورد بحث قرار می‌گیرد. همچنین در بخش ۴ نتایج حاصل شده مورد ارزیابی قرار می‌گیرند. در نهایت در بخش ۵ نتیجه‌گیری انجام می‌شود.

۲- محیط رانندگی

در این بخش سناریوی رانندگی مورد مطالعه در بزرگراه یعنی سناریو مهم و پرکاربرد سبقت‌گیری معرفی می‌شود. در این کار برای طراحی سناریو از محیط نرم افزار متلب نسخه ۲۰۲۲ استفاده شده است، به‌اینصورت که ابتدا یک محیط بزرگراه سه بانده ساخته می‌شود. سپس عامل و محیط ترافیکی اطراف آن طراحی می‌شود. علاوه بر این، یک کنترل کننده حرکت سلسله مراتبی برای مدیریت حرکات جانبی و طولی عامل و وسایل نقلیه اطراف معرفی می‌شود.

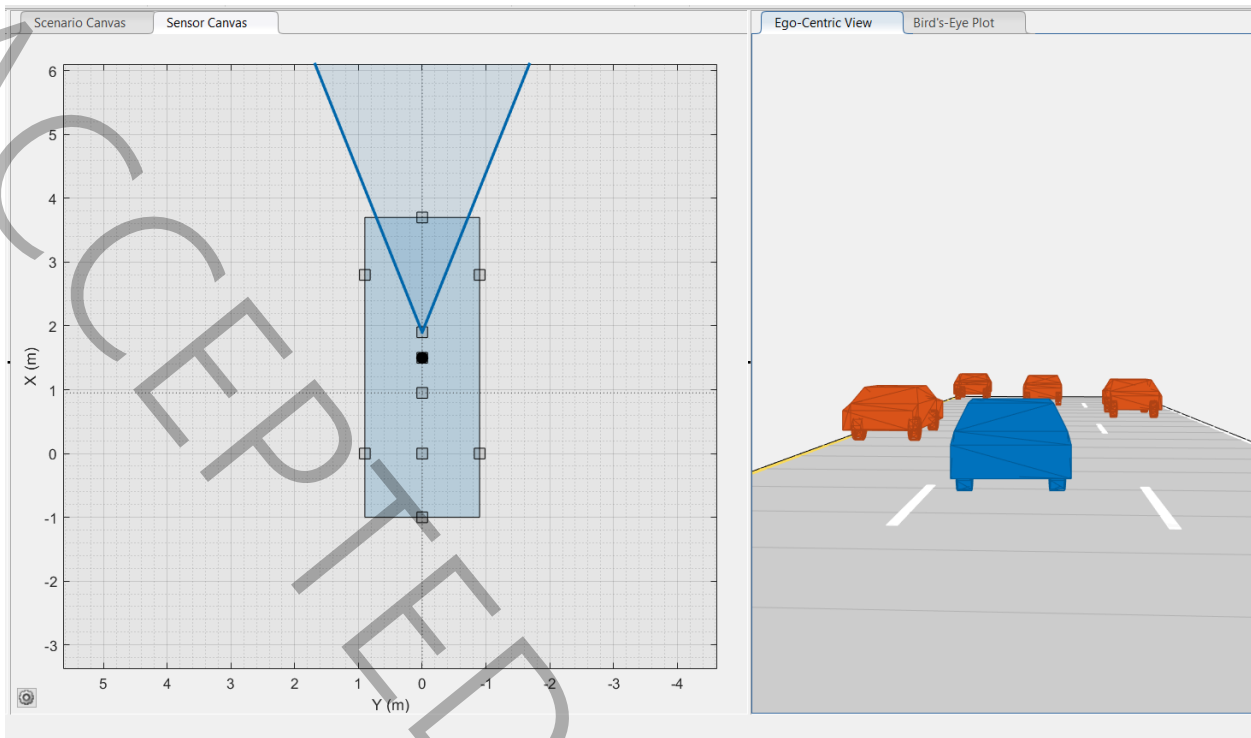
تصمیم‌گیری در رانندگی خودکار به معنای انتخاب رشته‌ای از رفتارهای رانندگی منطقی برای دستیابی به مقاصد رانندگی خاص است. در مانور سبقت‌گیری در بزرگراه، این رفتارها شامل تغییر خط، حفظ خط، شتاب‌گیری، ترمزگیری است. اهداف اصلی اجتناب از برخورد، حرکت موثر و سریع و رانندگی در خط سرعت می‌باشد. به عبارت دیگر، شتاب گرفتن و پیشی گرفتن از سایر وسایل نقلیه یک رفتار معمول در رانندگی است که سبقت‌گیری نامیده می‌شود.

این کار مساله تصمیم‌گیری در بزرگراه برای وسایل نقلیه خودران را مورد بحث قرار می‌دهد و سناریوی رانندگی در شکل ۲ نشان داده شده است. وسیله نقلیه نارنجی رنگ عامل است و سایر خودروهای سبز به عنوان وسایل نقلیه اطراف نامیده می‌شوند. عامل در خط میانی با سرعت تصادفی شروع به رانندگی می‌کند. هدف عامل، رانندگی با بیشترین سرعت ممکن و بدون تصادف با خودروهای اطراف می‌باشد. از این رو، این اهداف به ترتیب به کارایی و ایمنی الگوریتم تصمیم‌گیری تعبیر می‌شوند.

سرعت و موقعیت اولیه خودروهای اطراف به صورت تصادفی طراحی شده است. این شرایط گویای عدم قطعیت‌ها در محیط ترافیکی واقعی و نزدیک بودن محیط ترافیکی بزرگراه به شرایط واقعی می‌باشد.

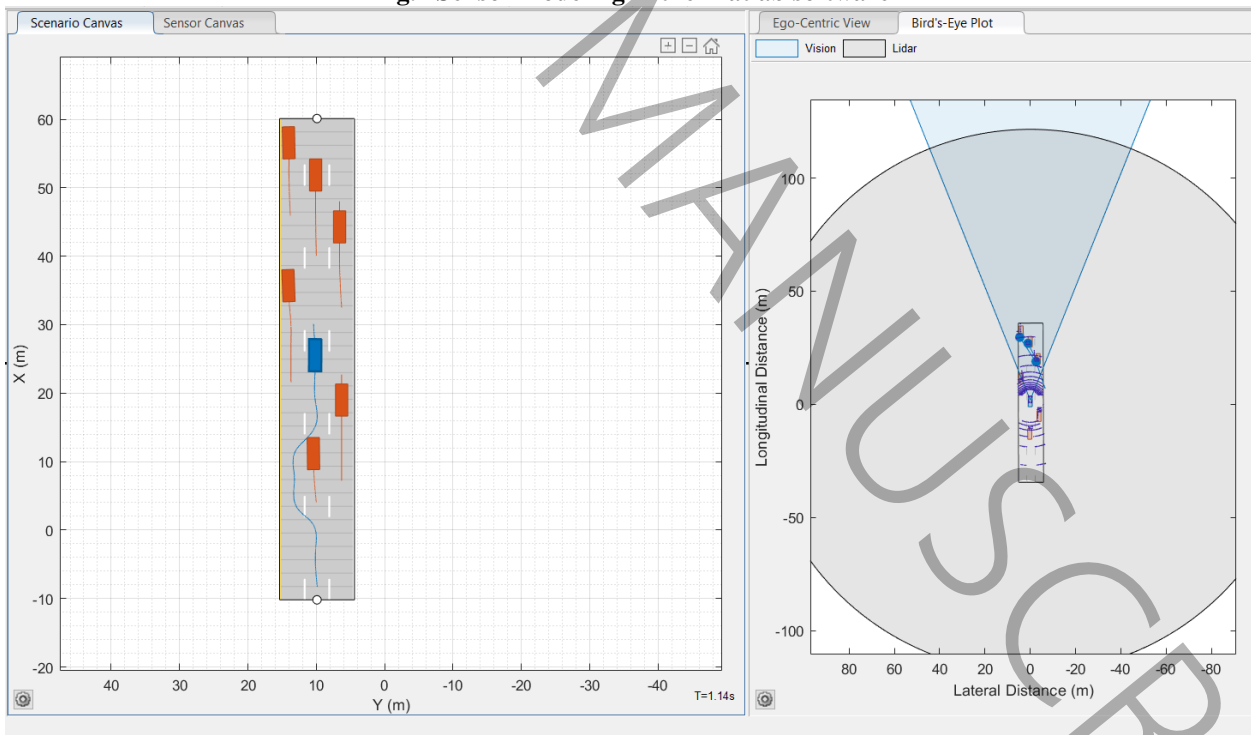
در ابتدای این وظیفه رانندگی، تمام خودروهای اطراف در جلوی عامل قرار می‌گیرند. همچنین در هر خط، دو خودرو وجود دارد. دو شرط باعث توقف عامل می‌شود که اولی تصادف با وسایل نقلیه دیگر و دومی رسیدن به محدودیت زمانی می‌باشد. پروسه رانندگی از نقطه شروع تا پایان در این نوشتار به عنوان یک اپیزود نامیده می‌شود.

در این پژوهش جهت طراحی سناریوی رانندگی از نرم افزار متلب ۲۰۲۲ استفاده شده است، برای طراحی سناریو ابتدا محیط بزرگراهی به صورت سه بانده طراحی می‌شود، سپس محیط ترافیکی که شامل عامل و خودروهای اطراف عامل می‌باشند طراحی شده اند، سپس مدلسازی سنسورها مطابق شکل ۲ بر روی عامل انجام می‌شود که سنسورها شامل دوربین و لیدار می‌باشند. در نهایت مطابق شکل ۳ سناریوی سبقت‌گیری بر محیط ترافیکی طراحی شده، اعمال می‌شود.



