

پشتیبانی تصمیم گیری مبتنی بر یادگیری تقویتی عمیق در مانور نبردهای هوایی

مهدی ذوالفقاری^{۱*}، حجت نورمحمدی^{۲*}، گنورگ قره‌پتیان^۳، حمید رادمنش^۴

تاریخ دریافت: ۱۴۰۲/۰۲/۲۲

تاریخ پذیرش: ۱۴۰۲/۰۵/۲۸

چکیده

در سال های اخیر، نبردهای هوایی گروهی و جمعی با استفاده از پهپادها در نیروهای مسلح کشورهای توسعه یافته رواج پیدا کرده است. در این گونه از نبردها انتظار می رود که پهپادها، مانورهای چابک و ایمنی را با توجه محیط پویا و پیچیده میدان نبرد انجام دهند. تکنیک یادگیری تقویتی عمیق^۵ (DRL)، که ابزاری مناسب برای تصمیم گیری های پیوسته در محیط های پویا و با عدم قطعیت بالا می باشد، گزینه ای مناسب برای پشتیبانی تصمیم گیری در مانورهای نبردهای هوایی^۶ (ACMD) خواهد بود. علیرغم مطالعات فراوان در سال های اخیر بر روی کارکردهای یادگیری تقویتی در پشتیبانی تصمیم گیری، تاکنون نقشه راهی برای بکارگیری این تکنیک در پشتیبانی تصمیم گیری نبردهای هوایی ترسیم نگردیده است. بنابراین در این مقاله ابتدا تصویر کاملی از موضوعات موجود در این حوزه ترسیم خواهد شد. ابتدا یادگیری تقویتی عمیق تشریح خواهد شد و سپس کارکرد آن در پشتیبانی تصمیم گیری نبردهای هوایی بیان خواهد شد. سپس یک تکنیک تصمیم گیری در نبردهای هوایی یک به یک برای پیروزی پهپادها در مانور نبردهای هوایی ارائه خواهد شد. در طراحی مدل پیشنهادی از الگوریتم DQN که حالت پیشرفته ای از یادگیری تقویتی عمیق است، بهره گرفته شده است. آموزش مدل در سه حالت توازن، ضعف و برتری پهپاد در نبرد هوایی انجام شده است. همچنین یک تابع پاداش متناسب با حوزه پژوهش طراحی شده و در نهایت طراحی و ساخت مدل و ارزیابی آن نیز اجرا گردیده است. در شبیه سازی انجام شده، مدل ایجاد شده با هفت مانور مختلف پهپاد مورد آزمایش و ارزیابی قرار گرفته است. ارزیابی مدل ایجاد شده بیانگر عملکرد مطلوب آن در مدیریت مانورهای پهپادها در نبردهای هوایی پیچیده و با عدم قطعیت بالا دارد.

واژگان کلیدی: یادگیری تقویتی عمیق، مانور هوایی، پشتیبانی تصمیم.

^۱ استاد پژوهشگر همکار، مرکز بهره برداری ایمن شبکه، دانشگاه صنعتی امیرکبیر، تهران (mahdizolfaghari@aut.ac.ir)

^{۲*} مدرس مهندسی نرم افزار، گروه مهندسی کامپیوتر دانشگاه فنی و حرفه ای لرستان، خرم آباد ([hz657091@gmail.com](mailto:haz657091@gmail.com))

^۳ استاد دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر، تهران (grptian@aut.ac.ir)

^۴ دانشیار، دانشکده مهندسی برق، دانشگاه علوم. فنون هوایی شهید ستاری، تهران، ایران (radmanesh@ssau.ac.ir)

^۵ Deep Reinforcement Learning

^۶ Air Combat Maneuver Decision-Making

۱. مقدمه

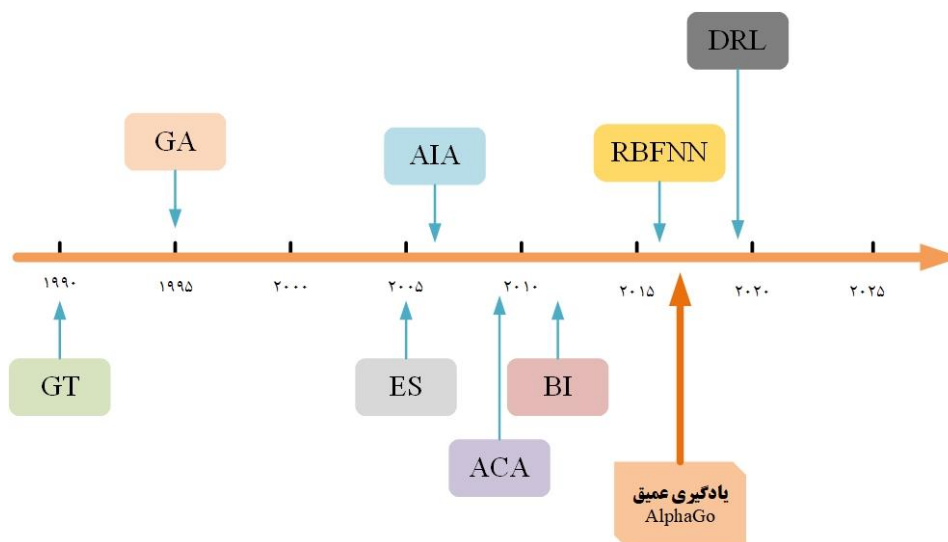
تعیین شده انجام دهد. یک پهپاد با آگاهی موقعیتی کامل، برنامه ریزی مأموریت قوی و توانایی های تصمیم گیری مانور، شانس بیشتری برای پیروزی در نبرد هوایی دارد.

۱-۱. انواع الگوریتم های تصمیم گیری در نبرد

هوایی

الگوریتم های تصمیم گیری جنگ هوایی هوشمند را می توان به چهار دسته تقسیم بندی کرد. اولین دسته، الگوریتم های مبتنی بر نظریه بازی^۷ (GT) است. دومین دسته، الگوریتم های بهینه سازی اکتشافی هستند که شامل الگوریتم ژنتیک^۸ (GA)، الگوریتم کلونی مورچه ها^۹ (ACA)، الگوریتم مصونیت مصنوعی^{۱۰} (AIA) می شود. دسته سوم روش های مبتنی بر دانش مانند استنتاج بیزی^{۱۱} (BI) و سیستم های خبره^{۱۲} (ES) هستند. چهارمین دسته نیز روش های مبتنی بر شبکه عصبی مانند شبکه عصبی تابع پایه شعاعی^{۱۳} (RBFNN) و یادگیری عمیق هستند. سیر زمانی توسعه این الگوریتم ها در شکل ۱ نشان داده شده است.

تکنیک های تصمیم گیری هوشمند برای پهپادها در انجام مأموریت های مختلف (مانند شناسایی، هشدار اولیه، اقدامات متقابل الکترونیکی و غیره) و در عین حال بقای خود ضروری است. [۱] تجهیزات هوایی سرشنین دار نیز می تواند از چنین تکنیک هایی استفاده نمایند. این قابلیت به فرماندهان و خلبانان اجازه می دهد تا به جای برنامه ریزی های پیچیده، انتخاب های خود را از لیست پیشنهادات بهینه انجام دهند و تمرکز خود را بر روی کلیت مأموریت حفظ نمایند. تصمیم گیری هوشمند برای پهپادها از سه مرحله اصلی تشکیل می شود، آگاهی از موقعیت، برنامه ریزی مأموریت و تصمیم گیری مانور. [۲][۳] [۴] [۵] آگاهی از موقعیت به پهپاد این امکان را می دهد تا موقعیت دقیق خود و تهدیدات را در فضای میدان نبرد درک نماید. برنامه ریزی مأموریت، طرح ها و برنامه های عملیاتی مناسبی را متناسب با محیط پویای نبرد ارائه می دهد. تصمیم گیری مانور نیز پهپاد را قادر می سازد تا اقدامات تاکتیکی حیاتی مانند فرار از صحنه یا حمله به اهداف را مطابق سیاست های



شکل ۱. سیر زمانی توسعه الگوریتم های تصمیم گیری

دنبال می کند. نتیجه بازی، ارزش پرداختی شرکت کنندگان است. در این نظریه می توانیم بهترین استراتژی برای یک بازیکن را در مقابل رقبایش تعیین نماییم. [۶] در این رویکرد معمولاً از روش

۱-۱-۱. الگوریتم های مبتنی بر نظریه بازی

نظریه بازی، روشی سیستماتیک برای تجزیه و تحلیل مسائل تصمیم گیری با چندین عامل است که هر عامل هدف خود را

¹¹ Bayesian Inference

¹² Expert Systems

¹³ Radial Basis Function Neural Network

⁷ Game Theory

⁸ Genetic Algorithm

⁹ Ant Colony Algorithm

¹⁰ Artificial Immunity Algorithm

[۱۱] نیز از الگوریتم **AIA** برای توانمند کردن پهپادها برای تسلط بر استراتژی‌های واکنش پیشرفته برای موقعیت‌های مختلف نبرد هوایی استفاده کرده‌اند. آن‌ها پایگاه داده‌ای از سناریوهای موفقیت و شکست مأموریت ایجاد کرده‌اند تا زمانی که پهپاد در آینده با موقعیت‌های مشابه مواجه می‌شود، به سرعت پاسخ دهد. با این حال در حالتی که ابعاد نبرد هوایی بالاست، اکثر الگوریتم‌های بهینه‌سازی اکتشافی نمی‌توانند بصورت بلادرنگ اجرا شوند و با محدودیت‌های عملی سازگار شوند.

۳-۱-۱. الگوریتم‌های مبتنی بر دانش

روش‌های مبتنی بر دانش، حجم زیادی از داده‌ها را ادغام می‌کنند و تجربه را برای سیستم‌های استدلال خودکار فراهم می‌کنند. [۱۲] دانش ارتباط بین داده‌ها در پایگاه داده، به شکل قوانین ارتباطی بیان می‌شود. رایانه از قوانین تداعی برای محاسبه بهترین استراتژی برای موقعیت مربوطه استفاده می‌کند. در مطالعه هوانگ و همکارانش [۱۳]، نبرد هوایی به عنوان یک فرآیند مارکوف مدل‌سازی شده است و از **BI** برای محاسبه وضعیت نبرد هوایی استفاده می‌شود. بر اساس نتایج ارزیابی موقعیتی، وزن عوامل تأثیرگذار به طور انطباقی تنظیم می‌شود تا تابع هدف معقول‌تر شود.

