DOR: https://csj.isi.org.ir/20.1001.1.2538161.1405.11.41.2.5

DOI : https://doi.org/10.22034/csj.2026.251328

تابستان ۱۴۰۵، دوره ۱۱، شماره ۲، پیاپی ۴۱

دریافت مقاله: ۱۴۰۴/۰۸/۱۱

پذیرش مقاله: ۱۴۰۴/۰۹/۲۷

نوع مقاله: پژوهشی

مقایسه عملکرد الگوریتم‌های DQN و Double DQN در محیط‌های یادگیری تقویتی با منابع محدود

سما نوری وراغولی

کارشناسی مهندسی کامپیوتر ، دانشکده مهندسی برق و کامپیوتر ، دانشگاه تبریز ، تبریز ، ایران

پست الکترونیکی: samanit2005@gmail.com

کاویان گچ کار

کارشناسی ارشد مهندسی مخابرات امن و رمزنگاری، پژوهشکده فضای مجازی، دانشگاه شهید بهشتی، تهران، ایران

پست الکترونیکی: k.gachkar@mail.sbu.ac.ir

سینا صمدی قره‌ورن*

دکتری مهندسی برق دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

پست الکترونیکی: s.samadi@tabrizu.ac.ir

کیمیا شیرینی

دکتری مهندسی کامپیوتر، دانشکده چند رسانه ای، دانشگاه هنر اسلامی تبریز، تبریز، ایران

پست الکترونیکی: k.shirini@tabrizu.ac.ir

چکیده

یادگیری تقویتی عمیق طی سال‌های اخیر به‌عنوان یکی از رویکردهای پیشرفته در حل مسائل تصمیم‌گیری پیچیده مطرح شده است و توانسته در حوزه‌هایی مانند بازی‌های ویدئویی، رباتیک و کنترل خودکار عملکردی چشمگیر نشان دهد. با این حال، بررسی کارایی این الگوریتم‌ها در محیط‌هایی با منابع محاسباتی محدود کمتر مورد توجه قرار گرفته است. در این مقاله، دو الگوریتم مطرح الگوریتم یادگیری تقویتی عمیق شامل شبکه کیو عمیق Deep Q-Network (DQN) و شبکه کیو عمیق دوگانه مورد مقایسه و ارزیابی قرار گرفتند. هر دو الگوریتم با بهره‌گیری از شبکه‌های عصبی هم‌آمیختی (CNN) و در محیط ساده و کم منبع گوگل کولَب پیاده‌سازی شدند. آزمایش‌ها بر روی بازی Breakout انجام شد و با تنظیم حداقلی ابرپارامترها، عملکرد هر الگوریتم تحلیل گردید. نتایج نشان دادند که شبکه کیو عمیق دوگانه در برخی شبیه‌سازی‌ها دقت و پایداری بیشتری دارد، در حالی که vanilla DQN با وجود سادگی ساختار خود، همچنان کارایی قابل‌قبولی ارائه می‌دهد. به‌طورکلی، می‌توان نتیجه گرفت که در محیط‌های با محدودیت منابع، استفاده از رویکردهای ساده‌تر الگوریتم یادگیری تقویتی عمیق می‌تواند گزینه‌ای بهینه‌تر و کارآمدتر باشد.

کلید واژه‌ها: یادگیری تقویتی عمیق، شبکه‌های عصبی هم‌آمیختی، شبکه کیو عمیق، شبکه کیو عمیق دوگانه، Breakout

* نویسندهٔ مسئول

۱- مقدمه

در سال‌های اخیر، یادگیری تقویتی عمیق[1] به‌عنوان یک ساختار قوی برای تصمیم‌گیری‌های متوالی شناخته شده است [۱]. این روش با ترکیب قابلیت تصمیم‌گیری ریاضی در یادگیری عمیق و ساختار یادگیری آزمون وخطای یادگیری تقویتی توسعه یافته است [۲]. از زمان معرفی شبکه کیو عمیق[2] توسط منیه و همکارانش در سال ۲۰۱۳، که از شبکه‌های عصبی هم‌آمیختی برای تقریب مقادیر Q مستقیماً از ورودی‌های تصویری با ابعاد بالا استفاده می‌کرد، الگوریتم یادگیری تقویتی عمیق توانست با موفقیت‌های چشمگیری در محیط‌های پیچیده‌ای مانند مجموعه Atari 2600 به دست آورد. توانایی شبکه کیو عمیق در یادگیری منطق کنترل از نقاط تصویر، شروع عصری جدید در توسعه عامل‌های چندمنظوره بود [۳].

با این‌حال، معماری اصلی شبکه کیو عمیق دارای محدودیت‌هایی بود همچون ناپایداری در آموزش، برآورد بیش از حد مقادیر عمل Q به دلیل خاصیت افزایش حداکثری در یادگیری کیو[3] است [٤]. این محدودیت‌ها انگیزه‌ای جدی برای توسعه و طراحی نسخه‌های پیشرفته شبکه کیو عمیق ایجاد کرده‌اند. شبکه کیو عمیق دوگانه مشکل برآورد بیش از حد را با جداکردن دو مرحله انتخاب عمل و مرحله ارزیابی عمل در به‌روزرسانی مقدار Q کاهش می‌دهد [٦،٥]

این مقاله بر ارزیابی الگوریتم شبکه کیو عمیق ساده و دوتا از پرکاربردترین نسخه آن، یعنی شبکه کیو عمیق و شبکه کیو عمیق دوگانه تمرکز دارد. عملکرد این الگوریتم‌ها بر روی بازی Breakout از مجموعه Atari 2600 با استفاده از رابط‌های ALE-py و Gymnasium بررسی شده است [۷]. بازی Breakout چالشی قابل‌مشاهده فراهم می‌کند که در آن میزان قابلیت شبکه برای استخراج ویژگی‌ها از نماهای متوالی و همچنین پایداری در تخمین ارزش نقش مهمی ایفا می‌کند در نتیجه می‌تواند محیطی ایده‌آل برای مقایسه شبکه کیو عمیق و شبکه کیو عمیق دوگانه باشد.

بازی Breakout و بازی‌هایی مانند آن تأثیرات گسترده‌ای در علوم‌شناختی دارند؛ به‌طور مثال می‌توان گفت، این نوع بازی‌ها باعث بهبود هماهنگی چشم و دست، افزایش تمرکز بینایی و توانایی کنترل حرکتی شده‌اند. هدف این پژوهش توصیف روابط میان نسخه‌های مختلف یادگیری کیو و ارزیابی مزایا و محدودیت‌های آنها در محیط‌های پویاست. نتایج این مطالعه، دید بهتری از تعادل‌های الگوریتمی در الگوریتم یادگیری تقویتی عمیق ارائه داده و دستورالعمل انتخاب ساختار مناسب در وظایف گوناگون را فراهم می‌کند [۸].