شکل ۲: مدل‌سازی سنسورها بر روی عامل در نرم افزار متلب

Fig.2 Sensor modeling in the Matlab software



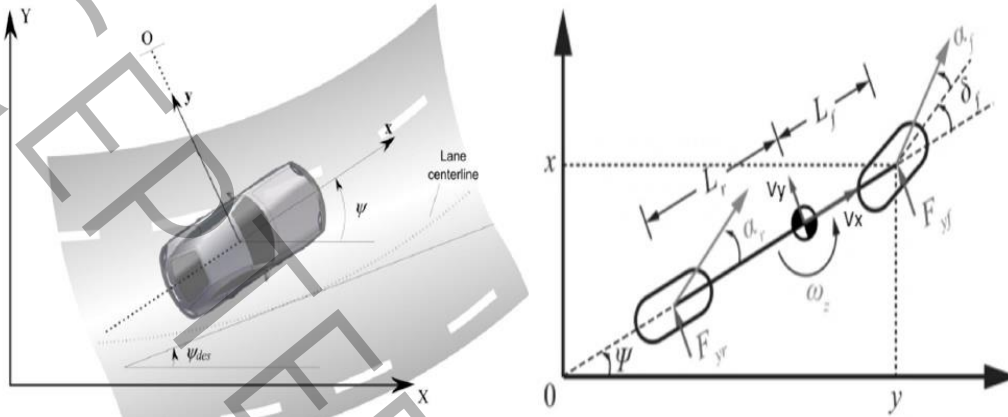
شکل ۳: طراحی محیط ترافیکی بزرگراه در نرم افزار متلب

Fig.3 Highway traffic environment design in the MATLAB software

فرض بر این است که موقعیت، سرعت و شتاب خودروهای اطراف برای عامل مشخص است. این محدودیت‌ها عامل را به یادگیری رانندگی در سناریو طراحی شده از طریق روش آموزش و خطا سوق می‌دهد. در بخش بعدی، رویکرد یادگیری تقویتی عمیق برای محقق ساختن فرآیند یادگیری و استخراج سیاست تصمیم‌گیری در بزرگراه معرفی می‌شود.

۳- مدل دینامیکی خودرو

در این پژوهش از مدل دو درجه آزادی یا مدل دوچرخه مطابق شکل ۴ برای بیان دینامیک خودرو استفاده شده است.



شکل ۴: مدل دو درجه آزادی خودرو

Fig.4 Two degree-of-freedom vehicle model

معادلات فضای حالت دینامیک خودرو مطابق زیر می‌باشند:

$$\begin{bmatrix} \dot{y} \\ \ddot{y} \\ \dot{\psi} \\ \ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \frac{2C_{\alpha_f} + 2C_{\alpha_r}}{mV_x} & 0 & V_x + \frac{2L_f C_{\alpha_f} - 2L_r C_{\alpha_r}}{mV_x} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{2L_f C_{\alpha_f} - 2L_r C_{\alpha_r}}{I_z V_x} & 0 & \frac{2L_f^2 C_{\alpha_f} + 2L_r^2 C_{\alpha_r}}{I_z V_x} \end{bmatrix} \begin{bmatrix} y \\ \dot{y} \\ \psi \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2C_{\alpha_f}}{m} \\ 0 \\ \frac{2L_f C_{\alpha_f}}{I_z} \end{bmatrix} \quad (1)$$

که پارامترهای موجود در معادله (۱) و همچنین مقادیر آنها در جدول ۱ قابل مشاهده می‌باشند.

جدول ۱: پارامترهای خودرو دو درجه آزادی

Table 1. Vehicle parameters

مقدار	نماد	پارامتر
۱۶۵۰ kg	m	جرم کلی خودرو
۲۸۷۵ kg.m ²	I _z	ممان اینرسی زاویه یابو
۱/۴ m	l _f	فاصله مرکز جرم تا محور جلو
۱/۶ m	l _r	فاصله مرکز جرم تا محور عقب
۲۹۰۰۰ N/rad	C _{αf}	سختی لاستیک جلو
۳۳۰۰۰ N/rad	C _{αr}	سختی لاستیک عقب

۴- روش یادگیری تقویتی عمیق

در این بخش ابتدا روش یادگیری تقویتی معرفی می‌شود و سپس الگوریتم‌های خاص یادگیری تقویتی عمیق مورد بررسی قرار می‌گیرند. ابتدا تعامل در روش یادگیری تقویتی بین عامل و محیط توضیح داده می‌شود، سپس الگوریتم‌های مبتنی بر ارزش و مبتنی بر سیاست مورد ارزیابی قرار می‌گیرند. سپس، الگوریتم یادگیری عمیق کیو که شبکه عصبی و الگوریتم یادگیری کیو را در بر می‌گیرد، معرفی می‌شود. سپس الگوریتم سارسا^۱ معرفی می‌شود. پس از آن الگوریتم‌های مبتنی بر سیاست مانند بهبود سیاست بازه اعتماد^۲ و بهبود سیاست نزدیک^۳ مورد ارزیابی قرار می‌گیرند. در نهایت الگوریتم گرادیان سیاست قطعی عمیق که در پژوهش حاضر مورد استفاده قرار گرفته است، مورد تبیین قرار می‌گیرد.

۴-۱ روش یادگیری تقویتی

رویکرد یادگیری تقویتی فرآیندی را توصیف می‌کند که یک عامل هوشمند با محیط خود تعامل دارد. روش یادگیری تقویتی برای حل مسائل تصمیم‌گیری متوالی بسیار قدرتمند و مفید می‌باشد. هدف عامل جستجوی یک توالی بهینه از اقدامات کنترلی بر اساس بازخورد خود از محیط می‌باشد. با توجه به ویژگی‌های ذاتی خود ارزیابی و خود ارتقای روش یادگیری تقویتی، این روش به طور گسترده در بسیاری از کارهای تحقیقاتی از جمله [۳۵] استفاده شده است.

در سناریو تصمیم‌گیری در بزرگراه، عامل و محیط به ترتیب خودرو قبلی (نارنجی رنگ) و خودروه‌های اطراف (سبز رنگ) هستند. این مساله را می‌توان توسط فرآیندهای تصمیم مارکوف^۲ مدل‌سازی نمود، در فرایند مارکوف متغیر حالت بعدی فقط به وضعیت و عملکرد فعلی مربوط می‌شود [۳۶]. این بدان معناست که مسئله تصمیم‌گیری متوالی مورد بحث در رانندگی خودران دارای ویژگی مارکوف است.

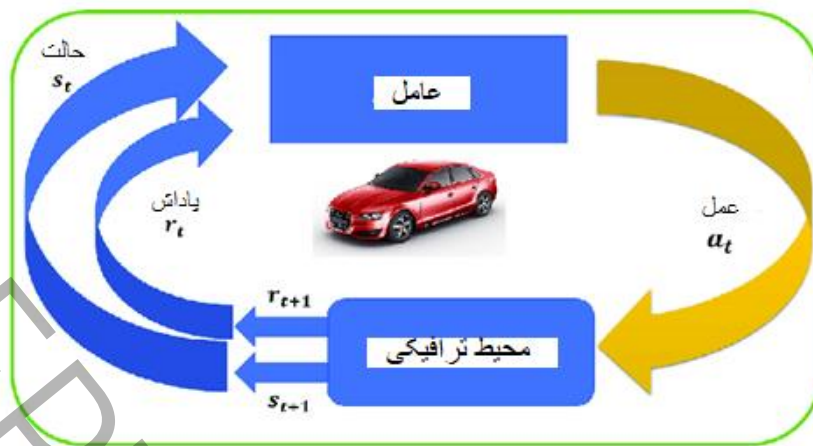
مطابق شکل ۵ فرآیند تصمیم مارکوف مربوطه اغلب تعامل یادگیری تقویتی را به صورت یک تاپل (S, A, P, R, γ) نشان می‌دهد. که در آن S و A به ترتیب مجموعه حالت‌ها و کنش‌های کنترلی هستند. همچنین P و R عناصر مهم محیط در روش یادگیری تقویتی هستند و به ترتیب به معنای مدل گذار و پاداش هستند.

¹ State Action Reward State Action

² Trust Region Policy Optimization

³ Proximal Policy Optimization

² Markov Decision Process



شکل ۵: مدل یادگیری تقویتی (RL)

Fig.5 Reinforcement learning (RL) method

در روش یادگیری تقویتی، عمل فعلی به طور همزمان بر پاداش‌های حال و آینده تأثیر می‌گذارد. از این رو، γ یک ضریب ثابت برای متعادل کردن دو پاداش ذکر شده می‌باشد.

برای نمایش مجموع پاداش‌های آینده، پاداش انباشته R_t به صورت زیر تعریف می‌شود:

$$R_t = \sum_t \gamma^t \cdot r_t \quad (۲)$$

که در آن t زمان و r_t پاداش مربوطه می‌باشد.

برای ثبت ارزش حالت s هنگام انجام عمل a و ارزیابی پاداش به‌دست آمده، دو تابع ارزش و تابع کیو Q به‌صورت زیر تعریف می‌شوند:

$$V^\pi(s_t) = E_\pi [R_t | S_t, \pi] \quad (۳)$$

$$Q^\pi(s_t, a_t) = E_\pi [R_t | S_t, a_t, \pi] \quad (۴)$$

که π به سیاست عمل کنترلی، V تابع ارزش، و Q تابع حالت-عمل (که به اختصار جدول Q نامیده می‌شود) می‌باشند.

برای به‌روزرسانی آسان، تابع حالت-عمل معمولاً به صورت بازگشتی مطابق زیر بازنویسی می‌شود:

$$Q^\pi(s_t, a_t) = E_\pi [R_t + \gamma \max_{a_{t+1}} Q^\pi(s_{t+1}, a_{t+1})] \quad (۵)$$

در نهایت، عمل کنترلی بهینه با توجه به سیاست کنترلی اتخاذ شده توسط تابع حالت-عمل به‌صورت زیر تعیین می‌شود:

$$\pi(s_t) = \arg \max Q(s_t, a_t) \quad (۶)$$

بنابراین، ماهیت الگوریتم‌های مختلف روش یادگیری تقویتی به‌روزرسانی تابع حالت-عمل $Q(s_t, a_t)$ به روش‌های مختلف است. با توجه به سبک به‌روزرسانی قوانین، الگوریتم‌های یادگیری تقویتی طبقه‌بندی‌های متنوعی دارند، از جمله می‌توان الگوریتم‌های یادگیری تقویتی را به دو دسته مبتنی بر سیاست و مبتنی بر ارزش دسته‌بندی نمود [۳۷].

۴-۲ روش مبتنی بر ارزش^۱:

برای یافتن سیاست بهینه، تابع حالت-عمل را می‌توان به صورت معادله (۷) بیان نمود:

$$Q^{\pi^*}(s_t, a_t) = \max_{\pi} (Q^{\pi}(s_{t+1}, a_{t+1})) \quad (7)$$

به طوری که سیاست بهینه را به صورت معادله (۸) می‌توان بیان نمود:

$$\pi^*(s) = \arg \max_a (Q^*(s, a)) \quad (8)$$

یکی از روش‌های حل مسائل یادگیری تقویتی با تلاش برای یافتن تابع مقدار-عمل بهینه، یادگیری کیو نام دارد. روش یادگیری کیو یک رویکرد خارج از سیاست و بدون مدل است. مقادیر حالت و عمل نیز به عنوان مقادیر کیو شناخته می‌شوند.