۴-۱-۱. الگوریتم‌های مبتنی بر شبکه‌های عصبی

الگوریتم‌های مبتنی بر شبکه‌های عصبی در پشتیبانی تصمیم‌گیری نبرد هوایی را می‌توان به دو زیرمجموعه، یعنی یادگیری تحت نظارت و یادگیری تقویتی عمیق طبقه‌بندی کرد. هدف یادگیری تحت نظارت پیش‌بینی نتایج با داده‌های برچسب‌گذاری شده است. [۱۴] ژائو و هوانگ اطلاعات پهپاد و دشمن را به عنوان ورودی درخت تصمیم برای ارزیابی وضعیت نبرد هوایی از جنبه‌های مختلف مدل‌سازی کرده‌اند. نتایج مطالعه آن‌ها نشان می‌دهد که روش درخت تصمیم جدید در مقایسه با شبکه سنتی بیزی عملکرد بهتری دارد. [۱۵] یادگیری تقویتی عمیق (**DRL**) یک سیاست تصمیم‌گیری را با مکانیسم

ماتریس بازی و استراتژی تعادل نش^{۱۴} برای اتخاذ تصمیمات هوشمند برای پهپاد استفاده می‌شود. یک برنامه کامپیوتری فرآیند تصمیم‌گیری را در صورت بلادرنگ انجام می‌دهد تا حالت پیش‌بینی شده را محاسبه نماید. سپس امتیازات محاسبه شده برای انتخاب بهترین استراتژی مقایسه می‌شوند. به عنوان مثال در کار [۷]، انهدام مواضع دشمن به عنوان هدف پهپاد در نظر گرفته شده است. این پهپاد از رویکرد تئوری بازی پویا با جمع غیر صفر برای مدل‌سازی تابع هدف، متغیرهای کنترل و متغیرهای حالت استفاده می‌کند و تصمیمات هوشمندانه‌ای برای پهپاد به دست می‌آورد. همچنین در کار [۸]، از یک رویکرد ماتریس بازی برای ایجاد تصمیمات هوشمند برای پرواز پهپادها در ارتفاع کم در نبردهای هوایی یک به یک و در زمین‌های تپه‌ای استفاده شده است. با این حال استقرار مدل‌های ریاضی مبتنی بر نظریه بازی دشوار است و اثبات کارایی و ضرورت راه حل‌های آنها دشوار است.

۲-۱-۱. الگوریتم‌های بهینه‌سازی اکتشافی

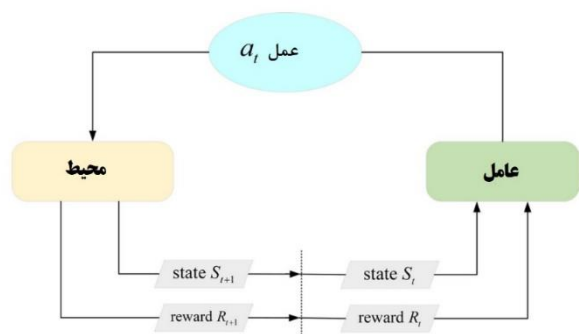
الگوریتم‌های بهینه‌سازی اکتشافی به دنبال راه‌حل‌های عملی خوب برای مسائل بهینه‌سازی در شرایطی هستند که مشکل پیچیده است یا حالتی که زمان برای محاسبه یک راه‌حل دقیق محدود است. [۹] این الگوریتم‌ها مسائل نبرد هوایی را به مسائل بهینه‌سازی نگاشت می‌کنند و راه حل‌های بهینه را پیدا می‌کنند. روند نبرد بین پهپاد و دشمن را می‌توان با تابع هزینه ارزیابی کرد. در نهایت پهپاد بهترین راه حل را دریافت می‌کند. به عنوان مثال، یک الگوریتم ژنتیک همجوشی مجارستانی در کار [۱۰] برای یک مدل تخصیص هدف با چند پهپاد پیشنهاد شده است. هو و همکارانش از یک الگوریتم کلونی مورچه‌های بهبود یافته استفاده کرده‌اند که بهبودهایی را در قوانین انتخاب مسیر، به روز رسانی فرومون و فاصله غلظت فرومون ایجاد می‌کند تا از مشکل بهینه محلی این الگوریتم سنتی جلوگیری کند. آنها مشکل تخصیص سلاح به هدف را در نبرد هوایی مشترک با چند پهپاد حل کرده‌اند و تصمیم‌گیری هوشمندانه چند پهپادی را در میدان جنگ پیچیده و متغیر ایجاد کرده‌اند. کریشناکومار و همکارانش

¹⁴ Nash equilibrium strategy

Trials (ADT) دارپا استفاده کرده است. این روش در رقابت نهایی **ADT** در رتبه دوم قرار گرفت و فارغ التحصیل دوره مربی اسلحه **F-16** نیروی هوایی ایالات متحده را شکست داد. [۱۷] هر یک از این مثال‌ها نشان می‌دهد که **DRL** دارای پتانسیل بالایی در سرعت و دقت تصمیم‌گیری است، بنابراین بسیاری از متخصصان و محققان تلاش کرده‌اند تا **DRL** را در زمینه نبرد هوایی اعمال کنند و به دستاوردهای قابل توجهی نیز دست یافته‌اند.

۱-۲. مفاهیم پایه یادگیری تقویتی عمیق

فرآیند تصمیم‌گیری مارکوف^{۱۵} (**MDP**) معمولاً به عنوان چارچوب نظری برای یادگیری تقویتی استفاده می‌شود. در مسائل یادگیری تقویتی بدون مدل، **MDP** با یک ۵ تایی (S, A, P, R, γ) توصیف می‌شود، که در آن **S** نشان‌دهنده فضای حالت، **A** نشان‌دهنده فضای رفتار، **P** نشان‌دهنده احتمال انتقال از حالت **S** به حالت بعدی **S'** است. **R** نشان‌دهنده تابع پاداش و γ نشان‌دهنده ضریب تخفیف است. فرآیند اجرای یادگیری تقویتی به این شرح است. در زمان t ، پس از انجام یک عمل توسط عامل در محیط، حالت از S_t به S_{t+1} تغییر می‌کند و عامل در لحظه بعد مقدار پاداش R_{t+1} را دریافت می‌کند. ساختار کلی یادگیری تقویتی در شکل نشان داده شده است.



شکل ۲. ساختار کلی یادگیری تقویتی.

یادگیری تقویتی روشی برای عامل با هدف بهینه‌سازی رفتار خود در یک محیط ناشناخته است. [۱۸] با توجه به هدف بهینه‌سازی، الگوریتم‌های یادگیری تقویتی را می‌توان به

آزمون و خطا می‌آموزد. یو و همکارانش [۱۶] با استفاده از چارچوب تصمیم‌گیری ناوبری مبتنی بر **DRL** در یک چارچوب مبارزه الکترونیکی شناختی، یک شبیه‌ساز ردیابی هدف ایجاد کرده‌اند. ردیابی اهداف توسط پهپادها در محیط‌های مختلف محقق می‌شود. به منظور مقابله با فضای بزرگ حالت در طول نبرد هوایی، لی و همکارانش الگوریتمی را برای بهبود عملکرد تصمیم‌تصمیمات کنترل پهپاد پیشنهاد کرده‌اند. نتایج شبیه‌سازی نشان می‌دهد که الگوریتم بهبودیافته می‌تواند تصمیم‌گیری در مانور نبردهای هوایی و اشتراک استراتژی را در محیط‌های عملیاتی متعدد تحقق بخشد. اگر شبکه عصبی به خوبی آموزش داده شود، الگوریتم‌های مبتنی بر شبکه عصبی می‌توانند تصمیمات بلندرنج با کیفیت بالا را ارائه دهند.

۱-۲. سازماندهی مقاله

ادامه مقاله در پنج بخش تنظیم گردیده است. در بخش اول مروری بر مطالعات انجام شده مربوط به **ACMD** مبتنی بر **DRL** شده است. در این بخش ابتدا توصیف مختصری از **DRL** آمده است، سپس به کاربرد آن در تصمیم‌گیری‌های جنگ هوایی هوشمند پرداخته شده است. در بخش دوم، پشتیبانی تصمیم در مانور نبردهای هوایی مبتنی بر یادگیری تقویتی با سه روش مختلف تشریح شده است. در بخش چهارم طراحی تابع پاداش مورد ارزیابی قرار گرفته است که هسته **ACMD** مبتنی بر **DRL** است. در این بخش همچنین یک نمونه پیاده‌سازی **ACMD** مبتنی بر **DRL** بر اساس سناریوی نبرد سگ (dogfight) آورده شده است. بخش پنجم مؤلفه‌های کلیدی برای ساخت مدل تصمیم‌گیری مانور، از جمله فضای حالت، فضای عمل و طراحی تابع پاداش را شرح می‌دهد.

۲. بررسی یادگیری تقویتی عمیق

یادگیری تقویتی عمیق در سال‌های اخیر به یکی از موضوعات مورد توجه در هوش مصنوعی تبدیل شده است. لاکهید از روشی مبتنی بر معماری سلسله‌مراتبی و حداکثر یادگیری تقویت آنتروپی در آزمایش‌های **AlphaDogfight**

¹⁵ Markov Decision Process

حالت های مشاهده شده به اقدامات اجرایی مختلف بیان می‌شود. الگوریتم های مبتنی بر تابع ارزش از DNN برای تقریب تابع ارزش پاداش استفاده می‌کنند. تابع ارزش پاداش بلندمدت مورد انتظار از هر حالت را تخمین می‌زند. این اقدام، عامل را قادر می‌سازد تا ارزش هر حالت را ارزیابی کند و اقداماتی را انتخاب کند که پاداش تجمعی آینده را در یک فرآیند تصمیم‌گیری متوالی به حداکثر می‌رساند. اقدامات انتخاب شده با توجه به تابع ارزش پاداش، آینده نگر هستند و خط مشی به دست آمده از نظر تئوری در دراز مدت بهینه است. یک سیاست بهینه وجود دارد که پاداش انباشته بلند مدت را به حداکثر می‌رساند:

$$\pi^*(a|s) = \arg \max_{a \in A} Q_{\pi}(s, a) \quad (1)$$

بهترین خط مشی از طریق تکرارها پیدا می‌شود.

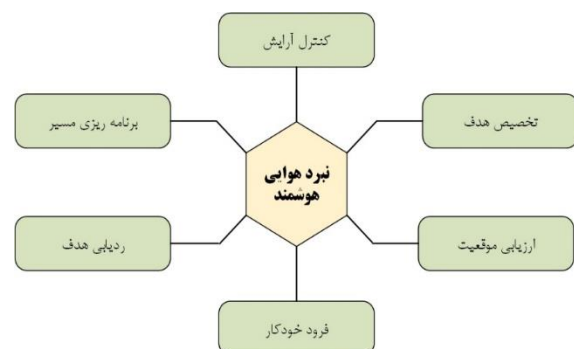
شکل ۳. کاربردهای DRL در نبرد هوایی

همراه با توسعه تجهیزات نظامی، روش های جدید استفاده از پهپاد مانند تبدیل جنگنده های از رده خارج شده به حالت بدون سرنشین، عملیات مشترک تجهیزات سرنشین دار/ بدون سرنشین و فناوری خوشه ای پدیدار شده است. پهپادها می‌توانند مانورهای چابک تری نسبت به پرنده های سرنشین دار انجام دهند، زیرا محدودیت های فیزیکی خلبانان را ندارند. همچنین در شرایط پیچیده نبرد هوایی، پهپادها با قابلیت تصمیم‌گیری مانور با کیفیت بالا می‌توانند در نبردهای هوایی برتری محسوسی داشته باشند. در حال حاضر، DRL به طور فزاینده ای در زمینه پشتیبانی تصمیم‌گیری در مانورهای هوایی استفاده می‌شود. اگرچه آموزش شبکه های عصبی به تعداد زیادی منابع محاسباتی نیاز دارد، مکانیسم آزمون و خطای منحصر به فرد یادگیری تقویتی و همچنین قابلیت تطبیق داده های قدرتمند شبکه های عصبی عمیق به DRL اجازه می‌دهد تا از استراتژی های مانور با کیفیت بالا بهره برداری کند.