با وجود پیشرفت‌های چشمگیر در حوزه یادگیری تقویتی عمیق، بیشتر پژوهش‌های انجام شده در محیط‌هایی با منابع محاسباتی گسترده (مانند پردازنده‌های گرافیکی قدرتمند یا کارسازهای ابری) صورت گرفته‌اند. در چنین شرایطی، عملکرد الگوریتم‌های پیچیده؛ مانند شبکه کیو عمیق دوگانه و نسخه‌های پیشرفته‌تر آن به‌خوبی به نمایش گذاشته می‌شود. با این‌حال، در بسیاری از کاربردهای واقعی نظیر سامانه‌های تعبیه‌شده، ربات‌های سبک، یا بن‌سازه‌های آموزشی و شبیه‌سازی مبتنی بر ابر، منابع پردازشی، حافظه و زمان آموزش به‌شدت محدود هستند [۹].

این محدودیت‌ها می‌توانند موجب تغییر در رفتار یادگیری، ناپایداری و حتی افت عملکرد الگوریتم‌های پیچیده الگوریتم یادگیری تقویتی عمیق شود. در نتیجه، پرسش کلیدی مطرح می‌شود که آیا الگوریتم‌های ساده‌تر مانند شبکه کیو عمیق می‌توانند در این محیط‌ها عملکردی هم‌تراز یا حتی بهتر از نسخه‌های پیشرفته‌تر خود داشته باشند؟

این مقاله با هدف پاسخ به این پرسش، به مقایسه تجربی الگوریتم‌های شبکه کیو عمیق و شبکه کیو عمیق دوگانه در محیط بازی Breakout تحت شرایط محدود منابع می‌پردازد. در این راستا، تلاش شده است تا تأثیر محدودیت‌های محاسباتی بر پایداری، سرعت یادگیری و

1-Deep Reinforcement Learning (DRL)
2-DQN
3-Q-learning

دقت تصمیم‌گیری این الگوریتم‌ها تحلیل شود. یافته‌های این مطالعه می‌توانند راهنمایی کاربردی برای انتخاب معماری و الگوریتم مناسب در سناریوهای واقعی با محدودیت منابع فراهم کنند.

در این پژوهش، هر دو الگوریتم یادگیری کیو عمیق و شبکه کیو عمیق دوگانه در محیط بازی Breakout از مجموعه Atari ۲۶۰۰ شبیه‌سازی شدند. فرایند آموزش در سامانهٔ گوگل کولَب با ۵۰۰ دورهٔ آموزشی انجام گرفت. در این فرایند از شبکه‌های عصبی هم‌آمیختی برای استخراج ویژگی‌های بصری از نماهای بازی و از مجموعه‌ای از ابرپارامترهای کمینه برای کاهش بار محاسباتی استفاده شد. نتایج شبیه‌سازی نشان داد که الگوریتم شبکه کیو عمیق مضاعف از نظر پایداری در یادگیری و دقت تصمیم‌گیری عملکرد بهتری دارد، در حالی که شبکه کیو عمیق معمولی باوجود سادگی، کارایی قابل‌قبولی را در محیط‌های کم منبع حفظ می‌کند.

نوآوری این پژوهش در آن است که برای نخستین‌بار یک مطالعه تجربی مقایسه‌ای جامع بین دو الگوریتم یادگیری کیو عمیق و یادگیری کیو عمیق مضاعف در شرایط محدود منابع محاسباتی انجام‌گرفته است. بیشتر پژوهش‌های پیشین این الگوریتم‌ها را در بسترهایی با منابع غنی (واحد پردازش گرافیکی و کارسازهای قدرتمند) بررسی کرده‌اند، در حالی که این پژوهش تأکید دارد حتی در محیط‌های سبک و کم منبع نیز می‌توان با تنظیم مناسب ساختار شبکه و اَبَرپارامترها به کارایی قابل‌توجهی دست‌یافت.

ساختار مقاله به این ترتیب تنظیم شده است: در بخش دوم پیشینه و مبانی نظری مرتبط با یادگیری تقویتی عمیق مرور می‌شود، بخش سوم روشگان پژوهش شامل نحوه طراحی مدل، معماری شبکه و تنظیم پارامترها را شرح می‌دهد، بخش چهارم به تحلیل و ارزیابی نتایج اختصاص دارد و در بخش پنجم جمع‌بندی و پیشنهادهای آتی ارائه شده است.

۲- پیشینه پژوهش

یادگیری تقویتی عمیق طی یک دهه اخیر به یکی از موضوعات بنیادین در هوش مصنوعی و علوم تصمیم‌گیری تبدیل شده است. این رویکرد با ترکیب قدرت یادگیری بازخوردی در یادگیری تقویتی[4] و توانایی تعمیم و تقریب توابع در شبکه‌های عصبی عمیق[5]، توانسته است زمینه را برای توسعه عامل‌هایی فراهم کند که بدون نظارت مستقیم انسان، از طریق تعامل با محیط یاد می‌گیرند[۱۰].

مبنای نظری الگوریتم یادگیری تقویتی عمیق به الگوریتم کلاسیک یادگیری کیو بازمی‌گردد که توسط ساتون و بارتو معرفی شد. در این روش، عامل یادگیرنده با تعامل مستمر با محیط و با هدف بیشینه‌سازی مجموع پاداش‌های بلندمدت، تابع ارزش عمل یا Q-Function را یاد می‌گیرد. با این حال، یادگیری کیو در محیط‌های با فضای حالت یا عمل بزرگ دچار مشکل می‌شود، زیرا ذخیره‌سازی و به‌روزرسانی جدول Q برای تمام حالت‌ها و اعمال ممکن عملاً غیرممکن است [۱۱].

برای رفع این محدودیت،Mnih و همکاران در پژوهشی پیشگامانه، شبکه عصبی هم‌آمیختی را برای تقریب تابع Q معرفی کردند و الگوریتم شبکه Q عمیق را توسعه دادند. این مدل توانست در محیط‌های پیچیده‌ای مانند Atari ۲۶۰۰ که ورودی آن نماهای تصویری با ابعاد بالا بود، بدون نیاز به مهندسی ویژگی‌های دستی به تصمیم‌گیری مؤثر بپردازد [۱۲]. موفقیت شبکه کیو عمیق موجب شد تا یادگیری تقویتی عمیق به‌عنوان روشی عملی برای یادگیری سیاست‌های پیچیده در محیط‌های بصری و پویا مورد توجه گسترده قرار گیرد [۱۳].