در حقیقت، مقادیر کیو بر اساس داده‌های جمع‌آوری شده توسط عامل در هنگام تعامل با محیط طبق معادله بهینه‌سازی بلمن معادله (۴)، به روزرسانی می‌شوند.

در روش‌های مبتنی بر ارزش، به منظور یادگیری تابع کیو، از سیاست ϵ حریصانه در انتخاب اقدام جدید استفاده می‌شود، به طوری که عامل یک عمل جدید (بر اساس کیو) با احتمال $1-\epsilon$ انتخاب می‌کند. (مقدار ϵ بین ۰ و ۱ است) و یک اقدام تصادفی با احتمال ϵ انتخاب می‌شود.

در واقع، در روش‌های مبتنی بر ارزش عامل بدون توجه به سیاست ورودی به آن فقط در جهت بهینه‌سازی تابع ارزش یا تابع کیو گام برداشته و آموزش دیده می‌شود.

۴-۲-۱ الگوریتم شبکه عمیق کیو

همانطور که گفته شد، در روش یادگیری تقویتی عمیق از شبکه‌ی عصبی برای تقریب تابع کیو استفاده می‌شود، در حقیقت در استراتژی تصمیم‌گیری مبتنی بر یادگیری تقویتی عمیق، یک عامل می‌تواند در یک محیط تصادفی با انتخاب یک سری اقدامات در طول یک توالی از مراحل زمانی عمل کند، سپس از بازخوردها (یعنی پاداش) یاد بگیرد تا پاداش تجمعی را به حداکثر برساند. شبکه عمیق کیو یکی از مهم‌ترین الگوریتم‌های مورد استفاده در مسائل مربوط به روش یادگیری تقویتی عمیق می‌باشد.

شبکه عمیق کیو اولین بار برای اجرای بازی‌های آتاری ارائه شده است. این شبکه نقاط قوت یادگیری عمیق (شبکه عصبی) و یادگیری کیو را برای به دست آوردن تابع جدید ارزش-حالت ترکیب می‌کند. در یادگیری کیو، قانون به روز رسانی به شرح زیر می‌باشد:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad (9)$$

که در آن $\alpha \in [0, 1]$ به عنوان نرخ یادگیری برای مبادله تجربیات آموخته شده قدیمی و جدید از محیط نام‌گذاری شده است. همچنین s' و a' حالت و عمل در مرحله زمانی بعدی هستند. یادگیری کیو معمولی قادر به حل مسأله با فضای بزرگی از متغیر حالت نیست، زیرا به زمان زیادی برای به دست آوردن جدول Q که متغیر است، نیاز دارد. بنابراین، در روش یادگیری عمیق کیو، یک شبکه عصبی برای تقریب جدول کیو به صورت $Q(s, a; \theta)$ استفاده می‌شود. برای شبکه عصبی، ورودی‌ها آرایه‌های متغیرهای حالت و اقدامات کنترلی هستند و خروجی تابع حالت-مقدار است. برای اندازه‌گیری اختلاف بین جدول کیو تقریبی و واقعی، تابع ضرر مانند عبارت زیر معرفی می‌شود:

¹ Value-based

$$L(\theta) = E\left[\sum_{t=1}^N (y_t - Q(s, a; \theta))^2\right] \quad (10)$$

که

$$y_t = r_t + \gamma \max_a Q(s', a'; \theta') \quad (11)$$

همانطور که مشاهده می‌شود، دو پارامتر (θ, θ') در شبکه عصبی وجود دارد که دو شبکه را در شبکه عمیق کیو تفویض می‌کنند. این شبکه‌ها، شبکه‌های پیش‌بینی و هدف می‌باشند. اولی برای تخمین عمل کنترل فعلی اعمال می‌شود و دومی برای ایجاد مقدار هدف است. به طور کلی، شبکه هدف پارامترها را در هر تعداد معینی از مراحل زمانی از شبکه پیش‌بینی تقلید می‌کند. با انجام این کار، جدول کیو هدف همگرا می‌شود تا یک نمونه را تا حدی پیش‌بینی کند به طوری که ناپایداری شبکه برطرف شود. در شبکه عمیق کیو، شبکه عصبی برخط با کاهش گرادبان به شکل زیر به روز می‌شود:

$$\nabla_{\theta} L(\theta) = E[(y_i - Q(s, a; \theta)) \nabla_{\theta} Q(s, a; \theta)] \quad (12)$$

این عملیات شبکه عمیق کیو را به عنوان یک الگوریتم مبتنی بر ارزش تبدیل می‌کند و همچنین وضعیت‌ها و پاداش‌ها با معیار خاصی به دست می‌آیند.

۲-۲-۴ الگوریتم سارسا

علاوه بر روش یادگیری عمیق کیو، نمونه دیگری از روش مبتنی بر ارزش، روش سارسا است. در سارسا، مقادیر کیو به صورت معادله (۱۳) به روزرسانی می‌شوند:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r + \gamma Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t) \quad (13)$$

در حالی که یادگیری کیو از تجربیات ایجاد شده می‌آموزد که ممکن است از سیاست فعلی نیز نباشند (خارج از سیاست)، اما روش سارسا از تجارب جمع‌آوری شده که مبتنی بر سیاست اولیه وارده بر عامل می‌باشد، می‌آموزد.

تمایز دیگر این است که یادگیری کیو، مقادیر کیو را با انتخاب حداکثر یک تقریب در مرحله بعد به روزرسانی می‌کند، اما روش سارسا با پیروی از سیاست ϵ حریصانه^۱ فرایند به روزرسانی را انجام می‌دهد.

همچنین هر دو روش اقدام را با استفاده از سیاست ϵ حریصانه برای وظیفه‌ی اکتشاف انتخاب می‌کنند. با این حال، روش یادگیری کیو، مقادیر کیو را با یک سیاست حریصانه به روزرسانی می‌کند، در حالی که روش سارسا مقادیر کیو را با پیروی از سیاست ϵ حریصانه به روزرسانی می‌کند.

۳-۴ روش مبتنی بر سیاست^۲

در حالیکه روش‌های مبتنی بر ارزش، تابع ارزش (کیو) را بهینه کرده و سپس خروجی را به سیاست بهینه منتج می‌کنند، روش‌های مبتنی بر سیاست مطابق معادله (۱۴) عموماً یاد می‌گیرند که یک سیاست را مستقیماً بر اساس مجموع پاداش‌های دریافتی بهینه کنند.

$$\max J(\theta) = \max V_{\pi_{\theta}} = \max E_{\pi_{\theta}} [G_t | S_t = S] \quad (14)$$

¹ greedy

² Policy-based

³ Kullback-Leibler

⁴ critic

⁵ actor

که در آن θ پارامترهای سیاست است که با گرادیان سیاست مطابق معادله (۱۵) بهروزرسانی می‌شود.

$$\nabla J(\theta) = E_{\pi} [G_t \nabla \ln \pi_{\theta}(a_t | s_t)] \quad (15)$$

بهروز رسانی پارامترها احتمال اقدامات انجام شده توسط عامل را به صورت لگاریتمی افزایش می‌دهد.

روش‌های گرادیان سیاست، سیاست را از تندترین شیب بهروزرسانی می‌کنند تا بازدهی را به حداکثر مقدار برسانند.

از جمله الگوریتم‌های مورد استفاده در روش‌های مبتنی بر سیاست می‌توان به الگوریتم‌های بهبود سیاست بازه اعتماد و بهبود سیاست نزدیک اشاره نمود.

الگوریتم بهبود سیاست بازه اعتماد با در نظر گرفتن قید بهروز رسانی سیاست بر اساس روش واگرایی^۳ پارامترهای شبکه را بهروزرسانی می‌کند به طوری که سیاست جدید تفاوت قابل توجه‌ای با سیاست فعلی نداشته باشد.

برخلاف الگوریتم بهبود سیاست بازه اعتماد که قید را برای مسئله بهینه‌سازی در نظر می‌گیرد. الگوریتم بهبود سیاست نزدیک شامل قید در تابع هدف بهینه‌سازی در قالب یک تابع هدف جایگزین جدا می‌باشد. در حقیقت، در روش بهبود سیاست نزدیک اگر سیاست پیشنهادی جدید با سیاست قبلی تفاوت زیادی داشته باشد، تابع هدف جریمه شده و پاداش منفی دریافت خواهد کرد.

۴-۴ الگوریتم گرادیان سیاست قطعی عمیق

اگرچه عملکرد الگوریتم یادگیری عمیق کیو در مسائلی که ابعاد زیادی دارند بسیار پربازده می‌باشد، اما فضای اقدام در این شبکه گسسته است و از آنجایی که بسیاری از مسائل واقعی و حائز اهمیت برای ما از جمله مسائل مربوط به کنترل فیزیکی، دارای فضای اقدام پیوسته هستند، گسسته بودن فضا در شبکه یادگیری عمیق کیو یک نقطه ضعف به‌شمار می‌آید.

روش گرادیان سیاست قطعی عمیق یک الگوریتم خارج از سیاست است و در فضای عملکردی پیوسته قابل استفاده است. الگوریتم گرادیان سیاست قطعی عمیق از دو بخش اصلی تشکیل شده است: یادگیری تابع کیو که توسط شبکه منتقد^۴ انجام می‌شود و یادگیری سیاست انجام شده توسط شبکه سیاست عملگر^۵.