الگوریتم‌های مبتنی بر تابع ارزش و مبتنی بر گرادیان خط مشی تقسیم کرد. الگوریتم های مبتنی بر تابع ارزش، خط مشی بهینه را با به حداکثر رساندن تابع ارزش-حالت یا تابع ارزش-عمل پیدا می‌کنند. از جمله این الگوریتم ها، Q-Learning و SARSA هستند. الگوریتم های مبتنی بر خط مشی از یک تابع غیرخطی برای پارامترسازی خط مشی استفاده می‌کنند. حداکثر کردن پاداش تجمعی با تکرار خط مشی، مانند گرادیان خط مشی (PG)، REINFORCE و سایر الگوریتم ها به دست می‌آید. الگوریتم های مبتنی بر گرادیان خط مشی از شبکه های عصبی عمیق (DNN) برای تقریب خط مشی استفاده می‌کنند و از روش گرادیان خط مشی برای یافتن خط مشی بهینه استفاده می‌کنند. به طور گسترده ای در مسائل یادگیری تقویتی در فضای پیوسته استفاده می‌شود. هدف از یادگیری تقویتی یادگیری یک خط مشی برای به حداکثر رساندن پاداش است. خط مشی، رفتار عامل را در یادگیری تقویتی تعیین می‌کند، که به صورت نگاشت

۲-۲. کاربرد DRL در تصمیم‌گیری نبرد هوایی

پهپادها در نبردهای هوایی با فضای حالت عظیمی در فضای سه بعدی مواجه هستند که توانایی تعمیم شبکه های عصبی با ساختارهای ثابت را به چالش می‌کشد. DRL بازخورد پاداش فوری و بازخورد پاداش آتی را در نظر می‌گیرد که پهپاد را قادر می‌سازد تا سیاست بهینه را در دراز مدت یاد بگیرد. کاربردهای DRL در نبرد هوایی در شکل (۳) نشان داده شده است. این کاربردها عبارت اند از: برنامه ریزی مسیر [۱۹] و [۲۰]، ردیابی هدف [۲۱] و [۲۲]، فرود خودکار [۲۳]، کنترل آرایش [۲۴] و [۲۵]، تخصیص هدف [۲۶] و [۲۷] و ارزیابی موقعیت [۲۸].



۳. طراحی تابع پاداش در پشتیبانی تصمیم مانور نبردهای هوایی مبتنی بر یادگیری تقویتی عمیق

یادگیری تقویتی دارای چالش هایی در آموزش الگوریتم و کنترل دقت می باشد. در پشتیبانی تصمیم نبردهای هوایی که استقلال عامل بالاترین اولویت را دارد، طراحی تابع پاداش از اهمیت ویژه ای برخوردار است. [۲۹] پاداش پراکنده، مشکل رایج DRL در حل وظایف عملی است. طراحی تابع پاداش نبرد هوایی ساده، آسان است اما آموزش الگوریتم ضعیف و عملکرد پائینی خواهد داشت. پهباد قادر به یادگیری استراتژی های تصمیم گیری مانور نیست. به عنوان مثال، عامل تنها بر اساس پیروزی یا شکست در پایان نبرد هوایی پاداش دریافت می کند. به دلیل سیگنال پاداش پراکنده، دستیابی به نمونه های آموزشی موثر برای عامل دشوار است و در نتیجه زمان زیادی صرف آموزش تعداد زیادی نمونه بی معنی می شود. نبردهای هوایی دارای فضای بزرگی هستند که طراحی تابع پاداش متمرکز را دشوار می کند. بنابراین نحوه طراحی پاداش در حین نبرد هوایی و راهنمایی عامل برای تصمیم گیری در جهت پیروزی در نبرد هوایی، تمرکز و دشواری عملکرد پاداش است. [۳۰] حل مشکل پاداش پراکنده می تواند به عامل کمک کند تا تعامل کمتری با محیط داشته باشد و از نمونه ها بهتر استفاده کند. پژوهش هایی در مورد حل مشکل پاداش پراکنده ACMD مبتنی بر RL صورت گرفته است که شامل موارد زیر است.

۳-۱. طراحی تابع پاداش

در مطالعات [۳۱]، [۳۲] و [۳۳] یک مکانیسم کلیدی شکل دهی پاداش در خلال نبرد هوایی پیشنهاد شده است. در این مطالعات، پاداش های پراکنده اما عینی را فراتر از سیگنال برد یا باخت اپیزودیک برای تسریع فرآیند یادگیری تقویتی ارائه شده است. در کنار آنها، لیو و همکارانش [۳۱] از یک تابع پیاده سازی توزیع شده و روش آموزشی متمرکز برای بهبود تصمیم گیری پهباد در یک شبیه سازی واقعی نبرد هوایی فراتر از برد بصری استفاده کرده اند. آنها زاویه انحراف، سرعت، ارتفاع و فاصله سازندهای پهباد و مناطق حمله موشکی را در نظر گرفته اند. فن و همکارانش نیز یک مکانیسم پاداش دو لایه را با

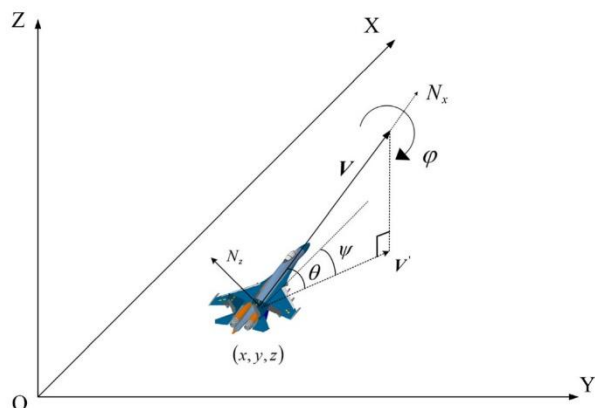
ترکیب پاداش های داخلی و پراکنده برای بهبود سرعت همگرایی مدل با در نظر گرفتن وضعیت نبرد هوایی طراحی کرده اند.

۳-۲. مکانیسم تکرار تجربه اولویت

مکانیسم تکرار تجربه اولویت به این معنی است که نمونه هایی با خطای TD بزرگتر در مجموعه تجربه ترجیحاً نمونه برداری می شوند. این به عامل اجازه می دهد تا به طور موثرتری از تجربیات مهم بیاموزد و راه حل موثری برای رسیدگی به ناکارآمدی های نمونه است. [۳۴] در کار [۲۲]، الگوریتم DDPG با ساخت مخزن تجربه موفقیت و مخزن تجربه شکست برای بهینه سازی استراتژی نمونه گیری موفق بوده است. آنها از مجموعه تجربه شکست و مجموعه تجربه موفقیت به نسبت معینی نمونه می گیرند. نتایج شبیه سازی نشان می دهد که الگوریتم بهبود یافته از نظر همگرایی و پایداری بهتر عمل می کند. ژانگ و همکارانش [۳۵] یک روش تصمیم گیری مانور نبرد هوایی را بر اساس تخمین پاداش نهایی پیشنهاد کرده اند. آنها از تخمین پاداش نهایی و الگوریتم PPO برای جایگزینی تابع تخمین برتری اصلی با تخمین پاداش نهایی استفاده می کنند. این روش، عملکرد آموزشی یادگیری تقویتی را بهبود می بخشد.

۳-۳. وظیفه کمکی

عامل می تواند با کارهای ساده و مرتبط شروع کند و سپس برای یادگیری استراتژی های پیچیده تر، سختی کارها را افزایش دهد. در کار [۳۶]، به منظور حل چالش بار محاسباتی عظیم مورد نیاز برای نبرد هوایی، ابتدا "مقابله اساسی" با مسیرهای ثابت جنگنده های دشمن انجام می شود. سپس اطلاعات دشمن ارتقاء می یابد و عامل در سطح "مقابله اساسی" بازآموزی می شود. این مکانیزم در نهایت پهباد را قادر می سازد تا تصمیمات هوشمندانه ای در نبرد هوایی بگیرد و استراتژی های موثری برای شکست دادن حریف به دست آورد. نبرد هوایی یک مسئله چالش برانگیز با فضای حالت بزرگ است. برای کنار آمدن با این چالش، سه روش بصورت خلاصه بیان می شود که می توانند نرخ یادگیری نمونه را افزایش دهند و از نیاز به طراحی تابع پاداش پیچیده اجتناب کنند. همچنین این سه روش را می توان ترکیب کرد تا پهبادها بتوانند استراتژی های هوشمند را سریعتر یاد بگیرند.



شکل ۴. شرح وضعیت پهپاد در فضای سه بعدی

زاویه گردش زاویه اطراف بردار سرعت است (بدنه به سمت راست مثبت است). θ ، ψ و ϕ در مجموع به عنوان زوایای اوایلر^{۱۶} پهپاد نامیده می‌شوند که منعکس کننده نگرش پهپاد نسبت به سیستم مختصات اینرسی زمین هستند. برای ساده‌تر کردن فرآیند تحلیل، معادلات حرکت فضایی سه‌بعدی پهپاد در سیستم مختصات اینرسی زمین به صورت زیر تعریف می‌شوند:

$$\begin{cases} \dot{v} = g(N_x - \sin \theta) \\ \dot{\psi} = \frac{g N_z \sin \phi}{v \cos \theta} \\ \dot{\theta} = \frac{g}{v} (N_z \cos \phi - \cos \theta) \\ \dot{x} = v \cos \theta \cos \psi \\ \dot{y} = v \cos \theta \sin \psi \\ \dot{z} = v \sin \theta \end{cases} \quad (2)$$

هر عبارت در مجموعه معادلات به ترتیب بیانگر یک معادله دیفرانسیل مرتبه اول برای مقادیر سرعت پرواز، زاویه انحراف، زاویه گام و مقادیر مختصات فضایی سه بعدی پهپاد است. این عبارات روش محاسبه و به روز رسانی وضعیت پهپاد را در حین نبرد تعیین می‌کنند. g نشان دهنده شتاب گرانش است. N_x اضافه بار مماسی در جهت سرعت است که نیروی رانش و کاهش سرعت پهپاد را تعیین می‌کند. N_x کنترل شتاب، کاهش سرعت، یا سرعت یکنواخت پهپاد را امکان پذیر می‌کند. N_z اضافه بار معمولی پهپاد است که نشان دهنده اضافه بار رو به بالا عمود بر هواپیمای بدنه آن است که قادر است بالابر مورد نیاز پهپاد را