با وجود این موفقیت‌ها، شبکه کیو عمیق با چالش‌هایی از جمله ناپایداری در فرایند آموزش، تخمین بیش از حد مقادیر Q و حساسیت زیاد به انتخاب ابرپارامترها مواجه بود [۱۴]. پژوهشگران مختلف در سال‌های بعد تلاش کردند با اصلاح معماری و روش به‌روزرسانی شبکه، این

4-Reinforcement Learning
5-Deep Neural Networks

مشـــکلات را برطرف کنند. یکی از مهمترین این اصلاحات، ارائه الگوریتم شبکه کیو عمیق دوگانه توسط وان‌هاست و همکاران بود. در این روش، دو شبکه عصبی جداگانه برای انتخـــاب و ارزیابی عمل به کار گرفته می‌شـــوند تا خطای تخمیـــن بیش از حد کاهش یابد. این تغییر ســـاده اما مؤثر موجب افزایـــش پایداری و دقت آموزش شـــد و الگوریتم شـــبکه کیو عمیـــق دوگانه به یکی از پایه‌هـــای بسیاری از نســـخه‌های پیشـــرفته‌تر الگوریتم یادگیری تقویتی عمیق تبدیل گردید[۱۵].

در ادامه این مسیـــر، پژوهش‌های مختلفی برای بهبود کارایی الگوریتم یادگیری تقویتی عمیق پیشـــنهاد شده‌اند. بـــرای مثـــال، وانگ و همـــکاران با معـــرفی Dueling DQN ساختاری ارائه کردند که تفکیک ارزش وضعیت[۶] و مزیت عمل[۷] را ممکن ساخت و از این طریق، سرعت یادگیری در محیط‌های با پاداش پراکنده را بهبود داد [۱۶]. همچنین، بادیا و همکاران با توسعه Agent57 توانستند عملکردی فراتر از انســـان در بسیاری از بازی‌های آتاری ارائه دهند[۱۷]. در همین راستا، مرورهای جامعی مانند پژوهش فان و همکاران بـــه تحلیل چالش‌های موجود در الگوریتم یادگیری تقویتی عمیق از جملـــه محدودیت منابع داده، ناپایداری گرادیان و وابستگی زیاد به تنظیمات ابرپارامترها پرداخته‌اند [۱۸].

در کنـــار پیشـــرفت‌های الگوریتمی، مســـئله معماری شـــبکه‌های عصبی مورداســـتفاده در الگوریتـــم یادگیری تقویـــتی عمیق نیز بسیـــار موردتوجه قرار گرفته اســـت. استفاده از شبکه‌های هم‌آمیختی برای استخراج ویژگی‌های فضایی از نماهای تصویـــری در محیط‌های بصری مانند آتاری نتایج بسیار مؤثری داشـــته است [۱۹]. با این حال، از آنجا که شـــبکه عصبی هم‌آمیختی معمولاً وابستگی‌های زمانی بین نماها را نادیده می‌گیرند، پژوهشگران مختلفی از شبکه‌های بازگشـــتی مانند LSTM برای ثبت وابستگی‌های زمانی در الگوریتم یادگیری تقویتی عمیق استفاده کرده‌اند [۲۰]. این تلاش‌ها زمینه‌ساز توسعه مدل‌های چندعاملی و چندلایه در محیط‌های پیچیده‌تر بوده است.

از نظر بســـتر آزمایشی، بلمار و همـــکاران با معرفی محیط یادگیـــری آرکید (ALE) امکان ارزیابی اســـتاندارد الگوریتم‌های یادگیـــری تقویتی در محیط‌های بازی آتاری را فراهم کردند. این بن‌ســـازه تا امـــروز یکی از مهمترین معیارها برای ســـنجش عملکرد الگوریتم یادگیری تقویتی عمیق محســـوب می‌شـــود. در میان این بازی‌ها، Breakout بـــه دلیل ماهیت پویا، ســـاختار پاداش تدریـــجی و نیاز به هماهنگی دقیق بین بینایی و تصمیم‌گیری، به‌عنوان محیطی ایده‌آل بـــرای ارزیـــابی مدل‌های یادگیـــری تقویتی عمیق شناخته شده است [۲۱].

با وجود پیشرفت‌های چشمگیر در طراحی الگوریتم‌ها و شـــبکه‌های عصبی، بیشـــتر پژوهش‌ها بر محیط‌هایی با منابع محاسباتی غنی (GPU و کارسازهای قدرتمند) تمرکز داشـــته‌اند. در حالی که در بســـیاری از کاربردهای واقعی مانند سیستم‌های تعبیه‌شده، ربات‌های سبک، سامانه‌های آمـــوزشی و شبیه‌ســـازهای برخـــط، محدودیت‌هـــای محاســـباتی، حافظه و توان پردازشی وجود دارد. بررسی منابع اخیر نشان می‌دهد که پژوهش‌های اندکی به ارزیابی عملکرد الگوریتم یادگیری تقویتی عمیق در چنین شرایطی پرداخته‌اند. برای نمونه، در مطالعاتی که بر بهبود پایداری تمرکز داشـــته‌اند، معمولاً منابع ســـخت‌افزاری قدرتمند در دســـترس بوده و اثر محدودیت منابع بر دقت و همگرایی الگوریتم‌ها نادیده گرفته شده است [۲۲].

بنابراین، یک شـــکاف پژوهشی آشـــکار در این حوزه وجـــود دارد: تأثیر محدودیت منابع محاســـباتی بر رفتار و کارایی الگوریتم‌هـــای یادگیـــری تقویـــتی عمیق، به‌ویژه شـــبکه کیو عمیق و شـــبکه کیو عمیق دوگانه، به‌صورت نظام‌مند بررسی نشـــده است. این شـــکاف می‌تواند منجر به تفاوت قابل‌توجهی میان عملکرد الگوریتم‌ها در شـــرایط آزمایشگاهی و محیط‌های واقعی شود.