معادله بلمن برای الگوریتم گرادیان سیاست قطعی عمیق مطابق معادله (۱۶) می‌باشد:

$$L(\phi, \theta) = E_D \left[(Q_{\phi}(s, a) - (r + \gamma(1-d) \max_{a'} Q_{\phi}(s', a')))^2 \right] \quad (16)$$

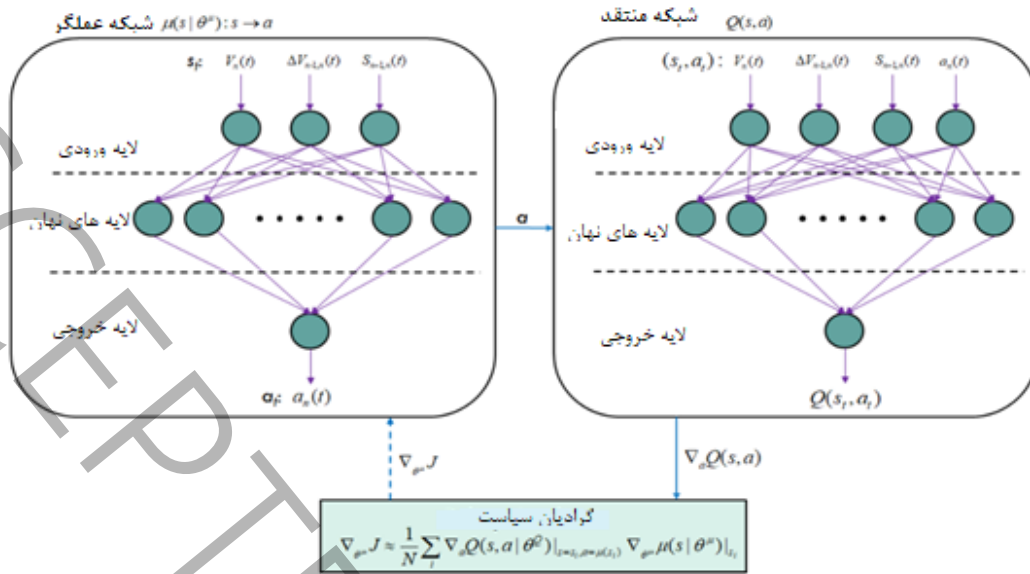
که $Q_{\phi}(s, a)$ شبکه منتقد با پارامتر ϕ می‌باشد.

مطابق معادله (۱۶)، در الگوریتم گرادیان سیاست قطعی عمیق، شبکه سیاست تلاش می‌کند تا یک سیاست قطعی $u_{\phi}(s)$ را بیاموزد به طوریکه عمل انتخاب شده توسط عامل در جهت به حداکثر رساندن تابع $Q^*(s, a)$ باشد.

برای کاهش مشکل پایداری شبکه در الگوریتم گرادیان سیاست قطعی عمیق، از بافر پخش مجدد و شبکه‌های هدف استفاده می‌شود. شبکه‌های هدف مورد استفاده در الگوریتم گرادیان سیاست قطعی عمیق یک شبکه منتقد هدف و یک شبکه هدف سیاست می‌باشند.

معماری هر دو شبکه هدف از نسخه اصلی خود کپی شده‌اند، اما پارامترهای ϕ از نسخه اصلی خود عقب‌تر هستند. در واقع، شبکه‌های هدف با میانگین بهروز می‌شوند، که سبب می‌شود تا شبکه‌های اصلی را به آرامی ردیابی کنند و در نتیجه پایداری را افزایش دهند.

نحوه عملکرد الگوریتم گرادیان سیاست قطعی عمیق در شکل ۶ قابل مشاهده می‌باشد.



شکل ۶: نحوه عملکرد الگوریتم گرادینان سیاست قطعی عمیق

Fig.6 DDPG algorithm

۴- مشخصات پارامترها

برای استخراج استراتژی تصمیم‌گیری مبتنی بر الگوریتم گرادینان سیاست قطعی عمیق، متغیرها جهت شبیه‌سازی محیط رانندگی به شرح زیر مقداردهی اولیه می‌شوند.

عمل‌های کنترلی دریچه گاز و زاویه فرمان خودرو هستند. همچنین متغیرهای حالت مطابق معادله (۱۷) و (۱۸) فاصله و سرعت نسبی بین عامل و خودروهای اطراف آن هستند:

$$\Delta s = |s_{ag} - s_{su}| \quad (17)$$

$$\Delta V = |V_{ag} - V_{su}| \quad (18)$$

که در آن s و v اطلاعات موقعیت و سرعت به دست آمده از دینامیک خودرو هستند. همچنین اندیس‌های ag و su به ترتیب عامل و خودروهای اطراف را نشان می‌دهند. لازم به ذکر است که معادله (۱۷) و (۱۸) را می‌توان به عنوان مدل انتقال P در چارچوب روش یادگیری تقویتی نیز در نظر گرفت.

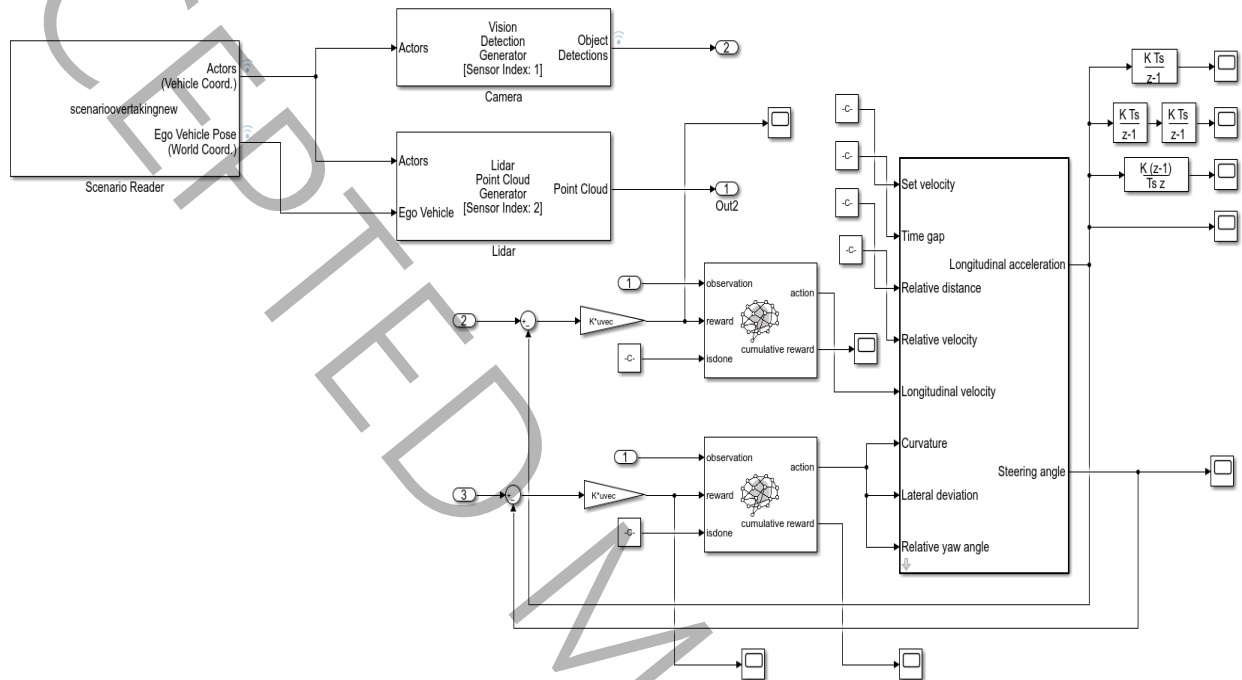
در نهایت، تابع پاداش R در این پژوهش شامل سه مورد است که نشان‌دهنده کارایی، ایمنی و اهداف رانندگی است. در حقیقت، عامل باید با حداکثر سرعت ممکن رانندگی کند، به‌طوری‌که در خط راست بماند و از تصادف با سایر خودروهای اطراف اجتناب کند. پاداش طراحی شده در هر مرحله زمانی (t) به صورت معادله (۱۹) تعریف می‌شود:

$$R_t = -100(\text{collision}) - 40(L-1)^2 - 10(V_{ag} - V_{ag}^{\max})^2 \quad (19)$$

که در آن برخورد به صورت $\{0, 1\}$ تعریف می‌شود که شرایط برخورد را برای عامل نشان می‌دهد. همچنین تعداد لاین‌ها به صورت $\{1, 2, 3\}$ است که نشان‌دهنده شماره خط در بزرگراه می‌باشد.

در این پژوهش سیاست کنترلی تصمیم‌گیری پیشنهادی در محیط نرم افزار متلب شبیه‌سازی، آموزش و ارزیابی شده است. تعداد خطوط و وسایل نقلیه اطراف به ترتیب ۳ و ۶ عدد است. همچنین مقادیر ضریب تخفیف و نرخ یادگیری به ترتیب ۰/۸ و ۰/۲ می‌باشد. تعداد لایه های شبکه ارزش ۱۲۸ هستند. همچنین تعداد اپیزودها ۱۰۰ عدد می‌باشد.

در شکل ۷ نمودار بلوکی سناریو سبقت‌گیری شبیه‌سازی شده در نرم افزار متلب- سیمولینک قابل مشاهده می‌باشد.



شکل ۷: نمودار بلوکی سناریو سبقت‌گیری در نرم افزار سیمولینک

Fig.7 Block diagram in the Simulink

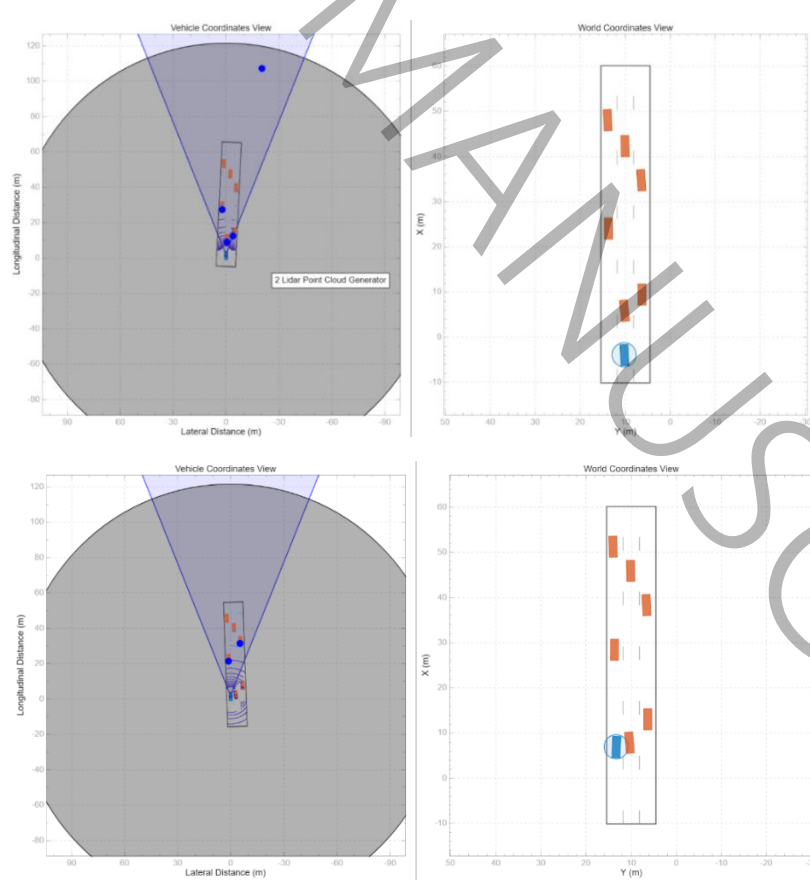
مطابق شکل ۷ دیاگرام بلوکی شامل سه زیرسیستم سناریو، یادگیری تقویتی و دینامیک خودرو به همراه کنترل آن می‌باشد. در بخش بعدی عملکرد الگوریتم تصمیم‌گیری ارائه شده در زیرسیستم یادگیری تقویتی مورد بحث قرار می‌گیرد.