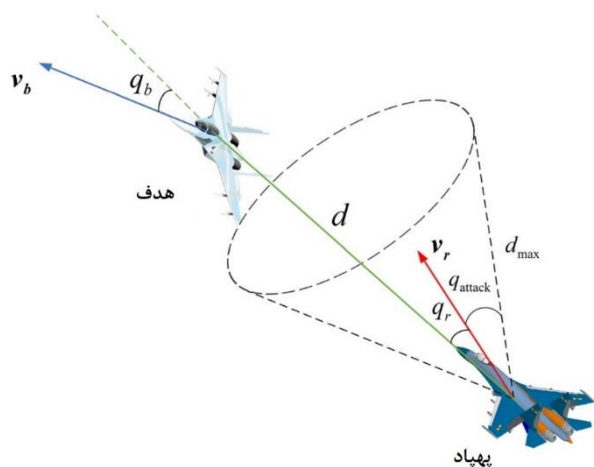
۴. مدل پشتیبان تصمیم گیری مانور نبردهای هوایی پیشنهادی

Dogfight یک نوع از نبرد هوایی است که در آن مسلسل سلاح اصلی است و هر دو جنگنده سعی می‌کنند برای هدف گیری و شلیک پشت سر یکدیگر قرار بگیرند. علیرغم پیشرفتهایی که در موشک‌ها و رادارها در انجام عملیات فراتر از برد بصری صورت گرفته است، **Dogfight** در نبرد هوایی اجتناب‌ناپذیر است [۳۷]. نبرد هوایی فرآیندی پیچیده و پویا است. با بهبود عملکرد پهپاد، تاکتیک‌های نبرد هوایی نیز تکامل یافته‌اند. هسته اصلی تصمیم گیری مانور هوشمند، تولید و ارزیابی توالی مانورهای رزمی متعدد بر اساس تجربه رزمی و موقعیت بلادرنگ است [۳۸]. پهپاد باید مانورهای خود را انتخاب کند و بر اساس تغییرات موقعیتی در طول مأموریت‌های رزمی، توالی‌های رزمی مؤثری را تشکیل دهد.

۴-۱. معادلات حرکتی پهپاد

پهپاد دارای شش درجه آزادی حرکت در طول پرواز است. می‌توان آن را به یک نقطه جرم تقلیل داد که در سه بعد با مقادیر فیزیکی مانند سرعت پرواز، مختصات سه بعدی و زوایای نگرش توصیف می‌شود. وضعیت لحظه‌ای پهپاد را می‌توان مطابق شکل (۴) تجسم کرد. بردار واحد V نشان دهنده سرعت پهپاد است. x و y و z به ترتیب مختصات پهپاد را در فضای سه بعدی در مقایسه با زمین به عنوان سیستم مرجع نشان می‌دهند. θ ، ψ و ϕ به ترتیب زاویه گام، زاویه انحراف و زاویه چرخش پهپاد هستند. زاویه گام، زاویه بین جهت سرعت پرواز V پهپاد و صفحه افقی است. زاویه انحراف نشان دهنده زاویه بین برآمدگی V بردار سرعت و جهت محور Y است.

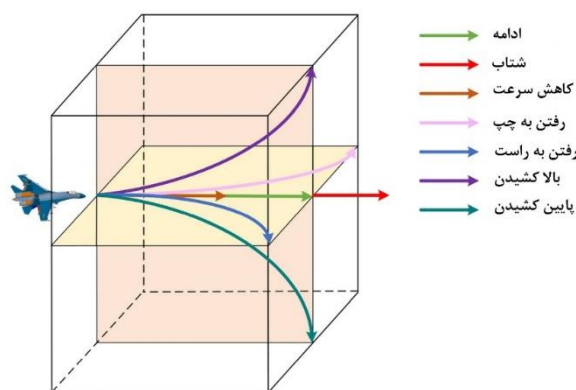
حمله سلاح d_{max} باشد، حداکثر زاویه پرتاب خارج از محور q است، پس برد حمله پهپاد یک منطقه مخروطی Ω است.



شکل ۵. محدوده حمله پهپاد

۴-۳. فضای عمل

مانورهای تاکتیکی پهپادهای با بال ثابت را می‌توان به طور کلی به دو دسته تقسیم کرد. دسته اول، مانورهای پیچیده تر مانند سالتو، کبرا، رول و غیره است. دسته دیگر، مانورهای اساسی است که مانورهای پیچیده را تشکیل می‌دهند، یعنی مجموعه مانورهای پایه مبارزه (BFM^{17}) پیشنهاد شده توسط اداره ملی هوانوردی و فضایی (ناسا)، که شامل هفت الگوی اساسی است که در شکل (۶) نشان داده شده است. چنین انتخابی به طور گسترده در مطالعات [۳۲] و [۴۰] استفاده شده است.



شکل ۶. مانورهای پایه پهپاد

تأمین کند. N_z می‌تواند شیرجه را کنترل کند و برای تغییر ارتفاع عمودی پهپاد بالا بکشد. این مدل به طور گسترده در مطالعات مختلفی استفاده شده است. [۳۹]

۴-۲. فضای حالت

با ترکیب حالات پهپاد قرمز و آبی می‌توان وضعیت نبرد هوایی را ارزیابی کرد. با توجه به تنظیمات در نظر گرفته شده در کار [۳۱]، فضای حالت ۱۲ بعدی زیر را انتخاب می‌کنیم:

$$S = \{v_r, x_r, y_r, z_r, \theta_r, \psi_r, v_b, x_b, y_b, z_b, \theta_b, \psi_b\} \quad (۳)$$

که در آن عناصر با زیرنویس r و b نشان دهنده وضعیت سمت قرمز و آبی هستند. در اجرای نبرد هوایی هوشمند، از وضعیت لحظه ای رنگ قرمز و آبی به عنوان منبع سیگنال پاداش استفاده خواهد شد. دو بردار واحد (v_b و v_r) برای نشان دادن جهت پرواز دو طرف استفاده می‌شود. d_r و d_b نشان دهنده بردار موقعیت نسبی بین دو طرف هستند. این متغیرها را می‌توان به صورت زیر بیان کرد:

$$\begin{cases} v_r = [\cos \psi_r, \cos \theta_r, \sin \psi_r, \cos \theta_r, \sin \theta_r] \\ v_b = [\cos \psi_b, \cos \theta_b, \sin \psi_b, \cos \theta_b, \sin \theta_b] \\ d_r = -d_b = [x_b - x_r, y_b - y_r, z_b - z_r] \end{cases} \quad (۴)$$

رابطه بین اطلاعات موقعیتی نسبی و سیستم مختصات زمین در معادله (۵) نشان داده شده است.

$$\begin{cases} d = |d_r| \\ q_r = \cos^{-1} \frac{v_r \cdot d_r}{d} \\ q_b = \cos^{-1} \frac{v_b \cdot d_b}{d} \end{cases} \quad (۵)$$

که در آن q_b و q_r به ترتیب زاویه بین بردار سرعت و مرکز را نشان می‌دهند. d نشان دهنده فاصله نسبی بین قرمز و آبی است. هدف از نبرد هوایی کوتاه برد، مانور دادن پهپاد به موقعیتی است که بتواند به هدف حمله کند و در عین حال از محدوده حمله هدف نیز جلوگیری کند. محدوده حمله پهپاد در این مقاله در شکل (۵) نشان داده شده است. با فرض اینکه حداکثر فاصله

¹⁷ Basic Fight Maneuvering

برای هفت مانور فوق، اضافه بار مماسی، اضافه بار معمولی و زاویه چرخش پهپاد را برای کنترل مانور انتخاب می‌کنیم. سه گانه $[N_x, N_z, \phi]$ برای نمایش اقدامات انجام شده توسط پهپاد در هر لحظه در شبیه‌سازی استفاده می‌شود. با توجه به محدودیت‌های فیزیکی که پهپاد با آن‌ها روبرو است، فرامین مربوط به هفت مانور مختلف در جدول ۱ نشان داده شده است.

جدول ۱. دستورات مانورهای اساسی مبارزه

فرامین			مانور انتخابی
ϕ	N_z	N_x	
۰	۱	۰	پرواز عادی
۰	۱	۲	افزایش شتاب
۰	۱	-۲	کاهش شتاب
$-\pi/3$	۵	۰	چرخش به چپ
$\pi/3$	۵	۰	چرخش به راست
۰	۵	۰	بالا کشیدن
۰	-۵	۰	پائین کشیدن

۴-۴. توابع پاداش

نبرد هوایی کوتاه برد را به عنوان محیط شبیه‌سازی در نظر می‌گیریم و فرض می‌کنیم که هر دو طرف قرمز و آبی از مسلسل‌های هوایی به عنوان سلاح استفاده می‌کنند. عوامل موثر بر وضعیت در طول نبرد هوایی عمدتاً شامل موقعیت فضایی، ارتفاع پرواز، سرعت پهپاد و سایر پارامترهای پرواز است. اطلاعات موقعیتی مانند زاویه، فاصله، ارتفاع و سرعت پهپادهای دو طرف به عنوان شاخصی برای ارزیابی برتری نبرد هوایی استفاده می‌شود. [۳۳] در اینجا دو نوع تابع پاداش را تعریف می‌شود، توابع پاداش فرآیند و نتیجه. تابع پاداش فرآیند هر دو طرف نبرد و موقعیت‌های مربوط به آن‌ها را در نظر می‌گیرد، در حالی که تابع پاداش نتیجه به نتیجه هر قسمت بستگی دارد. دو تابع پاداش فوق با هم ترکیب می‌شوند تا مشکل پاداش پراکنده را حل کنند.

۴-۴-۱. تابع پاداش فاصله زاویه

معادله حرکتی پهپاد در فضای سه بعدی و به روزرسانی موقعیت آن در طی شبیه‌سازی با استفاده از رابطه زیر تعریف می‌شود.

$$S_{t+1} = f(S_t, a_r, a_b) \quad (6)$$

در این معادله، a_r و a_b نشان دهنده مانورهای انتخاب شده توسط پهپادهای قرمز و آبی در تصمیم‌گیری تک مرحله‌ای هستند. f نشان دهنده تابع به روزرسانی حالت است که بر اساس معادله دیفرانسیل حرکتی ساخته شده است. دیفرانسیل مرتبه اول سرعت پهپاد، مقادیر مختصات فضایی سه بعدی، زوایای گام و انحراف را می‌توان با ترکیب حل عددی معادلات دیفرانسیل اوایلر به دست آورد. در حالت S_t ، با توجه به مانورهای هر دو طرف همراه با واحدهای زمان شبیه‌سازی d_t می‌توان حالت جدید S_{t+1} را محاسبه کرد.