در این راســـتا، مقاله حاضر با هدف پر کردن این خلأ علمی، به مطالعه مقایســـه‌ای عملکرد الگوریتم‌های شـــبکه

6-State Value
7-Advantage Function

جدول ۱- مروری بر پژوهش‌های مرتبط در حوزه یادگیری تقویتی عمیق

مرجع	الگوریتم مورداستفاده	محیط آزمایشی	هدف پژوهش	نقاط قوت	محدودیت‌ها
[۱۸]	DQN	Atari 2600	معرفی شبکه Q عمیق برای یادگیری مستقیم از ورودی تصویری	یادگیری از نقاط تصویری خام، دقت بالا در کنترل	تخمین بیش از حد مقدار Q، ناپایداری آموزش
[۱۹]	Double	Atari / ALE	کاهش خطای برآورد Q در DQN	پایداری بالاتر و دقت بیشتر	پیچیدگی بیشتر در پیاده‌سازی
[۲۰]	Dueling DQN	Atari / Gym	تفکیک ارزش وضعیت و مزیت عمل	سرعت یادگیری بهتر در پاداش‌های پراکنده	افزایش بار محاسباتی
[۲۱]	Agent57	Atari benchmark	دستیابی به عملکرد فراتر از انسان در بازی‌های آتاری	بالاترین دقت یادگیری	نیاز شدید به منابع محاسباتی زیاد
[۲۲]	شبکه عصبی پیچشی در الگوریتم یادگیری تقویتی عمیق	بینایی ماشینی	ارزیابی نقش شبکه عصبی هم‌آمیختی در یادگیری بصری	استخراج ویژگی قوی از تصاویر	نادیده‌گرفتن وابستگی زمانی نماها

کیو عمیق و شبکه کیو عمیق دوگانه در محیط بازی Breakout در شرایط محدود منابع محاسباتی می‌پردازد. این مقاله تلاش دارد تا تأثیر عوامل محدودکننده مانند توان پردازشی، اندازه حافظه بازپخش و تعداد گام‌های آموزش را بر دقت و پایداری مدل‌ها بررسی کند. نتایج حاصل می‌تواند مبنایی برای طراحی الگوریتم‌های سبک‌تر و کارآمدتر در محیط‌های واقعی و کم منبع فراهم کند.

برای درک بهتر روند پیشرفت و چالش‌های موجود در الگوریتم‌های یادگیری تقویتی عمیق، جدول ۱ خلاصه‌ای از مطالعات شاخص انجام شده در این حوزه را ارائه می‌دهد. این مرور نشان می‌دهد که گرچه الگوریتم‌های پیشرفته مانند شبکه کیو عمیق دوگانه و Agent57 عملکرد چشمگیری در محیط‌های پر منبع دارند، اما بررسی کارایی آنها در شرایط محدود منابع کمتر موردتوجه قرار گرفته است.

۲-۱ مبانی یادگیری تقویتی

با افزایش نیاز به کامپیوترها برای یادگیری و انجام وظایف خودکار بدون داده‌های از پیش تعریف شده، یادگیری تقویتی معرفی شد. یادگیری تقویتی گونه‌ای از یادگیری ماشینی است که در آن یک عامل از راه تعامل با محیط می‌آموزد چگونه پاداش‌های دریافتی خود را بیشینه کند [۲۳]. در این مدل عامل محیط را مشاهده کرده و با استفاده از مدل تصمیم‌گیری مارکوف (MDP)، بر اساس داده‌های دریافتی از محیط، عمل مناسب را انتخاب می‌کند. یک MDP توسط تاپل (S, A, P, R, γ) تعریف می‌شود:

S: حالت‌های ممکن

A: اقدامات ممکن

$P(s'|s,a)$: احتمال گذار از حالت S به حالت S' پس از انجام اقدام a

$R(s,a)$: مقدار پاداش پس از انجام اقدام a در حالت

$\gamma \in [0,1]$: ضریب تنزیل برای پاداش‌های درازمدت

هدف اصلی عامل، کشف دستورالعملی $\pi(a|s)$ است که بازده کل موردانتظار را به حداکثر برساند که به‌عنوان مجموع پاداش‌های بلندمدت که در طول زمان طبق معادله ۱ تخفیف داده شده‌اند، تعریف می‌شود.

$$E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_{t}\right] \quad (۱)$$

یک مفهوم کلیدی در یادگیری تقویتی، تابع Q یا تابع

مقدار عمل اســت که تخمیــنی از پاداش‌های جمعی مورد انتظاری را که یک عامل می‌تواند با انجام عمل a در حالت S با پیروی از دســتورالعمل $\pi(a|s)$ به دســت آورد، ارائه می‌دهد.

تابــع بهینهٔ مقــدار – عمل، معادلهٔ بهینــگی Bellman را برآورده می‌کند که حداکثر پاداش موردانتظار از هر جفت حالت – عمل را تعریف می‌کند معادلهٔ ۲ این مفهوم را نشان می‌دهد و تابع بهینهٔ مقدار – عمل و برآورده شــدن آن از معادلهٔ بهینگی Bellman را نشان می‌دهد [۲٤].

$$Q^*(s,a)=E\left[r_t+\gamma\right]\max Q^*(s_{t+1},a')\left|s_t=s,a_t=a\right| \quad (۲)$$

یک روش قدیمی به نام یادگیری کیو از معادله ۲ برای محاسبه $Q(s,a)$ اســتفاده می‌کند، اما با فضاهای حالت بزرگ یا پیوسته مشکل دارد. این محدودیت، در کنار ظهور شبکه‌های عصبی عمیق، منجر به توسعه یادگیری تقویتی عمیق شده است.

۲-۲ شبکه کیو عمیق

معرفی شبکه یادگیری کیو عمیق در سال ۲۰۱۳ توسط Minh و همکارانــش، با ترکیب یادگیری کیو با شــبکه‌های عصبی عمیق، پیشــرفت قابل‌توجــهی در یادگیری تقویتی ایجاد کرد. شبکه کیو عمیق تقریبی از تابع Q را با استفاده از یک شبکه عصبی $Q(s,a;\theta)$ محاسبه می‌کند که در آن θ ، پارامترهای مدل را نشان می‌دهد [۲۵]. تابع loss مورد استفاده در آموزش این شبکه‌ها مطابق معادله ۳ محاسبه می‌شود.

$$L(\theta)=E_{(s,a,r,s')}\left[(y_t^{DQN}-Q(s,a;\theta)^2)\right] \quad (۳)$$

که در آن هدف y_t^{DQN} را می‌توان با معادله ٤ محاسبه کرد.

$$y_t^{DQN}=r+\gamma\max_{a'}Q(s',a';\theta^-) \quad (٤)$$

۲-۳ شبکه کیو عمیق دوگانه

مشکل تخمین بیش از حد[8] در شبکه کیو عمیق که ناشی از اســتفاده از یک شــبکه برای انتخاب و ارزیابی اقدامات بود، بعداً با معرفی شبکه کیو عمیق دوگانه که یک محاسبه هدف جدا شده را با معادله ٥ پیشنهاد می‌کرد، کاهش یافت [۲٦].

$$y_t^{DoubleDQN}=r+\gamma Q(s',\arg\max_{a'}Q(s',a';\theta);\theta^-) \quad (٥)$$

۲-۴ معماری‌های یادگیری عمیق

انتخاب یک معماری عصبی مناسب برای محیط، به‌ویژه هنگام برخورد با ورودی بصــری، در الگوریتم یادگیری تقویتی عمیق بسیار مهم است. در شبکه کیو عمیق اصلی، از شــبکه عصبی هم‌آمیختی برای اســتخراج ویژگی‌های محیطی از نماهای تصویر خام استفاده می‌شد. شبکه های عصبی هم‌آمیختی در مقایســه با سایر معماری‌ها، به دلیل توانایی‌شــان در بهره‌برداری از الگوهای مکانی و محلی، برای این کار بسیار مناسب هستند. عیب شبکه های عصبی هم‌آمیخــتی از پردازش نماها به‌صورت مســتقل یا فقط با جمع‌آوری کوتاه‌مــدت (به‌عنوان‌مثــال، ورودی ٤ نمایی) ناشی می‌شــود که توانایی آنها را در ثبت وابســتگی‌های زمانی محدود می‌کند [۲۷].