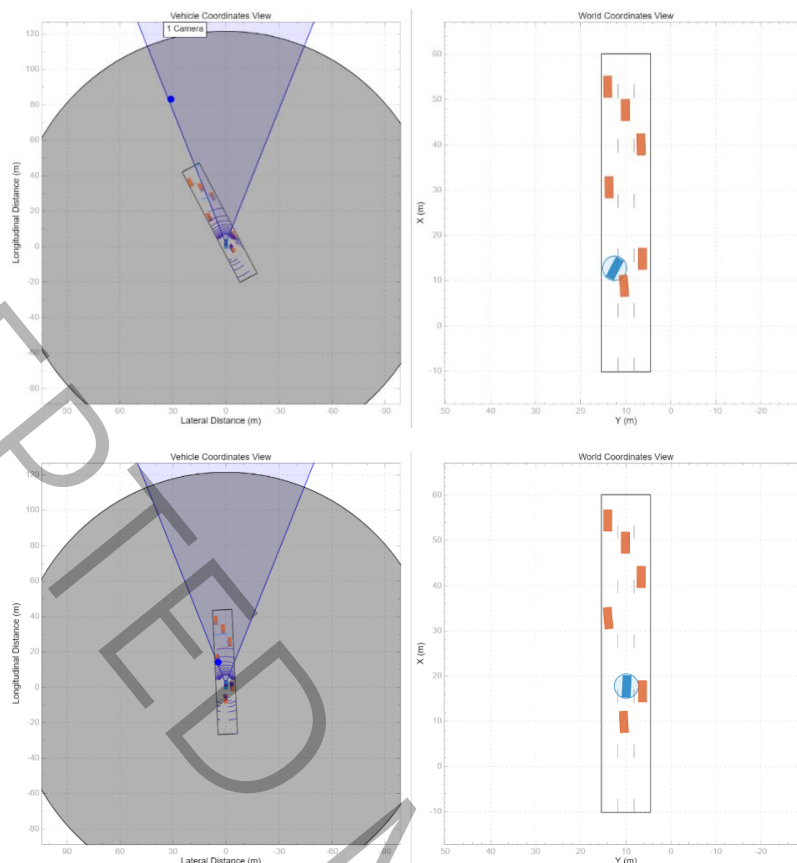
۵- نتایج و بحث

در این بخش عملکرد کنترلی الگوریتم گرادیان سیاست قطعی عمیق جهت فرایند تصمیم‌گیری عامل در محیط ترافیکی بزرگراه مورد ارزیابی و تحلیل قرار می‌گیرد.

در ارزیابی، ابتدا اثربخشی سیاست تصمیم‌گیری با روش دیگر مقایسه و راستی‌آزمایی می‌شود. نتایج شبیه‌سازی حاکی از بهینه بودن آن است. سپس، توانایی یادگیری الگوریتم گرادیان سیاست قطعی عمیق پیشنهادی با تجزیه و تحلیل پاداش‌های انباشته شده اثبات می‌شود. در شکل ۸ مانور سبقت‌گیری انجام شده توسط عامل (خودرو آبی رنگ) در محیط ترافیکی طراحی شده در بزرگراه قابل مشاهده می‌باشد.

ACCEPTED





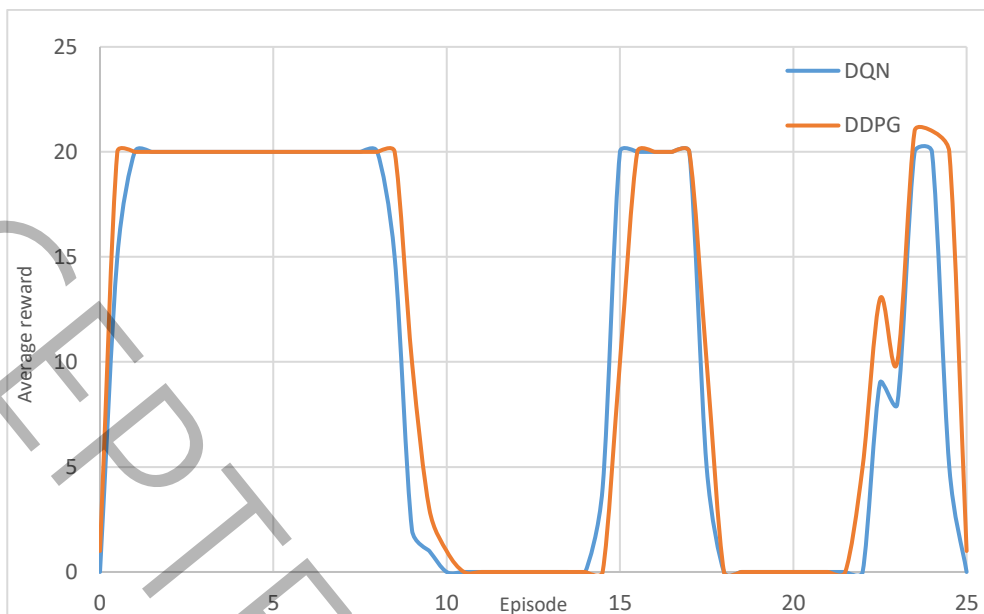
شکل ۸: مانور سبقت‌گیری انجام شده توسط عامل در محیط بزرگراه

Fig.8 Overtaking monouvre

۵-۱ اثربخشی الگوریتم گرادیان سیاست قطعی عمیق

در این بخش دو روش برای تصمیم‌گیری عامل در محیط ترافیکی بزرگراه ارائه شده است. ابتدا الگوریتم یادگیری عمیق کیو و سپس الگوریتم گرادیان سیاست قطعی عمیق مورد بررسی و ارزیابی قرار می‌گیرد. همچنین مزیت و برتری الگوریتم گرادیان سیاست قطعی عمیق نسبت به الگوریتم یادگیری عمیق کیو در این بخش نمایش داده می‌شود. در حقیقت الگوریتم یادگیری عمیق کیو به عنوان معیاری برای استنتاج بهینه بودن الگوریتم گرادیان سیاست قطعی عمیق در نظر گرفته می‌شود. لازم به ذکر است که پارامترهای دو الگوریتم یادگیری عمیق کیو و الگوریتم گرادیان سیاست قطعی عمیق یکسان در نظر گرفته شده‌اند.

مجموع پاداش به دست آمده در هر اپیزود نشان‌دهنده عملکرد سیاست کنترلی در روش یادگیری تقویتی عمیق می‌باشد. میانگین پاداش در دو روش یادگیری عمیق کیو و الگوریتم گرادیان سیاست قطعی عمیق در ۲۵ اپیزود در شکل ۹ نمایش داده شده است.

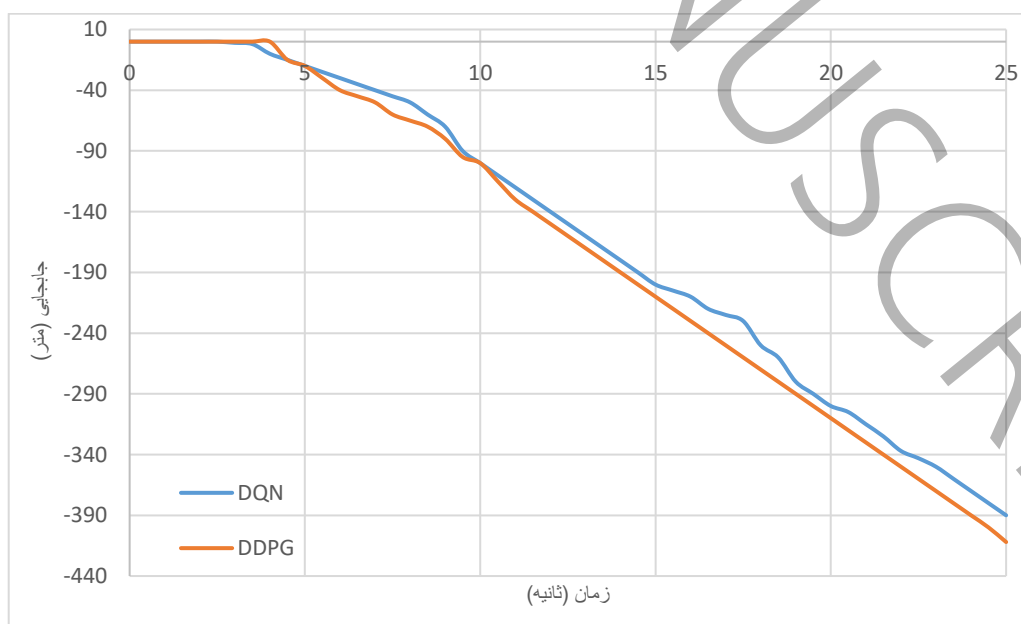


شکل ۹: پاداش میانگین در دو روش یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق

Fig.9 Average reward in DDPG and DQN methods

روند افزایشی این منحنی‌ها گویای عملکرد بهتر عامل در تعامل با محیط می‌باشد. همچنین روند کاهشی منحنی ناشی از احتمال برخورد عامل با خودروهای اطراف و عبور از خطوط موجود در بزرگراه در طی انجام مانور سبقت‌گیری می‌باشد که این امر ناشی از محیط ترافیکی پیچیده طراحی شده در بزرگراه است. طبق شکل ۹، میزان یادگیری الگوریتم گرادیان سیاست قطعی عمیق برای انجام مانور سبقت‌گیری طراحی شده در محیط بزرگراه بهتر از الگوریتم یادگیری عمیق کیو می‌باشد.