می‌یابد و شرایط برای حمله نامطلوب می‌شود. با توجه به تاثیر زاویه جداسازی و زاویه انحراف بر تسلط پهپاد قرمز در نبرد، تابع پاداش فاصله زاویه به شرح زیر بیان می‌شود:

$$r_a = \begin{cases} (q_b - q_r) \exp\left(-\left(\frac{d_0 - d_{max}}{d_{max}}\right)^2\right), & q_r \leq q_b, d_0 > d_{max} \text{ and } q_b \in \alpha_m \\ (q_b - q_r), & q_r \leq q_b, d_0 > d_{max} \text{ and } q_b \notin \alpha_m \\ 0, & \text{else} \end{cases} \quad (7)$$

ارتفاع همچنین باعث افزایش انرژی پتانسیل گرانشی می‌شود که می‌تواند برای دستیابی به سرعت پرواز بالاتر به انرژی جنبشی تبدیل شود. تابع پاداش ارتفاع به صورت زیر تعریف می‌شود:

$$r_{h1} = \begin{cases} \exp\left(-\left(\frac{z_r - h_{best}}{h_{best}}\right)\right), & z_r \leq h_{best} \\ \exp\left(\frac{z_r - h_{best}}{h_{best}}\right), & z_r > h_{best} \end{cases} \quad (9)$$

$$r_{h2} = \begin{cases} \exp\left(\frac{\Delta h - \Delta h_{best}}{\Delta h_{best}}\right), & \Delta h \leq \Delta h_{best} \\ \exp\left(\frac{\Delta h_{best} - \Delta h}{\Delta h_{best}}\right), & \Delta h > \Delta h_{best} \end{cases} \quad (10)$$

$$r_h = r_{h1} + r_{h2} \quad (11)$$

در این تابع پاداش، r_h بیانگر تابع پاداش ارتفاع، r_{h1} بیانگر تابع پاداش ارتفاع مطلق و r_{h2} نشان دهنده تابع پاداش ارتفاع نسبی می‌باشد. $\Delta h = z_r - z_b$ نشان دهنده اختلاف ارتفاع بین قرمز و آبی است. Δh_{best} ، h_{best} ارتفاع نبرد هوایی بهینه و ارتفاع نسبی بهینه هستند که کمیت های مشخصی هستند.

۴-۴-۴. تابع پاداش نتیجه

تابع پاداش نتیجه نشان دهنده پاداشی است که عامل زمانی که پهپاد به وضعیت خاصی می‌رسد دریافت می‌کند. این شامل مواردی می‌شود که پهپاد در یکی از حالات برنده یا بازنده، از محدوده خارج می‌شود. تابع پاداش نتیجه به صورت زیر تعریف می‌شود.

تابع پاداش فاصله زاویه اثرات فاصله، زاویه انحراف q_r و زاویه انحراف q_b را بر ارزیابی موقعیت در نظر می‌گیرد. تا حد زیادی نتیجه نبرد هوایی را تعیین می‌کند. هر دو پارامتر q_r و q_b مقداری در بازه $[0, \pi]$ می‌گیرند. زاویه جداسازی کوچک به این معنی است که دشمن در مقابل جهت پرواز پهپاد قرمز قرار دارد و به طرف قرمز امتیاز حمله و تهدیدی برای پهپاد آبی می‌دهد. این وضعیت مساعدترین موقعیت برای حمله تیم قرمز است. با افزایش تدریجی q_r از صفر به π ، زاویه حمله قرمز افزایش

در این معادله، r_a تابع پاداش زاویه است. d_0 و d_{max} نشان دهنده فاصله بین قرمز و آبی و حداکثر فاصله حمله سلاح است. α_m نشان دهنده زاویه ای است که در آن حمله مجاز است.

۴-۴-۲. تابع پاداش سرعت

تابع پاداش سرعت توانایی پهپاد را در استفاده از مزیت سرعت خود برای جلوگیری از تهدید دشمن و رسیدن سریع به منطقه حمله خود ارزیابی می‌کند. تابع پاداش سرعت به صورت زیر تعریف می‌شود:

$$r_v = \begin{cases} \exp\left(\frac{v_r - v_{best}}{v_{best}}\right), & v_r \leq v_{best} \\ \exp\left(\frac{v_{best} - v_r}{v_{best}}\right), & v_r > v_{best} \end{cases} \quad (8)$$

v_{best} سرعت نبرد هوایی بهینه را نشان می‌دهد که یک مقدار مشخص است.

۴-۴-۳. تابع پاداش ارتفاع

پرواز سه بعدی پهپاد در مقایسه با شبیه سازی پرواز در یک هواپیمای دو بعدی، یک بعد ارتفاع اضافی را شامل می‌شود. در نبرد هوایی در صورتی که پهپاد در ارتفاع معینی بالاتر از هدف قرار گرفته باشد، دارای مزیت انرژی پتانسیل گرانشی است. برتری در ارتفاع مناسب برای پهپاد برای تنظیم موقعیت مکانی مفید است تا دشمن را در محدوده حمله خود قرار دهد. برتری

$$\omega_1 + \omega_2 + \omega_3 = 1 \quad (\omega_1, \omega_2, \omega_3 \geq 0)$$

تابع پاداش یکپارچه r برای فرآیند نبرد هوایی به صورت زیر نشان داده می‌شود.

$$r = r_{process} + r_{result} \quad (14)$$

۵-۴. چارچوب تصمیم‌گیری در مانور نبرد هوایی

مبتنی بر DQN

DQN، یا همان حالت پیشرفته DRL، می‌تواند استراتژی‌های کنترل را مستقیماً از داده‌های خام با ابعاد بالا در نبرد هوایی بیاموزد. ایده کلیدی الگوریتم DQN استفاده از DNN برای تقریب تابع مقدار است. همچنین از خطای TD برای تنظیم پیوسته پارامترهای شبکه عصبی استفاده می‌شود، به طوری که مقدار خروجی شبکه به طور پیوسته به مقدار واقعی $Q(s, a) \approx Q(s, a|\theta)$ تقریب زده می‌شود. مدل محیط نبرد هوایی، ارزش پاداش پهباد را بر اساس وضعیت فعلی هر دو طرف محاسبه کرده و آن را به عامل ارسال می‌کند. سپس عامل اقداماتی را برای تغییر حالت در مدل محیط نبرد هوایی انجام می‌دهد. عامل دائماً با محیط نبرد هوایی در تعامل است تا وضعیت لحظه بعدی، عمل و ارزش پاداش را به دست آورد و آنها را در مخزن تجربه ذخیره کند. با انباشت تجربه، مدل DQN استراتژی مانور پهباد را بر اساس مخزن تجربه به روز می‌کند تا مقادیر عمل را بهینه کند. شکل (۷) مدل تصمیم‌گیری مبتنی بر DQN را برای نبرد هوایی کوتاه برد پهباد نشان می‌دهد.

$$r_{result} = \begin{cases} r_1, & q_r \in \alpha_m, d < d_{max} \text{ and } q_b \in q_0 \\ r_2, & q_b \in \alpha_m \text{ and } d < d_{max} \\ r_3, & s \neq s_r \\ r_4, & step = step_{max} \\ r_5, & else \end{cases} \quad (12)$$

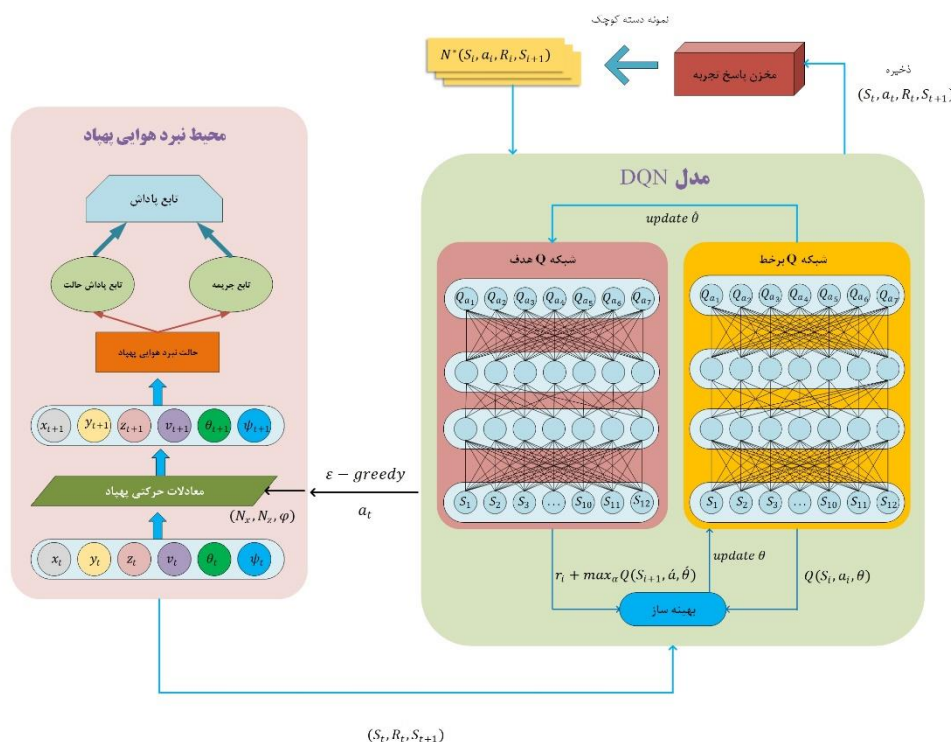
وقتی پهباد قرمز شرایط حمله را برآورده می‌کند و برنده می‌شود، عامل یک پاداش نتیجه مثبت r_1 دریافت می‌کند. برعکس، زمانی که پهباد آبی شرایط حمله را برآورده می‌کند و برنده می‌شود، عامل یک پاداش نتیجه منفی r_2 دریافت می‌کند. همچنین S_r محدوده مرزی حرکت پهباد و $step_{max}$ حداکثر مراحل تصمیم‌گیری است. برای بهبود راندمان آموزشی شبکه، زمانی که پهباد از محدوده مرزی فراتر می‌رود یا تکرارها به حداکثر مراحل تصمیم‌گیری می‌رسد، عامل پاداش‌های منفی r_3 یا r_4 را دریافت می‌کند.

۵-۴-۵. تابع پاداش تجمع شده

بر اساس تجزیه و تحلیل فوق از وضعیت نبرد هوایی، پاداش فاصله زاویه بالا، پاداش ارتفاع و پاداش سرعت به ترتیب به صورت r_a ، r_v و r_h نشان داده می‌شود. تابع پاداش تجمع شده، ترکیبی از این سه مقدار است.

$$r_{process} = \omega_1 r_a + \omega_2 r_v + \omega_3 r_h \quad (13)$$

ضرایب این تابع با شرایط زیر در نظر گرفته می‌شوند.



شکل ۷. مدل تصمیم گیری مبتنی بر DQN برای نبردهای هوایی کوتاه برد

استراتژی $\epsilon - greedy$ برای ایجاد تعادل بین اکتشاف و بهره برداری استفاده شده است. معادله (۱۵) این موضوع را اثبات می کند:

$$\epsilon = \epsilon_{end} + (\epsilon_{start} - \epsilon_{end}) \exp\left(-\frac{frame_index}{\epsilon_{decay}}\right) \quad (15)$$

در این معادله، ϵ_{start} نشان دهنده مقدار اولیه اکتشاف، ϵ_{end} نشان دهنده مقدار نهایی اکتشاف، $frame_index$ یکی عدد تجمعی و ϵ_{decay} بیانگر نرخ فروپاشی است. با افزایش تعداد تکرارهای آموزش، عامل کمتر و کمتر محیط را بررسی می کند و بیشتر به تجربه ذخیره شده قبلی تکیه می کند. الگوریتم DQN با تناسب تابع Q-value در هر حالت، تفاوت بین پیش بینی های شبکه و مقدار واقعی را به حداقل می رساند. تابع زیان بصورت زیر تعریف می شود:

وضعیت فعلی نبرد هوایی S_t است، شبکه Q برخط بر اساس استراتژی $\epsilon - greedy$ عمل می کند. پهباد عملیات را در a_t انجام می دهد و هدف طبق استراتژی از پیش تعیین شده پرواز می کند. سپس حالت به S_{t+1} تغییر می کند و مقدار پاداش R_t برگردانده می شود. (S_t, a_t, R_t, S_{t+1}) نیز به عنوان یک نمونه تجربه در مخزن تجربه ذخیره می شود.