۲-۵ محیط‌های شبیه‌سازی بازی (داده)

به‌منظــور مقایســه عملکرد هــر الگوریتــم یادگیری تقویــتی عمیــق ، از معیارهای مختلفی بــرای ارزیابی این پیاده‌سازی‌ها اســتفاده می‌شــود. محیط یادگیری آرکید (ALE)، به‌ویژه بازی‌های آتاری ۲٦۰۰، به بستری گسترده بــرای ایــن ارزیابی‌ها تبدیل شــده اســت. در میــان این محیط‌ها، Breakout بــه دلیل ماهیت پویا، پیچیدگی بصری

8-Overestimation

شکل ۱- تصویری از نما بازی Breakout

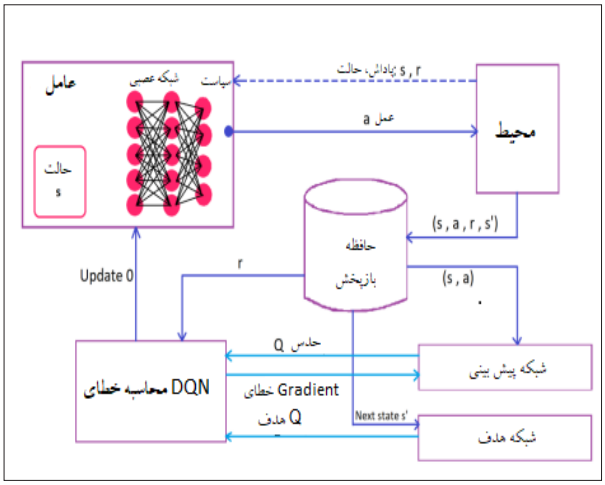
و ساختار پاداش آن که نیازمند دقت و راهبرد بلندمدت است، در این مطالعه انتخاب شده است. با شبیه‌سازی این بازی، این مطالعه نحوه عملکرد الگوریتم‌های مختلف الگوریتم یادگیری تقویتی عمیق را در یک محیط بصری چالش‌برانگیز ارزیابی می‌کند [۲۸].

۳- روشگان پژوهش

در این بخش، نحوه طراحی مدل و چگونگی پیاده‌سازی الگوریتم‌های یادگیری تقویتی عمیق شامل شبکه کیو عمیق و شبکه کیو عمیق دوگانه تشریح می‌شود. هدف، مقایسه عملکرد این دو الگوریتم در محیط‌های با منابع محاسباتی محدود است. به‌منظور دستیابی به نتایج دقیق، مراحل آماده‌سازی داده‌ها، معماری شبکه و تنظیم اَبَرپارامترها به تفکیک توضیح داده شده‌اند.

۳-۱ طراحی مدل پژوهش یا شرح مدل پیشنهادی

برای ارزیابی و سنجش عملکرد الگوریتم‌ها، محیط بازی Breakout از مجموعه ALE-Py به‌عنوان محیط شبیه‌سازی انتخاب شد. این محیط به دلیل پیچیدگی بصری و ساختار پاداش تدریجی خود، بستر مناسبی برای بررسی کارایی عامل‌های یادگیری تقویتی فراهم می‌کند[۲۹]. هر دو الگوریتم در بن سازهٔ Google Colab پیاده‌سازی و آموزش داده شدند. شکل ۱ نمایی از محیط بازی مورداستفاده را نشان می‌دهد.



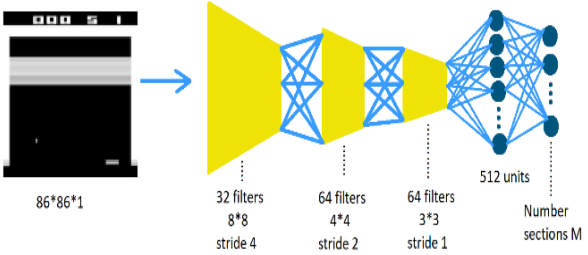
شکل ۲- عملکرد یادگیری کیو عمیق

۳-۲ معماری مدل

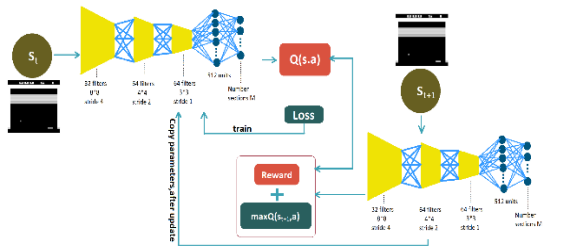
در معماری مدل، یک شبکه عصبی هم‌آمیخت مورد بهره‌برداری قرار گرفته است. شبکه عصبی هم‌آمیخت، یک رمزگذار بصری است که از سه‌لایه هم‌آمیختی و به دنبال آن لایه‌های متراکم به‌عنوان معماری پایه تشکیل شده است. شکل ۲ فرایند یادگیری شبکه شبکه کیو عمیق را نشان میدهد؛ در این چارچوب عامل محیط را نظاره کرده، عمل مبنی بر سیاست خود را انتخاب نموده و در نهایت پاداشی از محیط دریافت میکند. نتیجه هر عملکرد (شامل حالت، عمل، پاداش و حالت بعدی) در حافظه بازپخش، جهت پایداری یادگیری، ذخیره میشود. شبکه یادگیری کیو عمیق، وزنهای خود را با کمینه‌کردن خطای بین مقدار Q پیش‌بینی‌شده و مقدار Q هدف محاسبه شده، با استفاده از معادله Bellman بروزرسانی میکند [۳۰]. شکل ۳ (الف) نمایانگر ساختار کلاسیک شبکه کیو عمیق و شکل ۳ (ب) معماری نمایانگر ساختار شبکه کیو عمیق دوگانه را ارائه می‌دهد.