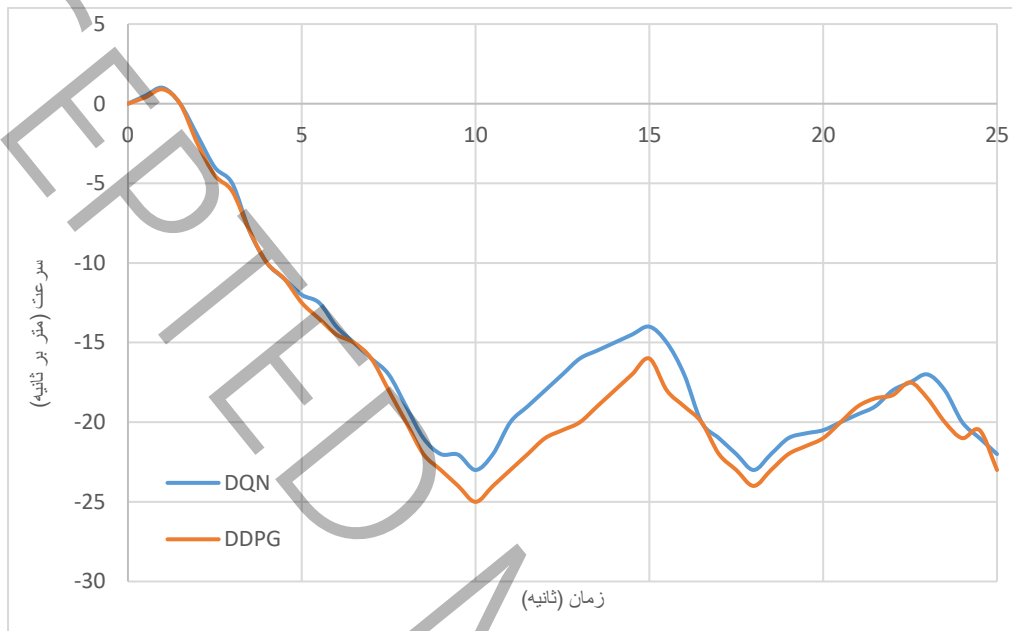
از آنجایی که سرعت و مسافت خودرو به عنوان متغیرهای حالت در این کار انتخاب شده‌اند، شکل ۱۰ مقادیر فاصله عامل با استفاده از دو روش یادگیری عمیق کیو و الگوریتم گرادیان سیاست قطعی عمیق همچنین شکل ۱۱ نیز به ترتیب مقادیر سرعت طولی عامل با استفاده از دو روش یادگیری عمیق کیو و الگوریتم گرادیان سیاست قطعی عمیق را در مدت زمان ۲۵ ثانیه نشان می‌دهند.



شکل ۱۰: جابه‌جایی عامل در روش یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق

Fig.10 Travel distance in DDPG and DQN methods

طبق شکل ۱۰، جابه‌جایی طولانی‌تر نشان‌دهنده‌ی این است که اقدامات کنترلی انتخاب شده عامل، خودرو را قادر می‌سازد تا طولانی‌تر حرکت کرده و از برخورد جلوگیری کند.



شکل ۱۱: سرعت طولی عامل در روش یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق

Fig.11 Longitudinal velocity in DDPG and DQN methods

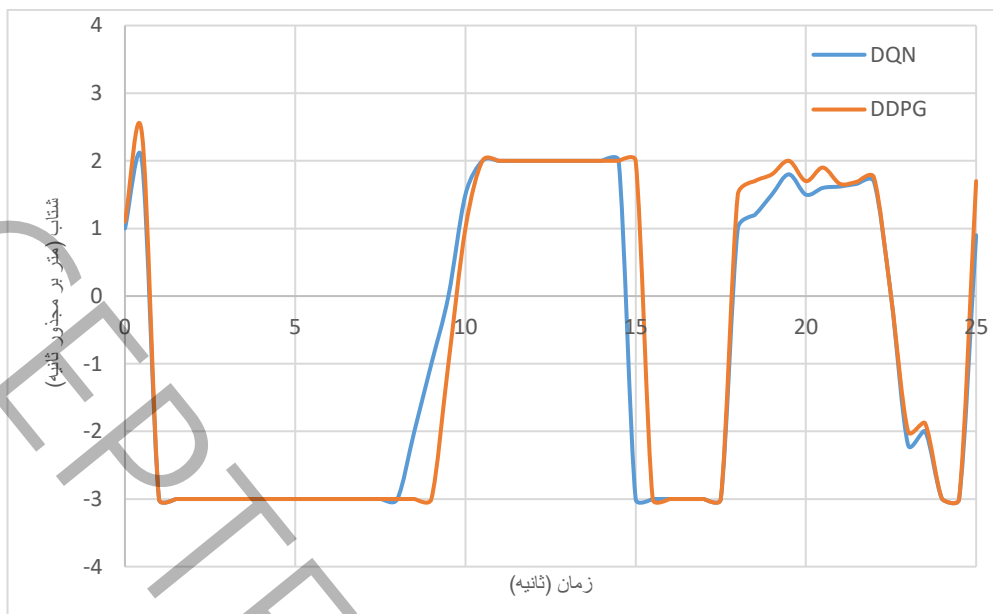
طبق نمودار ۱۱، بزرگی سرعت بالاتر به معنای پاداش‌های بیشتر است. نتایج گویای پاداش‌های بیشتر و در نتیجه پربازده بودن الگوریتم گرادیان سیاست قطعی عمیق نسبت به الگوریتم یادگیری عمیق کیو در محیط بزرگراه می‌باشد.

طبق نتایج شبیه‌سازی حاصل شده از شکل‌های ۹-۱۱، عامل (خودروی خودران) هدایت شده توسط الگوریتم گرادیان سیاست قطعی عمیق می‌تواند به طور موثرتر نسبت به الگوریتم یادگیری عمیق کیو به اهداف ایمنی و کارایی در محیط ترافیکی بزرگراهی دست پیدا کند.

۲-۵ نرخ یادگیری الگوریتم گرادیان سیاست قطعی عمیق

در این بخش، میزان یادگیری و نرخ همگرایی الگوریتم گرادیان سیاست قطعی عمیق ارائه شده مورد بحث و بررسی قرار می‌گیرد. همانطور که گفته شد، هدف اصلی در الگوریتم‌های یادگیری تقویتی عمیق به‌روزرسانی تابع حالت-عمل $Q(s, a)$ به روش‌های مختلف می‌باشد.

در شکل ۱۲ نمودار شتاب طولی با الگوریتم‌های تصمیم‌گیری یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق در مدت زمان ۲۵ ثانیه قابل مشاهده می‌باشد.

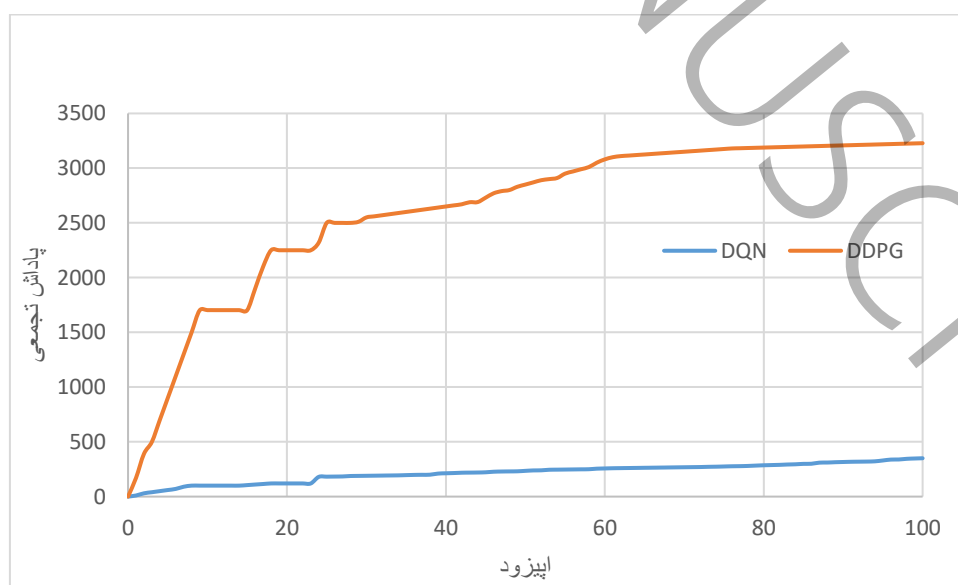


شکل ۱۲: شتاب طولی عامل در روش یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق

Fig.12 Longitudinal acceleration in DDPG and DQN methods

مطابق شکل ۱۲، عامل در الگوریتم گرادیان سیاست قطعی عمیق بیشتر از یادگیری عمیق کیو با محیط رانندگی آشنایی دارد. از اینرو، می‌توان بیان کرد که نرخ همگرایی الگوریتم گرادیان سیاست قطعی عمیق برای مساله تصمیم‌گیری در بزرگراه بهتر از یادگیری عمیق کیو می‌باشد. همچنین در الگوریتم گرادیان سیاست قطعی عمیق مقدار شتاب‌گیری عامل بیشتر از الگوریتم یادگیری عمیق کیو می‌باشد و این موضوع حاکی از میل بیشتر عامل جهت حرکت در لاین سرعت و در نتیجه بازدهی بیشتر الگوریتم گرادیان سیاست قطعی عمیق نسبت به یادگیری عمیق کیو در محیط بزرگراهی می‌باشد.

برای مقایسه نرخ یادگیری الگوریتم‌های گرادیان سیاست قطعی عمیق و یادگیری عمیق کیو، شکل ۱۳ مقدار پاداش‌های تجمعی را در این دو روش نشان می‌دهد.



شکل ۱۳: پاداش تجمعی عامل در روش یادگیری عمیق کیو و گرادیان سیاست قطعی عمیق

Fig.13 Accumulative reward in DDPG and DQN methods

مطابق شکل ۱۳، مقدار پاداش تجمعی در الگوریتم گرادیان سیاست قطعی عمیق همواره بزرگتر از یادگیری عمیق کیو است، که این موضوع نشان‌دهنده برتری سیاست کنترلی اتخاذ شده توسط الگوریتم گرادیان سیاست قطعی عمیق می‌باشد. به عبارت دیگر عامل در روش گرادیان سیاست قطعی عمیق می‌تواند دانش و تجربیات بیشتری در مورد محیط رانندگی بیاموزد. در حقیقت، عامل در الگوریتم گرادیان سیاست قطعی عمیق سیاست کنترل بهینه را سریع‌تر جستجو می‌کند.

در جدول ۲ نتایج حاصل شده از دو الگوریتم گرادیان سیاست قطعی عمیق و یادگیری عمیق کیو با یکدیگر مورد مقایسه قرار گرفته‌اند.