در طول فرآیند یادگیری، مدل DQN از استراتژی بازپخش تجربه اولویت دار برای نمونه برداری از دسته کوچکی از انتقال ها از مجموعه تجربه استفاده می کند و پارامترهای شبکه برخط θ را بر اساس اصل گرادین کاهشی به روزرسانی می کند. به صورت دوره ای، پارامترهای θ به شبکه هدف θ اختصاص داده می شود. یادگیری و به روز رسانی تا زمانی که خطای TD به صفر نزدیک شود ادامه می یابد و در نهایت یک استراتژی مانور هوایی کوتاه برد پیشنهاد می شود. در طول آموزش رزم هوایی، مشکل اکتشاف^{۱۸} و بهره برداری^{۱۹} باید حل شود. در این مقاله، از

¹⁹ Exploitation

¹⁸ Exploration

۵-۱. تنظیم پارامترها

پارامترهای مدل DQN به صورت زیر تنظیم می‌شوند: ابعاد فضای حالت و فضای عمل به ترتیب ورودی و خروجی شبکه عصبی یادگیری عمیق هستند. شبکه Q برخط و شبکه Q هدف دارای معماری یکسانی از چهار لایه کاملاً متصل شامل یک لایه ورودی، یک لایه خروجی و دو لایه پنهان هستند. هر لایه مخفی دارای ۳۲ واحد است و از تابع فعال سازی $ReLU$ استفاده می‌کند. اندازه دسته کوچک نمونه روی ۲۵۶ و اندازه مخزن تجربه روی 10^6 تنظیم شده است. پارامترهای سناریوی آزمایش و پارامترهای فرآیند آموزش در جداول (۲) و (۳) گزارش شده است.

$$Loss(\theta) = E_{(S_t, a_t, R_t, S_{t+1}) \sim D} \left[(R_t + \gamma \max_{a_{t+1} \in A} Q(S_{t+1}, a_{t+1}, \theta) - Q(S_t, a_t, \theta))^2 \right] \quad (16)$$

در این رابطه، $D(S_t, a_t, R_t, S_{t+1})$ نشان‌دهنده داده‌های نمونه‌برداری تصادفی از مجموعه تجربه D است و θ پارامتر مدل شبکه است.

۵. تجزیه و تحلیل نتایج

در این بخش نتایج آزمایش‌های نبرد هوایی هوشمند مبتنی بر DRL به همراه تجزیه و تحلیل نتایج به دست آمده ارائه می‌شود. همچنین تنظیمات پارامتر، فرآیند آموزش و یادگیری پهباد و نتایج نبرد با دشمن در این بخش تشریح خواهد شد.

جدول ۲. پارامترهای سناریوی آزمایش

پارامتر	مقدار	پارامتر	مقدار
d_{max}	5 km	z_{min}	1 km
q_{attack}	$\pi/6$	z_{max}	10 km
α_m	$[-\pi/6, \pi/6]$	v_{best}	300 m/s
q_0	$[2\pi/3, \pi] \cup [-2\pi/3, -\pi]$	h_{best}	5 km
v_{min}	100 m/s	Δh_{best}	1 km
v_{max}	400 m/s		

جدول ۳. پارامترهای فرآیند آموزش

پارامتر	مقدار	پارامتر	مقدار
r_1	8	ω_1	0.5
r_2	-8	ω_1	0.25
r_3	-5	ω_1	0.25
r_4	-1	dt	1 s

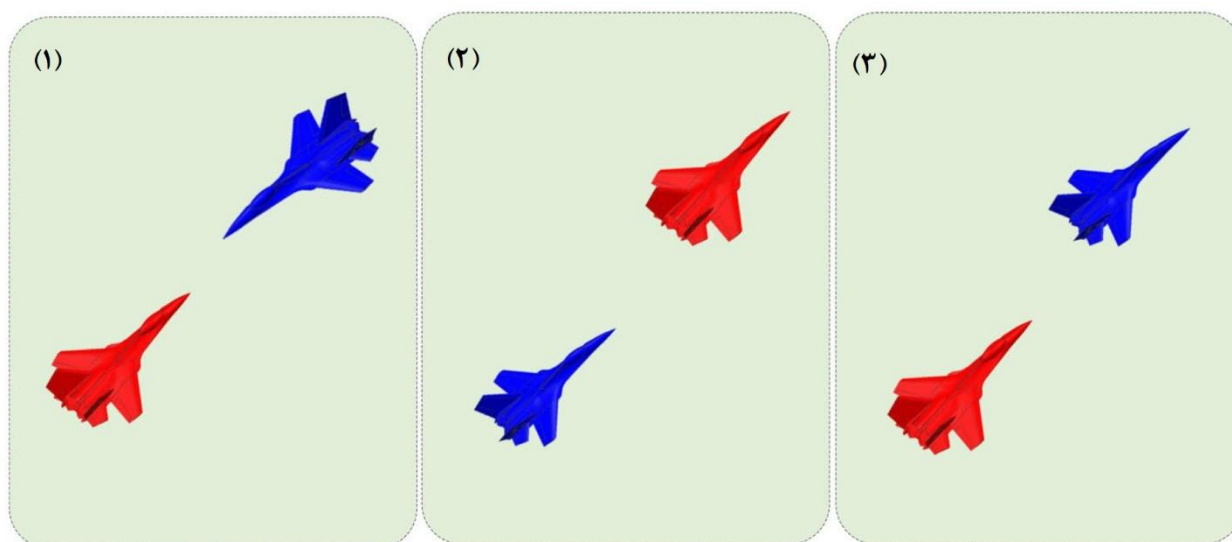
40	$step_{max}$	0	r_5
----	--------------	---	-------

۲-۵. آموزش

این بخش فرآیند آموزش و نتیجه نهایی پهپاد را در مراحل پیش‌آموزشی و تنظیم دقیق نشان می‌دهد. همچنین استراتژی مانور اصلی پهپاد در شرایط ساده و فرآیند یادگیری پهپاد در مرحله تنظیم دقیق هنگام مواجهه با حریف استراتژیک هوشمند را پوشش می‌دهد.

۱-۲-۵. مرحله قبل از آموزش

همانطور که در شکل (۸) نشان داده شده است، آزمایش نبرد هوایی در سه موقعیت اولیه مختلف انجام می‌شود: توازن، ضعف و برتری. هدف از مانور هوایی دستیابی به یک موقعیت برتر و رساندن هدف به محدوده سلاح خود و در عین حال دوری از برد سلاح آن است.



شکل ۸. سه حالت اولیه (۱) توازن، (۲) ضعف و (۳) برتری

ابتدا پهپاد را در سه حالت اولیه آموزش می‌دهیم. فرض بر این است که پهپاد آبی بدون هیچ تداخلی مستقیم و پایدار پرواز می‌کند. پهپاد قرمز از الگوریتم DQN برای یادگیری سیاست استفاده می‌کند. با آموزش ساده، پهپاد می‌تواند استراتژی حمله را در یک مأموریت ساده بیاموزد. جدول ۴ موقعیت اولیه هر دو پهپاد را در یک موقعیت ساده نشان می‌دهد.

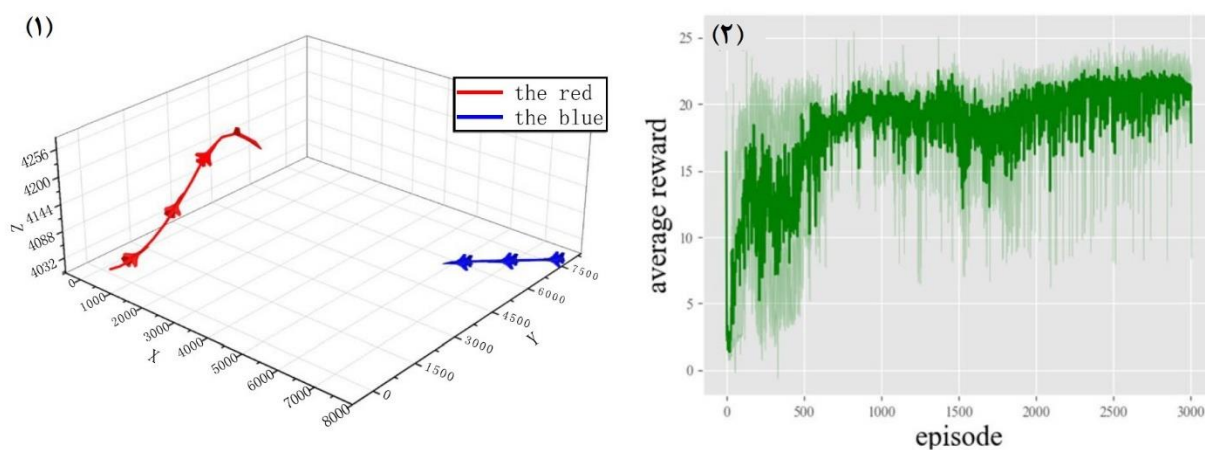
جدول ۴. وضعیت اولیه پهپادها در مرحله قبل از آموزش

مختصات شروع (متر)	سرعت (متر بر ثانیه)	زاویه استقرار (درجه)	زاویه انحراف (درجه)	
قرمز	(۰, ۰, ۴۰۰۰)	۲۰۰	۰.۰	۴۵
آبی	(۸۰۰۰, ۸۰۰۰, ۴۰۰۰)	۲۲۰	۰.۰	-۱۳۵

ضعف	قرمز	(۰,۰,۴۰۰۰)	۲۰۰	۰.۰	-۱۳۵
	آبی	(۴۰۰۰,۴۰۰۰,۴۰۰۰)	۲۲۰	۰.۰	-۱۳۵
برتری	قرمز	(۰,۰,۴۰۰۰)	۲۰۰	۰.۰	۴۵
	آبی	(۴۰۰۰,۴۰۰۰,۴۰۰۰)	۲۲۰	۰.۰	۴۵

شیرجه ای انجام می دهد تا زاویه حمله را به دست آورد و در نبرد هوایی پیروز شود. شکل (۹) بخش (۲)، میانگین تغییرات پاداش را در حالت توازن نشان می دهد. میانگین پاداش با پاداش مربوط به دانه های تصادفی مختلف محاسبه می شود.