۳-۳ آماده‌سازی داده و آموزش

جهت آموزش از چهارچوب PyTorch و طرح مینیمال سامانه گوگل کولَب مورد استفاده قرار گرفته است که مشخصات آن به شرح ذیل است:



شکل ۳ (الف) – ساختار کلاسیک شبکه کیو عمیق



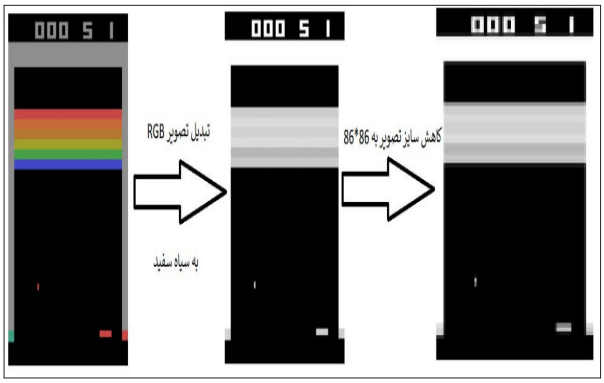
شکل ۳ (ب) – معماری نمایانگر ساختار شبکه کیو عمیق دوگانه

شکل ۳– الگوریتم‌های مورد استفاده در آموزش

در مرحله ورودی تصاویر سه‌کاناله RGB به نماهای تصویر Grayscale تبدیل شده و سپس به ابعاد ۸۶×۸۶ نقطه تصویری تغییر اندازه داده شده و برای ثبت تاریخچه اخیر روی‌هم انباشته شدند. جدول ۲ ابرپارامترهای مورد استفاده برای آموزش عامل‌ها و پیش‌پردازش نما در طول تجسم را نشان می‌دهد.

جدول ۲– ابرپارامترهای مورد استفاده

Parameter	Value
Gamma	۱.۰
Starting epsilon	۰.۰۱
Epsilon decay	۲۵۰۰۰
Learning rate	۴-۱e
Batch size	۳۲
Buffer size	۱۰۰۰
Target update	۱۰۰
Episodes	۵۰۰
max steps	۱۰۰۰
Frame skipping	۴
Optimizer	Adam



شکل ۴– فرایند پیش پردازش نما ورودی به مدل

در این مقاله داده‌های کسب شده از محیط Breakout به‌صورت (وضعیت، عمل، پاداش) در نظر گرفته شده است. وضعیت، حالت محیط فعلی را در نظر می‌گیرد (مثلاً در Breakout مقادیر پیکسل هر نما است). عمل تصمیمی است که توسط محیط پذیرفته شده و باید در نما فعلی اعمال شود و پاداش، مقدار Q است که با انجام عمل A در حالت S به دست می‌آید.

در مرحله پیش‌پردازش روش‌های مختلفی استفاده می‌شود [۳۱] در شروع فرایند , برای کاهش هزینه‌های محاسباتی، از پرش نما استفاده می‌شود. در مرحله بعد نما به مقیاس خاکستری تبدیل می‌شود تا مقادیر هر نقطهٔ تصویری نرمال‌سازی شود و در نهایت اندازه هر نما به ۸۶×۸۶ نقطهٔ تصویری کاهش می‌یابد تا ورودی مدل یادگیری عمیق به حداقل برسد. شکل ٤ مراحل پیش پردازش در محیط Breakout را نشان می‌دهد. این مراحل پیش‌پردازش کارایی اموزش را بهبود می‌بخشد و در عین حال اطلاعات بصری ضروری و نمادین کلیدی مورد نیاز Breakout را حفظ می‌کند.

۴– ارزیابی نتایج

این بخش نتیجه ارزیابی انجام شده در محیط مورداستفاده را ارائه می‌دهد. آموزش بر روی یک ایستگاه کاری Google colab free plan که فقط CPU آن در دسترس بود، اجرا شد. پیاده‌سازی‌ها با استفاده از پایتون ۳.۱۰، با PyTorch به‌عنوان چارچوب اصلی یادگیری عمیق و

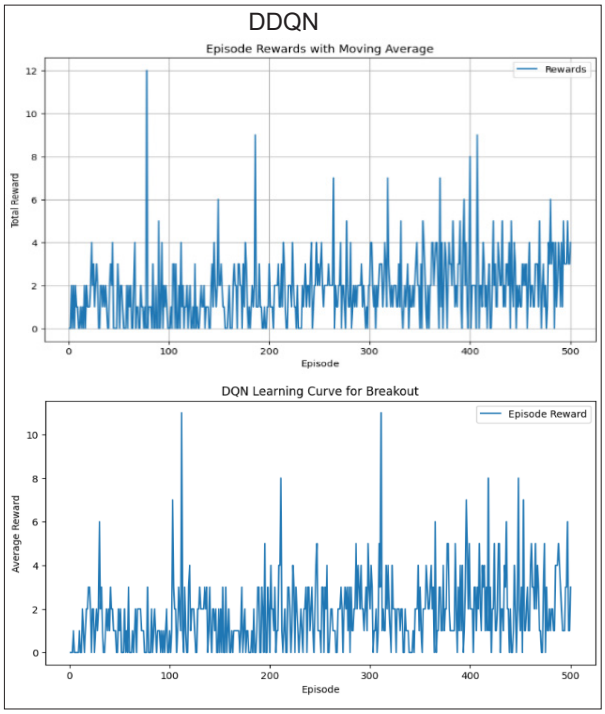
جـدول ۳- منابع سـختافزاری و محاسـباتی مورد استفاده

پارامتر	مقدار
پردازنده	Colab free plan CPU
حافظهٔ اصلی	12.7GB
حافظه سیستم	107.7GB
چهارچوب مورداستفاده	OpenAI Gym and Ale-Py
زبان برنامه‌نویسی	Python
کتابخانه یادگیری عمیق	PyTorch