جدول ۲: مقایسه‌ی دو الگوریتم گرادیان سیاست قطعی عمیق و یادگیری عمیق کیو

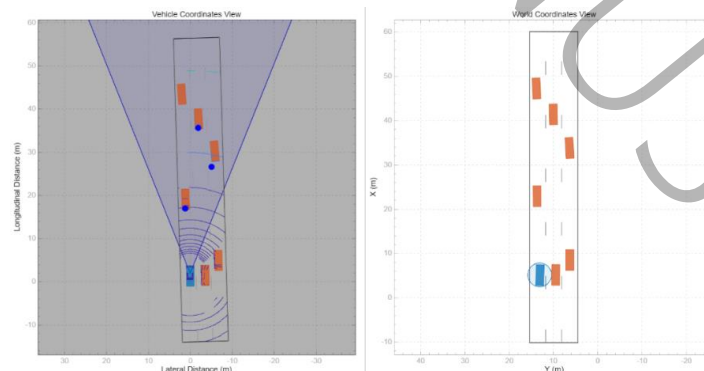
Table2 DDPG and DQN methods comparison

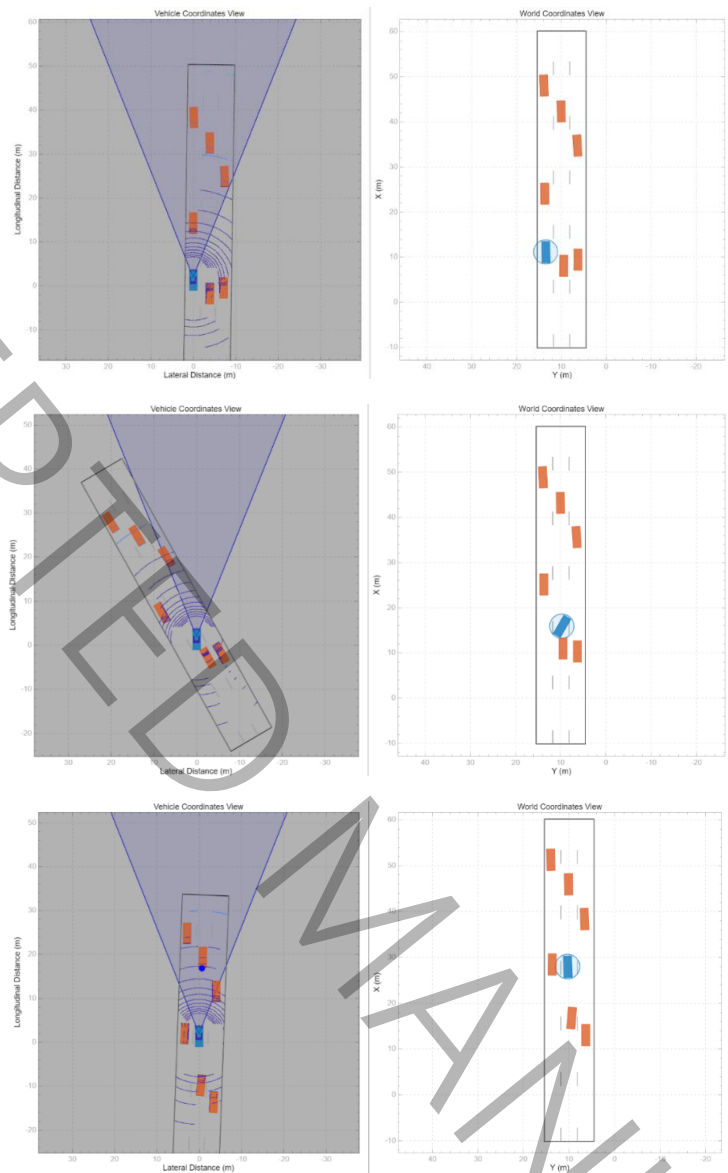
یادگیری عمیق کیو	گرادیان سیاست قطعی عمیق	پارامتر
۳۹۰	۴۱۲	بیشترین مقدار جابه‌جایی (متر)
۲۲/۰۴	۲۵	بیشترین مقدار سرعت طولی (متر بر ثانیه)
۳	۳/۱	بیشترین مقدار شتاب طولی (متر بر مجذور ثانیه)
۲۰	۲۱	بیشترین پاداش میانگین در یک اپیزود

در ادامه برای تشخیص سازگاری الگوریتم گرادیان سیاست قطعی عمیق، با تغییر در پارامترهای موثر در سناریو رانندگی، دو سناریو جدید طراحی شده و عملکرد الگوریتم در دو سناریوی جدید مورد ارزیابی و تبیین قرار خواهد گرفت.

در سناریو اول، موقعیت عامل جهت سبقت‌گیری تغییر داده شده است و عامل در کنار خودروی نارنجی رنگ جهت انجام مانور سبقت‌گیری قرار گرفته است. نتایج حاکی از انجام مانور بدون برخورد می‌باشد، اما به دلیل سبقت‌گیری خطرناکی که توسط عامل انجام شده، به عامل پاداش کمی داده شده است.

این پاداش پایین ناشی از دو عامل می‌باشد: ۱- قطع خطوط توسط عامل در هنگام انجام مانور سبقت‌گیری، ۲- بالا بودن احتمال وقوع تصادف توسط عامل در هنگام انجام مانور سبقت‌گیری در شکل ۱۴ مانور انجام شده توسط عامل قابل مشاهده می‌باشد.

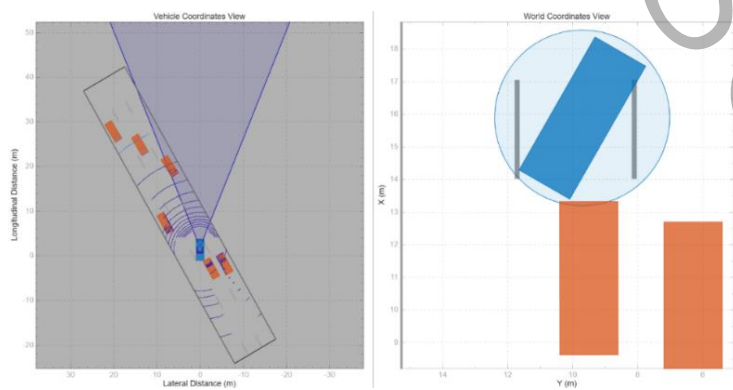




شکل ۱۴: مانور سبقت‌گیری خطرناک انجام شده توسط عامل در محیط بزرگراه

Fig.14 Dangerous overtaking scenario

در شکل ۱۵، علت دریافت پاداش کمتر توسط عامل در هنگام انجام مانور نشان داده شده است.

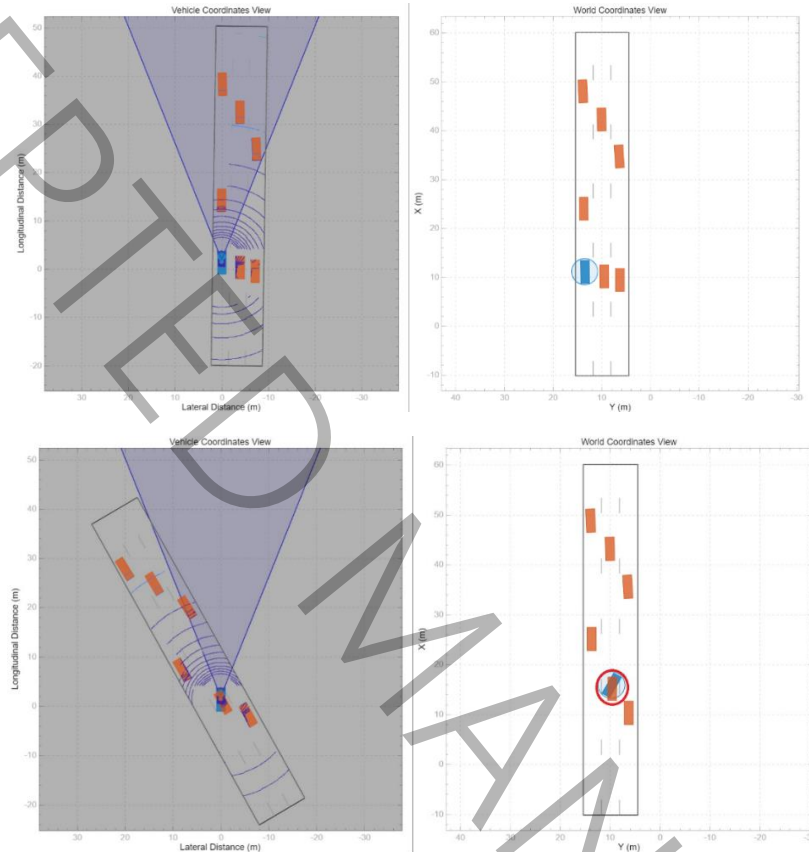


شکل ۱۵: عمل خطرناک انجام شده توسط عامل در محیط بزرگراه

Fig.15 Dangerous action

مطابق شکل ۱۵، عامل در هنگام سبقت‌گیری خط کشی جاده را قطع کرده و همچنین مانور را در وضعیت نزدیک به برخورد انجام داده است.

در سناریو دوم، سرعت خودرویی که از آن توسط عامل سبقت‌گیری می‌شود، افزایش پیدا کرده است، مطابق شکل ۱۶ در این سناریو، برخورد صورت گرفته است و عامل جریمه شده و پاداش منفی دریافت کرده است.



شکل ۱۶: وقوع برخورد در مانور سبقت‌گیری انجام شده توسط عامل در محیط بزرگراه

Fig.15 Collision in the overtaking scenario

برای افزایش توانایی عامل در هنگام مواجهه با شرایطی مشابه دو سناریوی طراحی شده، یا باید روند آموزش را طولانی‌تر کرد تا عامل بتواند دانش بیشتری از محیط‌های رانندگی یاد بگیرد. همچنین می‌توان اطلاعات اطراف را از طریق فناوری ارتباطات در دسترس عامل قرار داد، که به آن کمک کند تا تصمیمات درست را در بزرگراه بگیرد. همچنین می‌توان از دو عامل مجزا برای یادگیری سناریو استفاده نمود به طوری که یک عامل در جهت حرکت طولی خودرو و عامل دیگر در جهت حرکت عرضی خودرو آموزش ببیند، که این موضوع سبب معطوف شدن آموزش هر عامل به یک وظیفه‌ی خاص می‌شود.

۶- نتیجه‌گیری

در این پژوهش، یک الگوریتم تصمیم‌گیری کارآمد و ایمن مبتنی بر روش یادگیری تقویتی عمیق در بزرگراه برای خودروی خودران ارائه شده است. الگوریتم مذکور، گرادیان سیاست قطعی عمیق است. چارچوب کنترلی ساخته شده به سناریوهای رانندگی مشابه با تعداد خطوط مختلف و خودروهای اطراف تعمیم داده شد. نتایج شبیه‌سازی نشان می‌دهد که الگوریتم تصمیم‌گیری پیشنهادی می‌تواند بهینه

بودن و نرخ همگرایی را تضمین کند. همچنین میزان یادگیری الگوریتم پیشنهادی نسبت به الگوریتم یادگیری عمیق کیو بیشتر می باشد.