شکل (۹) بخش (۱)، مسیر مانور پهپاد را پس از آموزش در وضعیت توازن نشان می دهد. مشاهده می شود که پهپاد قادر است در اولین بار به سمت بالا بکشد تا ارتفاع خود را افزایش دهد و برتری ارتفاعی بدست آورد و از برد حمله دشمن دوری کند. سپس با کاهش فاصله بین دو طرف، پهپاد قرمز رنگ یک مانور

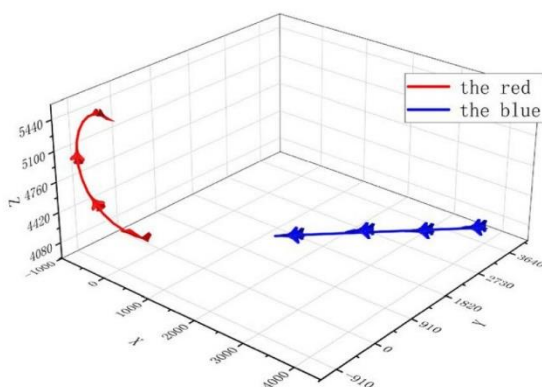


شکل ۹. نتایج تمرین وضعیت توازن (۱) مسیر مانور و (ب) میانگین پاداش

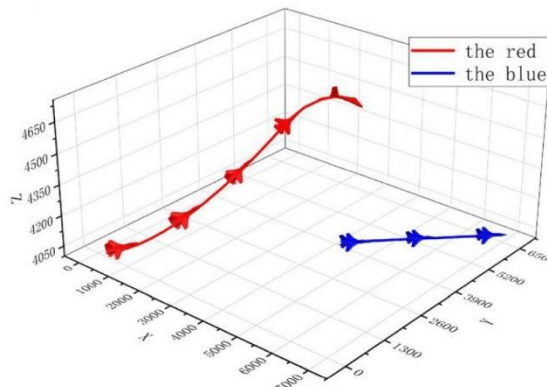
شکل (۱۰) قسمت های (۱) و (۲) مسیر آموزشی پهپاد را به ترتیب در موقعیت های نامطلوب و سودمند نشان می دهد. در مرحله پایانی آموزش ضعف ها، پهپاد برای به دست آوردن برتری ارتفاع بالا می آید و سپس به سرعت دماغه اش را می چرخاند و یک مانور تاکتیکی را انجام می دهد تا نقطه ضعف را معکوس کند و موقعیت حمله را به دست آورد. در مرحله پایانی آموزش برتری، پهپاد سرعت می گیرد و بالا می کشد تا موقعیت برتری را به موقعیت ضربه در برابر پهپاد دشمن تبدیل کرده و به پیروزی برسد.

مشاهده می شود که پهپاد در مرحله اولیه آموزش نتوانسته بر استراتژی حمله مناسب تسلط پیدا کند. به دلیل سرعت، ارتفاع بیش از مرز منجر به مشکلات و در نتیجه ارزش پاداش کمتر، تکرارها زود به پایان می رسد. با این حال، با افزایش تکرارهای آموزش، پهپاد به تدریج استراتژی مانور مناسب را یاد می گیرد. مشاهده می شود که پس از ۱۰۰۰ تکرار از آموزش، پهپاد نتوانسته است بر استراتژی مانور اولیه تسلط پیدا کند و ارزش پاداش همگرا شود.

(۱)



(۲)



شکل ۱۰. مسیر مانور در موقعیت های مختلف اولیه (۱) ضعف و (۲) برتری

قرمز مدل DQN آموزش دیده را برای یک مورد ساده استخراج می کند و با استفاده از استراتژی حریصانه ($\epsilon - greedy$) به کشف استراتژی مانور ادامه می دهد. همچنین برای اطمینان از تنوع نمونه های آموزشی، حالت اولیه هر دو طرف به طور تصادفی در محدوده ای مشخص، همانطور که در جدول (۵) نشان داده شده است، ایجاد شده است.

۲-۵. مرحله تنظیم دقیق

پس از آموزش برای سه سناریوی ساده، آموزش در برابر هدف هوشمند انجام می شود. پهباد قرمز از مدل DQN آموزش دیده با سناریوهای ساده برای مانور استفاده می کند، در حالی که پهباد آبی از یک استراتژی مانور قوی بر اساس اصول آماری برای انتخاب مانور استفاده می کند. در مرحله تنظیم دقیق، پهباد

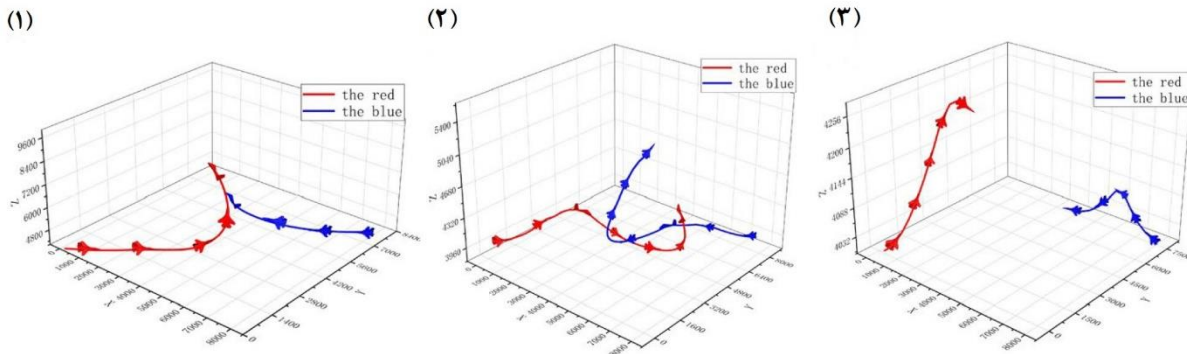
جدول ۵. وضعیت اولیه پهبادها در مرحله تنظیم دقیق

ψ (درجه)	θ (درجه)	V(m/s)	Z(m)	Y(m)	X(m)	
0	0	200	4000	[-300,300]	[-300,300]	قرمز
$[-\pi, -\pi/2]$	0	200	[3800,4200]	[7000,9000]	[7000,9000]	آبی

برای استفاده از برتری ارتفاع شتاب می گیرد. اما طرف قرمز سرعت و حرکت خود را به سرعت تنظیم می کند و در حین کاهش سرعت، یک مانور چرخش با زاویه بزرگ انجام می دهد. موقعیت تعقیب و گریز در برابر دشمن را شکل می دهد و پیروز می شود. شکل (۱۱) بخش (۳) مسیر رویارویی را در اواخر مرحله تنظیم دقیق نشان می دهد، زمانی که هر دو طرف بر استراتژی های مانور مسلط شده اند. هر دو طرف برای به دست آوردن برتری ارتفاع بالا می آیند، اما طرف قرمز سریعتر است. سپس طرف آبی دماغه خود را به سمت پایین می چرخاند تا از حملات جلوگیری کند و به دنبال فرصتی برای حمله باشد. با

شکل (۱۱) بخش (۱) مسیر رویارویی پهباد بین دو طرف را در ابتدای مرحله تنظیم دقیق نشان می دهد. پهباد قرمز به دلیل بی تجربگی، برای به دست آوردن برتری ارتفاع، ابتدا بالا می کشد. در همین حال سمت آبی نیز به سرعت بالا می آید و برای تنظیم دید به چپ می چرخد و به سمت قرمز حمله می کند. شکل (۱۱) بخش (۲) مسیر رویارویی را در وسط مرحله تنظیم دقیق نشان می دهد، زمانی که طرف قرمز بر برخی از استراتژی های مانور تسلط پیدا کرده است. دو طرف ابتدا سعی می کنند ارتفاع را بالا ببرند. طرف آبی پس از ناتوانی در یافتن زاویه حمله مناسب به چپ می پیچد و برای بالا بردن ارتفاع در تلاش

این حال، طرف قرمز شیرجه می زند و به سمت چپ می رود تا موقعیت هجومی به دست بیاورد و طرف آبی را شکست دهد.



شکل ۱۱. مسیرهای مانور آموخته شده در مرحله تنظیم دقیق، شامل (۱) دوره اولیه، (۲) دوره میانی و (۳) دوره پایانی

شکل ۱۲. نرخ پیروزی تحت دو استراتژی آموزش

۳-۵. تست

به منظور بررسی اثربخشی تنظیم دقیق، از مدل DQN برای انتخاب مانور در سمت قرمز پس از ۱۵۰۰۰ تکرار پیش تمرین و تنظیم دقیق استفاده شده است. شکل (۱۲) مقایسه نرخ پیروزی در ۱۰۰۰ قسمت از آموزش را پس از مرحله قبل از آموزش و تنظیم دقیق نشان می دهد. نتایج نشان می دهد که پس از بازآموزی در شبکه عصبی آموزش دیده بر روی سناریوهای ساده، نرخ برد عامل از ۱۳ به ۷۸ درصد افزایش می یابد. مدل DQN با یادگیری تقویتی روند آموزش می تواند به طور مداوم تصمیم گیری هوشمند پهاد را بهبود بخشد.

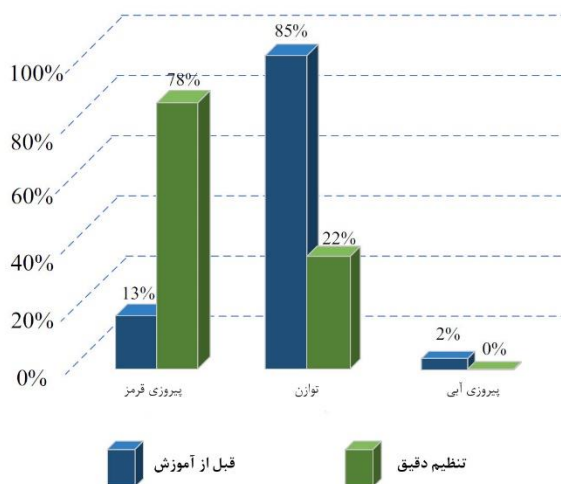
۴-۵. جهت گیری های تحقیقاتی آینده

فناوری هوشمند به ویژه هوش مصنوعی به سرعت در حال پیشرفت و تأثیرگذاری بر حوزه نظامی است. آموزش نبرد هوایی با چالش های زیادی مانند محدودیت های محل، آب و هوا، تجهیزات و پرسنل مواجه است. تکیه صرف به آموزش پرواز روزانه ناکارآمد است. علاوه بر این، با پیچیده تر شدن تجهیزات نظامی، خلبانان با بار عملیاتی بیشتری روبرو هستند. بنابراین نیاز شدیدی به پهاد هوشمند وجود دارد که بتواند سختی و پیچیدگی عملیات را کاهش دهد و به خلبانان کمک کند. جهت توسعه آینده نبرد هوایی هوشمند مبتنی بر DRL شامل سه جنبه اصلی است.

۱-۴-۵. عامل رزمی جامع تر

محیط میدان نبرد پیچیده تر و سخت تر می شود. عوامل مختلفی مانند انواع مختلف رادار، طیف الکترومغناطیسی، تسلیحات، ظرفیت سوخت، برد، تخصیص هدف چند پهاد و برنامه ریزی مشارکتی باید در نظر گرفته شوند. می توان از یادگیری تقویتی برای بهینه سازی مسیر پهادها در محیط های خاص که در معرض خطر ردیابی راداری هستند، استفاده کرد.