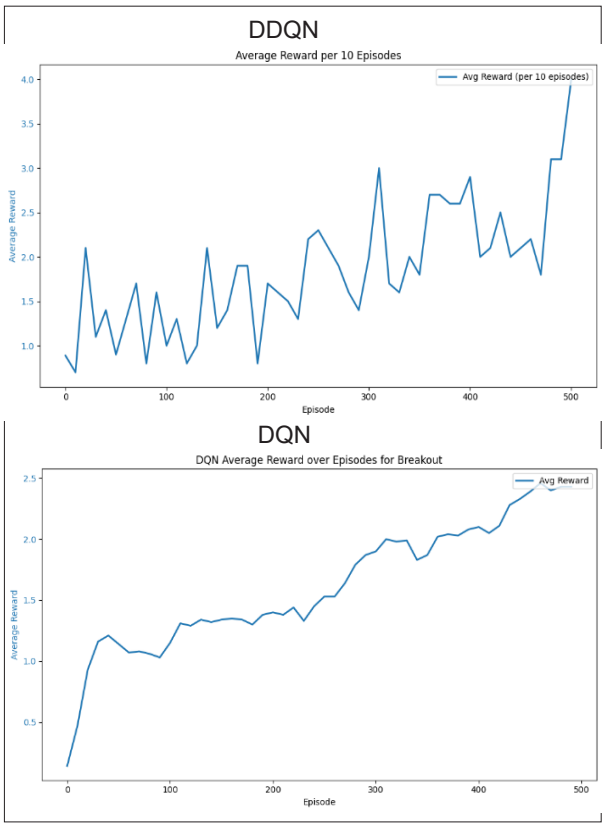
OpenAI Gym و ALE-py برای شبیه‌سـازی محیط، توسعه داده شـدند. جـدول ۳ خلاصه‌ای از منابع سـختافزاری و محاسـباتی مورد اسـتفاده برای آمـوزش و ارزیابی الگوریتم‌های یادگیری تقویتی عمیق را ارائه می‌دهد.

در بخـش بررسی عملکرد الگوریتم‌هـا، برای ارزیابی اثربخشی نسبی الگوریتم‌های یادگیری تقویتی، الگوریتم‌های شـبکه کیو عمیق و شـبکه کیو عمیق دوگانه را در محیط Breakout از مجموعه ALE-Py با معماری‌های شـبکه‌های عصبی هم‌آمیختی مقایسه کردیم. معیار ارزیابی، میانگین پاداش دوره‌ای در طول فرایند آموزش بود. نتایج نشـان می‌دهد که میانگین پاداش الگوریتم شـبکه کیو عمیق برابر با ۱/٤۱ و برای الگوریتم شـبکه کیو عمیق دوگانه برابر با ۱/۸۱ بوده اسـت که بیانگر عملکرد بهتر شبکه کیو عمیق دوگانه نسبت به شبکه کیو عمیق در این محیط است. برای تجسم بیشـتر فرایند تصمیم‌گیری هر الگوریتم، نموداری برای مرتبط کردن مقادیر Q با حالت‌های انتخاب شده رسم شد.

در شـکل ۵، روند تغییر مقدارQ در طول آموزش برای دو الگوریتم شـبکه کیو عمیق و شـبکه کیو عمیق دوگانه نمایش داده شـده اسـت. همان‌طور که مشاهده می‌شود، در هر دو روش نوسـانات قابل‌توجـهی در مقادیر پاداش دوره‌ها دیده می‌شود که ناشی از ماهیت تصادفی محیط و



شکل ۵- روند تغییر مقدار Q در آموزش



شکل ۶- روند تغییر مقدار میانگین Q در حین آموزش

فرایند اکتشاف است. با این حال، الگوریتم شبکه کیو عمیق دوگانه در مقایسـه با شبکه کیو عمیق رفتار باثبات‌تری از

خود نشـــان می‌دهد و دامنه نوسانات آن نسبتاً کمتر است. این امر نشان می‌دهد که استفاده از دو شبکه در شبکه کیو عمیق دوگانه باعث کاهش بیش‌برآورد مقدار Q شده و در نتیجه یادگیری پایدارتری حاصل شده است.

در شـــکل ٦، روند تغییر مقدار میانگین پاداش در طول دوره‌ها بـــرای دو الگوریتم نشان‌داده‌شـــده اســـت. نتایج بیانگر آن اســـت که هـــر دو روش به‌مـــرور زمان بهبود عملکرد را تجربه می‌کنند، اما نرخ رشـــد میانگین پاداش در شبکه کیو عمیق دوگانه ســـریع‌تر و یکنواخت‌تر از شبکه کیو عمیق اســـت. این موضوع بیانگر کارایی بهتر شـــبکه کیو عمیـــق دوگانه در تقریب مقدار واقـــعی Q و در نتیجه در تصمیم‌گیـــری بهینه‌تر عامل اســـت. در نهایت، می‌توان نتیجه گرفت که الگوریتم شبکه کیو عمیق دوگانه با کاهش بیش‌برآورد مقادیر Q، پایداری و کارایی بالاتری نسبت به شبکه کیو عمیق در فرایند آموزش ارائه می‌دهد.

۵- نتیجه‌گیری

این مطالعه مقایســـه عملکـــرد الگوریتم‌هـــای DQN و Double شـــبکه کیو عمیق را که هر کـــدام با معماری‌های شـــبکه‌های عصبی هم‌آمیختی پیاده‌ســـازی شـــده‌اند، در محیط‌هـــای با منابع محدود بررسی کرد. نتایج نشـــان داد که علی‌رغم سادگی، شـــبکه کیو عمیق معمولی پایدارتر و رقابتی‌تر است، حتی زمانی که محیط توسط منابع محدود می‌شود و از این‌رو تنظیم ابرپارامتر را نیز محدود می‌کند، با همتای پیچیده‌تر خود برابری می‌کند. در حالی که شبکه کیـــو عمیق دوگانه از نظر نظری مزایایی در کاهش تخمین بیش از حد دارد، اثرات آن در محیط‌های کمینه‌گرا به طور قابل‌توجهی کاهش می‌یابد. نتایج یافته‌ها نشـــان می‌دهد که در محیط‌هایی که محدودیت‌ها بر بودجه‌های محاســـباتی اعمال می‌شوند، الگوریتم‌هایی با معماری ساده‌تر می‌توانند در مقایسه با اتخاذ انواع پیشرفته که از منابع غنی‌تر بهره می‌برند، نتایج یادگیری کارآمدتری داشته باشند.