در کارهای آینده می توان بر روی کاربرد آنلاین سیایت تصمیم گیری پیشنهادی تمرکز نمود. همچنین، محیط متصل را می توان برای به اشتراک گذاشتن اطلاعات با وسایل نقلیه اطراف مورد مطالعه قرار داد. همچنین داده های رانندگی جمع آوری شده در دنیای واقعی را می توان برای ارزیابی عملکرد روش تصمیم گیری پیشنهادی در محیط رانندگی واقعی استفاده نمود.

مراجع

- [۱] A. Raj, J. A. Kumar, and P. Bansal, A multicriteria decision-making approach to study barriers to the adoption of autonomous vehicles, *Transp Res Part A, Policy Pract*, 133 (2020) 122-137.
- [۲] T. Liu, B. Tian, Y. Ai, L. Chen, F. Liu, and D. Cao, Dynamic states prediction in autonomous vehicles: Comparison of three different methods, *IEEE Intell Transp Syst Conf (ITSC)*, (2019) 3750-3755.
- [۳] A. Rasouli and J. K. Tsotsos, Autonomous vehicles that interact with pedestrians: A survey of theory and practice, *IEEE Trans Intell Transp Syst*, 21(3) (2020) 900-918.
- [۴] C. Gkartzonikas and K. Gkritza, What have we learned? A review of stated preference and choice studies on autonomous vehicles, *Transp Res Part C, Emerg Technol.*, 98 (2019) 323-337.
- [۵] C.J. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, and M. J. Kochenderfer, Combining planning and deep reinforcement learning in tactical decision making for autonomous driving, *IEEE Trans Intell Vehicles*, 5(2) (2020) 294-305.
- [۶] C. Yang, Y. Shi, L. Li, and X. Wang, Efficient mode transition control for a parallel hybrid electric vehicle with adaptive dual-loop control framework, *IEEE Trans Veh Technol*, 69(2) (2020) 1519-1532.
- [۷] C.-J. Hoel, K. Wolff, and L. Laine, Tactical decision-making in autonomous driving by reinforcement learning with uncertainty estimation, *IEEE Intelligent Vehicles Symposium (IV)*, (2020) 1292-1298.
- [۸] SAE On-Road Automated Vehicle Standards Committee, Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems, *SAE Standard J*, 3016, (2014) 1-16.
- [۹] Qin, Y., Tang, X., Jia, T., Duan, Z., Zhang, J., Li, Y., & Zheng, L., Noise and vibration suppression in hybrid electric vehicles: State of the art and challenges, *Renewable and Sustainable Energy Reviews*, 124, (2020) 109782.
- [۱۰] Hart, P., & Knoll, A., Using counterfactual reasoning and reinforcement learning for decision-making in autonomous driving, *Journal of Autonomous Vehicles*, 15(2), (2020) 123-145.
- [۱۱] W. Song, G. Xiong, H. Chen, Intention-aware autonomous driving decision-making in an uncontrolled intersection, *Math Problems Eng*, (2016) 1-15.
- [۱۲] Yang, C., You, S., Wang, W., Li, L., & Xiang, C, A stochastic predictive energy management strategy for plug-in hybrid electric vehicles based on fast rolling optimization, *IEEE Transactions on Industrial Electronics*, 67(11), (2020) 9659-9670.
- [۱۳] Furda, A., & Vlacic, L., Enabling safe autonomous driving in real-world city traffic using multiple criteria decision-making, *IEEE Intelligent Transportation Systems Magazine*, 3(1), (2011) 4-17.
- [۱۴] Nie, J., Zhang, J., Ding, W., Wan, X., Chen, X., & Ran, B, Decentralized cooperative lane-changing decision-making for connected autonomous vehicles, *IEEE Access*, 4, (2016) 9413-9420.

- [15] Li, L., Ota, K., & Dong, M, Humanlike driving: Empirical decision-making system for autonomous vehicles, *IEEE Transactions on Vehicular Technology*, 67(8), (2018) 6814-6823.
- [16] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D., Human-level control through deep reinforcement learning, *Nature*, 518(7540), (2015) 529-533.
- [17] Duan, J., Li, S. E., Guan, Y., Sun, Q., & Cheng, B., Hierarchical reinforcement learning for self-driving decision-making without reliance on labeled driving data, *IET Intelligent Transportation Systems*, 14(5), (2020) 297-305.
- [18] Kim, M., Lee, S., Lim, J., Choi, J., & Kang, S. G., Unexpected collision avoidance driving strategy using deep reinforcement learning, *IEEE Access*, 8, (2020) 17243-17252.
- [19] Hang, Q., Lin, J., Sha, Q., He, B., & Li, G., Deep interactive reinforcement learning for path following of autonomous underwater vehicle, *IEEE Access*, 8, (2020) 24258-24268.
- [20] Chen, C., Jiang, J., Lv, N., & Li, S., An intelligent path planning scheme of autonomous vehicles platoon using deep reinforcement learning on the network edge, *IEEE Access*, 8, (2020) 99059-99069.
- [21] Yang, C., Zha, M., Wang, W., Liu, K., & Xiang, C, Efficient energy management strategy for hybrid electric vehicles/plug-in hybrid electric vehicles: Review and recent advances under intelligent transportation system, *IET Intelligent Transportation Systems*, 14(7), (2020) 702-711.
- [22] Han, S., & Miao, F., Behavior planning for connected autonomous vehicles using feedback deep reinforcement learning, *Journal of Autonomous Systems*, 10(3), (2020) 112-134.
- [23] Nagesh Rao, S., Tseng, H. E., & Filev, D, Autonomous highway driving using deep reinforcement learning, In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2019) 2326-2331.
- [24] Li, G., Yang, Y., Zhang, T., Qu, X., Cao, D., Cheng, B., & Li, K, Risk assessment-based collision avoidance decision-making for autonomous vehicles in multi scenarios, *Transportation Research Part C: Emerging Technologies*, 122, (2021) 102820.
- [25] Li, G., Yang, L., Li, S., Luo, X., Qu, X., & Paul, G., Human-like decision-making of artificial drivers in intelligent transportation systems: An end-to-end driving behavior prediction approach, *IEEE Intelligent Transportation Systems Magazine*, 14(1), (2022) 24-36.
- [26] Duan, J., Guan, Y., Li, S. E., Ren, Y., & Cheng, B., Distributional Soft Actor-Critic: Off-Policy Reinforcement Learning for Addressing Value Estimation Errors, *IEEE Transactions on Neural Networks and Learning Systems*, 33(5), (2022) 2345-2357.
- [27] Li, G., Li, S., Li, S., & Qu, X., Continuous decision-making for autonomous driving at intersections using deep deterministic policy gradient, *IET Intelligent Transportation Systems*, 16(2), (2021) 1669-1681.
- [28] Liu, T., Huang, B., Deng, Z., Wang, H., Tang, X., Wang, X., & Cao, D., Heuristics-oriented overtaking decision making for autonomous vehicles using reinforcement learning, *IET Electrical Systems in Transportation*, 1(99), (2020) 1-8.
- [29] Treiber, M., Hennecke, A., & Helbing, D., Congested traffic states in empirical observations and microscopic simulations, *Physical Review E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 62(2), (2000) 1805-1824.
- [30] Zhou, M., Qu, X., & Jin, S., On the impact of cooperative autonomous vehicles in improving freeway merging: A modified intelligent driver model-based approach, *IEEE Transactions on Intelligent Transportation Systems*, 18(6), (2017) 1422-1428.
- [31] Kesting, A., Treiber, M., & Helbing, D, General lane-changing model MOBIL for car-following models, *Transportation Research Record: Journal of the Transportation Research Board*, 1999(1), (2007) 86-94.

[۳۲] Liu, T., Hu, X., Hu, W., & Zou, Y, A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles, IEEE Transactions on Industrial Informatics, 15(12), (2019) 6436-6445.

[۳۳] Liu, T., Tang, X., Wang, H., Yu, H., & Hu, X, Adaptive hierarchical energy management design for a plug-in hybrid electric vehicle, IEEE Transactions on Vehicular Technology, 68(12), (2019) 11513-11522.

[۳۴] Hu, X., Liu, T., Qi, X., & Barth, M, Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: Recent advances and prospects, IEEE Industrial Electronics Magazine, 13(3), (2019) 16-25.

[۳۵] Liu, T., Yu, H., Guo, H., Qin, Y., & Zou, Y, Online energy management for multimode plug-in hybrid electric vehicles, IEEE Transactions on Industrial Informatics, 15(7), (2019) 4352-4361.

[۳۶] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Hoboken, NJ, USA: Wiley, 2014.

[۳۷] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, Dueling network architectures for deep reinforcement learning, in Proc ICML, (2016) 1995-2003.

Highway decision-making strategy for autonomous vehicle for overtaking maneuver using deep reinforcement learning (DRL) method

Ali Rizehvandi ¹, Shahram Azadi ¹

1- Faculty of Mechanical Engineering, K.N.Toosi University of Technology, Tehran, Iran.

*Azadi@kntu.ac.ir

Abstract

Automated driving represents a novel technology aimed at reducing traffic accidents and enhancing driving efficiency. This research introduces a deep reinforcement learning (DRL) approach for autonomous vehicles, focusing on overtaking scenarios on highways. Initially, a highway traffic environment is established, to guide the agent through surrounding vehicles both efficiently and safely. A hierarchical control framework is outlined to manage high-level driving decisions alongside low-level control aspects like car speed and acceleration. Subsequently, a specialized DRL-based method known as Deep Deterministic Policy Gradient (DDPG) is employed to devise decision-making strategies on the highway. The DDPG offers continuous action space exploration, making it suitable for tasks like autonomous driving where actions are not discrete. Unlike DQN, it can handle high-dimensional action spaces more effectively, enhancing its applicability in complex environments like highway driving. The efficacy of the DDPG algorithm is compared to that of the DQN algorithm, with subsequent evaluation of the results. Simulation outcomes demonstrate that the DDPG algorithm adeptly handles highway driving tasks with efficiency and safety. The study underscores the potential of DRL techniques, particularly the DDPG approach, in advancing the capabilities of autonomous vehicles and improving their performance in complex driving scenarios.

Keywords

Autonomous vehicles, Decision making, DRL method, Overtaking, DDPG algorithm.