۲-۴-۵. الگوریتم های آموزشی کارآمدتر



بلادرنگ ایجاد کنند. در ابتدا یک بررسی سیستماتیک برای درک تصویر کاملی از این زمینه ارائه دادیم. سپس، یک روش تصمیم گیری مانور **dogfight** یک به یک طراحی شد که به پهباد اجازه می دهد در نبرد هوایی کوتاه برد پیروز شود. در نهایت، جهت گیری های احتمالی تحقیقات آینده شرح داده شد. خاطر نشان شده است که تمایل رو به رشدی وجود دارد که ساختارهای شبکه پیشرفته تری در مانورهای نبرد هوایی مبتنی بر **DRL** در حال استفاده می باشد که تا حدودی به بهبود کیفیت تصمیم گیری کمک می کند. به نظر ما، مطالعات آتی باید بیشتر بر روی بهبود سناریوهای خاص نبرد هوایی متمرکز شود. از یک طرف، برای ایجاد یک محیط شبیه سازی واقعی تر، باید اجزای جنگی لازم (مانند سنسورها، سلاح ها و مدل حرکت پهباد با شش درجه آزادی) در نظر گرفته شود. از سوی دیگر، تاکتیک های پیچیده تر (مانند حمله هماهنگ چند پهباد) باید برای انطباق با نیازهای نبرد هوایی مدرن مورد بهره برداری قرار گیرد.

یادگیری تقویتی یک الگوریتم قدرتمند و کلی است، اما همچنان با چالش هایی در بهبود کارایی نمونه و کاهش هزینه های یادگیری در محیط جنگ هوایی پویا و بلادرنگ مواجه است. برخی از روش های بهبود فعلی عبارتند از یادگیری تقویتی برنامه آموزش و طراحی توابع پاداش.

۳-۴-۵. ترکیبی از چندین الگوریتم تصمیم گیری

محیط های پیچیده نبرد هوایی اغلب با فضاهاى حالت بزرگ با روابط پیچیده چند عاملی همراه است. یادگیری تقویتی ممکن است به توابع پاداش بسیار پیچیده برای رسیدگی به رویارویی هوایی در مقیاس بزرگ نیاز داشته باشد. بنابراین، ادغام آن با سایر الگوریتم ها روشی موثر برای حل چنین مسائلی است.

۶. نتیجه گیری

کنترل نبردهای هوایی مبتنی بر **DRL** در سال های اخیر موفقیت های زیادی کسب کرده و توجه زیادی را به خود جلب کرده است. در مقایسه با روش های سنتی، انتظار می رود روش های مبتنی بر **DRL** مانورهایی با کیفیت بالا را بصورت

۷. مراجع

- based on deep reinforcement learning," in *2022 3rd Information Communication Technologies Conference (ICTC)*, 2022: IEEE, pp. 122-126.
- [6] J. Poropudas and K. Virtanen, "Game-theoretic validation and analysis of air combat simulation models," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 5, pp. 1057-1070, 2010.
 - [7] J. B. Cruz et al., "Game-theoretic modeling and control of a military air operation," *IEEE Transactions on aerospace and electronic systems*, vol. 37, no. 4, pp. 1393-1405, 2001.
 - [8] F. Austin, G. Carbone, M. Falco, H. Hinz, and M. Lewis, "Game theory for automated maneuvering during air-to-air combat," *Journal of Guidance, Control, and Dynamics*, vol. 13, no. 6, pp. 1143-1149, 1990.
 - [9] R. Chai, A. Tsourdos, H. Gao, S. Chai, and Y. Xia, "Attitude tracking control for reentry vehicles using centralised robust model predictive control," *Automatica*, vol. 145, p. 110561, 2022.
 - [10] J. Yan, W. Daobo, B. Tingting, and Y. Zongyuan, "Multi-UAV objective
 - [1] A. F. U. Din, I. Mir, F. Gul, and S. Akhtar, "Development of reinforced learning based non-linear controller for unmanned aerial vehicle," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 4005-4022, 2023.
 - [2] K. Zhou, R. Wei, Z. Xu, and Q. Zhang, "A brain like air combat learning system inspired by human learning mechanism," in *2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC)*, 2018: IEEE, pp. 1-6.
 - [3] K. Cui, W. Han, Y. Liu, X. Wang, X. Su, and J. Liu, "Model predictive control for automatic carrier landing with time delay," *International Journal of Aerospace Engineering*, vol. 2021, no. 1, p. 8613498, 2021.
 - [4] X. Gao et al., "Conditional probability based multi-objective cooperative task assignment for heterogeneous UAVs," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106404, 2023.
 - [5] L. Wang and H. Wei, "Research on autonomous decision-making of UCAV

- wireless data harvesting with deep reinforcement learning," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1171-1187, 2021.
- [20] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and J. Henry, "UAV trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9540-9554, 2021.
- [21] W. Zhao, Z. Meng, K. Wang, J. Zhang, and S. Lu, "Hierarchical active tracking control for UAVs via deep reinforcement learning," *Applied Sciences*, vol. 11, no. 22, p. 10595, 2021.
- [22] B. Li and Y. Wu, "Path planning for UAV ground target tracking via deep reinforcement learning," *IEEE access*, vol. 8, pp. 29064-29074, 2020.
- [23] J. Xie, X. Peng, H. Wang, W. Niu, and X. Zheng, "UAV autonomous tracking and landing based on deep reinforcement learning strategy," *Sensors*, vol. 20, no. 19, p. 5630, 2020.
- [24] D. Xu *et al.*, "PPO-Exp: keeping fixed-wing UAV formation with deep reinforcement learning," *Drones*, vol. 7, no. 1, p. 28, 2022.
- [25] W. Zhao *et al.*, "Research on the multiagent joint proximal policy optimization algorithm controlling cooperative fixed-wing UAV obstacle avoidance," *Sensors*, vol. 20, no. 16, p. 4546, 2020.
- [26] X. Zhao, R. Yang, Y. Zhang, M. Yan, and L. Yue, "Deep reinforcement learning for intelligent dual-UAV reconnaissance mission planning," *Electronics*, vol. 11, no. 13, p. 2031, 2022.
- [27] L. Yue, R. Yang, Y. Zhang, L. Yu, and Z. Wang, "Deep reinforcement learning for UAV intelligent mission planning," *Complexity*, vol. 2022, no. 1, p. 3551508, 2022.
- [28] J. Hu, L. Wang, T. Hu, C. Guo, and Y. Wang, "Autonomous maneuver decision making of dual-UAV cooperative air combat based on deep reinforcement learning," *Electronics*, vol. 11, no. 3, p. 467, 2022.
- [29] H. Zhou, X. Zhang, and Z. Zhang, "Reinforcement learning technology for air combat confrontation of unmanned aerial assignment using Hungarian fusion genetic algorithm," *IEEE Access*, vol. 10, pp. 43013-43021, 2022.
- [11] J. Kaneshige and K. Krishnakumar, "Artificial immune system approach for air combat maneuvering," in *Intelligent Computing: Theory and Applications V*, 2007, vol. 6560: SPIE, pp. 68-79.
- [12] R. Chai, A. Tsourdos, A. Savvaris, S. Chai, Y. Xia, and C. P. Chen, "Six-DOF spacecraft optimal trajectory planning and real-time attitude control: a deep neural network-based approach," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 11, pp. 5005-5013, 2019.
- [13] H. Changqiang, D. Kangsheng, H. Hanqiao, T. Shangqin, and Z. Zhuoran, "Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization," *Journal of Systems Engineering and Electronics*, vol. 29, no. 1, pp. 86-97, 2018.
- [14] X. Zhifei, L. Yue, K. Yingxin, L. Zhanwu, and L. You, "An online ensemble semi-supervised classification framework for air combat target maneuver recognition," *Chinese Journal of Aeronautics*, vol. 36, no. 6, pp. 340-360, 2023.
- [15] K. Zhao and C. Huang, "Air combat situation assessment for UAV based on improved decision tree," in *2018 Chinese Control And Decision Conference (CCDC)*, 2018: IEEE, pp. 1772-1776.
- [16] B. Li, Z. Gan, D. Chen, and D. Sergey Aleksandrovich, "UAV maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning," *Remote Sensing*, vol. 12, no. 22, p. 3789, 2020.
- [17] L. Lyu, Y. Shen, and S. Zhang, "The advance of reinforcement learning and deep reinforcement learning," in *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 2022: IEEE, pp. 644-648.
- [18] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied energy*, vol. 235, pp. 1072-1089, 2019.
- [19] H. Bayerlein, M. Theile, M. Caccamo, and D. Gesbert, "Multi-UAV path planning for

- [38] B. Li, S. Bai, S. Liang, R. Ma, E. Neretin, and J. Huang, "Manoeuvre decision-making of unmanned aerial vehicles in air combat based on an expert actor-based soft actor critic algorithm," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 4, pp. 1608-1619, 2023.
- [39] D. Hu, R. Yang, J. Zuo, Z. Zhang, J. Wu, and Y. Wang, "Application of deep reinforcement learning in maneuver planning of beyond-visual-range air combat," *IEEE Access*, vol. 9, pp. 32282-32297, 2021.
- [40] Y. Cao, Y.-X. Kou, Z.-W. Li, and A. Xu, "Autonomous maneuver decision of UCAV air combat based on double deep Q network algorithm and stochastic game theory," *International Journal of Aerospace Engineering*, vol. 2023, no. 1, p. 3657814, 2023.
- vehicle," in *International conference on computer graphics, artificial intelligence, and data processing (ICCAID 2021)*, 2022, vol. 12168: SPIE, pp. 454-459.
- [30] S. Parisi, D. Tateo, M. Hensel, C. D'eraimo, J. Peters, and J. Pajarinen, "Long-term visitation value for deep exploration in sparse-reward reinforcement learning," *Algorithms*, vol. 15, no. 3, p. 81, 2022.
- [31] X. Liu, Y. Yin, Y. Su, and R. Ming, "A multi-UCAV cooperative decision-making method based on an MAPPO algorithm for beyond-visual-range air combat," *Aerospace*, vol. 9, no. 10, p. 563, 2022.
- [32] Z. Fan, Y. Xu, Y. Kang, and D. Luo, "Air combat maneuver decision method based on A3C deep reinforcement learning," *Machines*, vol. 10, no. 11, p. 1033, 2022.
- [33] H. Piao *et al.*, "Beyond-visual-range air combat tactics auto-generation by reinforcement learning," in *2020 international joint conference on neural networks (IJCNN)*, 2020: IEEE, pp. 1-8.
- [34] X. Cao, H. Wan, Y. Lin, and S. Han, "High-value prioritized experience replay for off-policy reinforcement learning," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019: IEEE, pp. 1510-1514.
- [35] H. Zhang, H. Zhou, Y. Wei, and C. Huang, "Autonomous maneuver decision-making method based on reinforcement learning and Monte Carlo tree search," *Frontiers in neurorobotics*, vol. 16, p. 996412, 2022.
- [36] Q. Yang, J. Zhang, G. Shi, J. Hu, and Y. Wu, "Maneuver decision of UAV in short-range air combat based on deep reinforcement learning," *IEEE Access*, vol. 8, pp. 363-378, 2019.
- [37] W. Ruan, H. Duan, and Y. Deng, "Autonomous maneuver decisions via transfer learning pigeon-inspired optimization for UCAVs in dogfight engagements," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 9, pp. 1639-1657, 2022.