این یافته‌هـــا بر اهمیت در نظر گرفتـــن محدودیت‌های عملیـــاتی محیط هنـــگام انتخـــاب الگوریتم‌هـــای یادگیری تقویـــتی تأکیـــد می‌کننـــد. در کاربردهای عمـــلی که منابع محاســـباتی، ظرفیت حافظه و زمان تنظیم محدود هستند- مانند سیســـتم‌های تعبیه‌شده، شبیه‌ســـازهای آموزشی یا بن‌ســـازه‌های سبک مبتنی بر ابر- استفاده از الگوریتم‌های ســـاده‌تر ماننـــد Vanilla شـــبکه کیو عمیـــق می‌تواند نتایج قابل‌اعتمادتـــر و تکرارپذیرتـــری ارائـــه دهـــد. این مطالعه همچنین تأکید می‌کند که پیچیدگی معماری، اگر چه از لحاظ نظری مفید اســـت، اما ممکن است در شـــرایط محدود به افزایش عملکرد عملی منجر نشود. در نتیجه، در محیط‌هایی که منابع محدود هســـتند، ترکیب الگوریتم‌های ســـاده‌تر با معماری‌های شـــبکه‌ای با پیچیدگی کمتـــر می‌تواند تعادل بهتری بین عملکـــرد و کارایی محاســـباتی ایجاد کند. این تحقیق یک اصل راهنما برای طراحی سیستم‌های یادگیری تقویتی عمیق در ســـناریوهای با محدودیت منابع پیشنهاد می‌کند: اولویت‌دادن به‌ســـادگی و پایـــداری الگوریتمی بر پیچیدگی معمـــاری. کارهای آینده می‌تواننـــد راهبردهای تطبیقی را بررسی کنند که به‌صورت پویا شبکه یا استفاده از میانگیر بازپخش را بر اســـاس در دسترس بودن منابع تنظیم می‌کنند و به‌طور بالقوه به الگوریتم‌های پیشـــرفته‌تر اجازه می‌دهند بدون نیاز به بودجه‌های محاسباتی گسترده، پایداری را حفظ کنند.

مراجع

1. Oudouar, F., Bir-Jmel, A., Grissette, H., Douiri, S. M., Himeur, Y., Miniaoui, S., Atalla, S., & Mansoor, W. 2025. An empirical evaluation of neural network architectures for 3D spheroid segmentation. Computers, 14, 86.
2. Aloufi, N., & Aljuhani, A. 2025. Empirical evaluation of prompting strategies for Python syntax error detection with LLMs. Applied Sciences, 15, 9223.
3. Donahue, J., et al. 2015. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
4. Shaheen, A., Badr, A., Abohendy, A., Alsaadawy, H., & Alsayad, N. (2025). Reinforcement learning in strategy-based and atari games: A review of google deepminds innovations. arXiv preprint arXiv:2502.10303.
5. Elmezain, M., Saoud, L. S., Sultan, A., Heshmat, M., Se-

traction to synthesize natural speech. Procedia Computer Science, 258, 2737-2747.
20. Van Hasselt, H., Guez, A., & Silver, D. 2016. Deep reinforcement learning with double یادگیری کیو. Proceedings of the AAAI Conference on Artificial Intelligence.
21. Guleria, P., Frnda, J., & Srinivasu, P. N. 2025. NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis. Array, 100467.
22. Tanis, L. J., Cunha, R. F., & Zullich, M. 2025. Bridging faithfulness of explanations and deep reinforcement learning: A Grad-CAM analysis of Space Invaders. In Proceedings of the 20th International Conference on the Foundations of Digital Games, 1-7.
23. Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. Machine Learning, 3, 9-44.
24. Retkowski, F. 2020. Reinforcement learning for sequence-to-sequence dialogue systems. Karlsruhe Institute of Technology, 2020.
25. Mnih, V., et al. 2015. Human-level control through deep reinforcement learning. Nature, 518, 529-533.
26. Gragnaniello, D., Greco, A., Sansone, C., & Vento, B. 2025. FLAME: Fire detection in videos combining a deep neural network with a model-based motion analysis. Neural Computing and Applications, 37, 6181-6197.
27. Cong, R., Yang, N., Liu, H., Zhang, D., Huang, Q., Kwong, S., & Zhang, W. 2025. TRNet: Two-tier recursion network for co-salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2025.
28. Singh, N. J., Singh, K. R., Hoque, N., & Bhattacharyya, D. K. 2025. Massive IoT network traffic analysis using ML and DL methods: An empirical evaluation. The Journal of Supercomputing, 81, 1107.
29. Subramanian, A., Price, S., Kumbhar, O., Sizikova, E., Majaj, N. J., & Pelli, D. G. 2025. Benchmarking the speed–accuracy tradeoff in object recognition by humans and neural networks. Journal of Vision, 25, 4-4.
30. Jang, H., Sinha, P., & Boix, X. 2025. Configural processing as an optimized strategy for robust object recognition in neural networks. Communications Biology, 8, 386.
31. Meng, W., et al. 2025. Advances in UAV path planning: A comprehensive review of methods, challenges, and future directions. Drones, 9, 5.

neviratne, L., & Hussain, I. 2025. Advancing underwater vision: A survey of deep learning models for underwater object recognition and tracking. IEEE Access, 2025.
6. Lipton, Z. C., Berkowitz, J., & Elkan, C. 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019, 2015.
7. Spoerer, C. J., et al. 2020. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. PLoS Computational Biology, 16, e1008215.
8. Penzkofer, A., Schaefer, S., Strohm, F., Bâce, M., Leutenegger, S., & Bulling, A. 2025. Int-HRL: Towards intention-based hierarchical reinforcement learning. Neural Computing and Applications, 37, 18823-18834.
9. Bellemare, M. G., et al. 2013. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47, 253-279.
10. Devlin, J., et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).
11. Yao, J., Zhang, J., & Zhuo, L. 2025. SGNet: Sequence grouping network via vision-language model for text-guided video summarization. IEEE Journal of Selected Topics in Signal Processing.
12. Baek, I.-C., et al. 2025. IPCGRL: Language-instructed reinforcement learning for procedural level generation. arXiv preprint arXiv:2503.12358.
13. Joshi, V., Xu, Z., Liu, B., Stone, P., & Zhang, A. 2025. Benchmarking massively parallelized multi-task reinforcement learning for robotics tasks. arXiv preprint arXiv:2507.23172, 2025.
14. Hasan, M. M., Das, R. K., Hassan, M., Razia, S., Ani, J. F., Khushbu, S. A., & Islam, M. 2025. Hybrid deep learning: A comparative study on AI algorithms in natural language processing for text classification. Bulletin of Electrical Engineering and Informatics, 14, 551-559.
15. Greff, K., et al. 2016. LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28, 2222-2232.
16. Fan, C.-H., Wu, R.-T., & Chang, Y.-I. 2025. Robotic inspection for autonomous crack segmentation and exploration using deep reinforcement learning. Automation in Construction, 106009.
17. Snodgrass, S., & Ontanón, S. 2016. Learning to generate video game maps using Markov models. IEEE Transactions on Computational Intelligence and AI in Games, 9(4), 410–422.
18. Ulaş, B., Szklenár, T., & Szabó, R. 2025. Detection of oscillation-like patterns in eclipsing binary light curves using neural network-based object detection algorithms. Astronomy & Astrophysics, 695, A81.
19. Mohanty, S., Sahoo, D., Das, S., Acharya, A. A., & Panda, N. 2025. Sentiment analysis using CNN for emotion ex-