

Design and Simulate an Emotional Deep Reinforcement Learning Controller Based on the Inception Network for Quadrotor Control

Peyman Norouzi, Hamed Shahbazi*, Keivan Torabi

Department of Mechanical Engineering, University of Isfahan, Isfahan, Iran

* shahbazi@eng.ui.ac.ir

- Abstract

Deep reinforcement learning, as one of the powerful approaches in machine learning, has found extensive applications in controlling nonlinear and multivariable systems, including flying robots. However, conventional deep reinforcement learning algorithms still face challenges such as instability in convergence, high computational complexity, and sensitivity to network configuration. In this research, an emotional deep reinforcement learning controller based on the Inception network is introduced and compared with a conventional deep reinforcement learning controller for the path tracking problem in quadrotors. The proposed approach, inspired by the architecture of parallel networks, reflects the dual-path structure of the brain, including the logical and emotional pathways, enabling the formation of a bidirectional decision-making model. This mechanism enhances reaction speed, reduces cognitive load, and improves generalization in unfamiliar conditions. Additionally, in designing the reward function, both simplicity and intelligence are considered so that the agent can remain adaptable to diverse trainable movements. Simulation results indicate that the proposed controller offers significant improvements in flight stability, path tracking accuracy and learning efficiency compared to the conventional version.

- KEYWORDS

Deep reinforcement learning algorithms, Emotional deep reinforcement learning controller, Quadrotor, Neural Networks.

- Introduction

Reinforcement learning (RL) is a fundamental approach in the field of intelligent systems that operates based on causal relationships. In this framework, an intelligent agent interacts with its environment and observes the consequences of its actions, which may manifest as either rewards or penalties[1]. A rewarding outcome indicates a desirable action and moves the agent closer to its objective within the environment, whereas a penalizing outcome indicates an undesirable action and moves the agent farther away from its objective. Therefore, the agent must learn to determine how to obtain higher rewards and ultimately achieve its goal. The primary objective of reinforcement learning is to train an agent to perform a specific task in an uncertain environment. The agent receives feedback and rewards from the environment and responds by taking appropriate actions, treating these rewards as a measure of the effectiveness of its actions in achieving the ultimate objective[2]. The primary objective of the Inception architecture is to increase the depth and width

of the network without significantly increasing the number of parameters or the computational burden[3]. This objective is achieved through the use of a modular, multi-branch block. Each Inception block consists of several parallel pathways for extracting features from the input data. These pathways simultaneously employ convolutions with different kernel sizes to identify features at various scales. At the end of the block, the outputs of all branches are concatenated along the channel axis (depth dimension), forming the final output of the Inception module. This parallel design enables the network to process information at different scales within a single layer, thereby enhancing its ability to identify fine-grained and detailed patterns alongside abstract representations. Furthermore, by employing dimensionality-reduction techniques, such as convolutional operations, the Inception architecture substantially reduces the computational cost while preserving the representational capacity of the network.

In the proposed architecture, the environmental state input is simultaneously fed into the PPO and Inception networks. The Inception network operates as a parallel perceptual processing module alongside the main PPO network, extracting deep and multi-scale features that are subsequently integrated with the intermediate outputs of the PPO network at the final layer. The final output used for action selection is generated based on this enriched and shared representation. In the proposed model, the PPO network is responsible for learning and updating the policy, while the Inception network performs multi-scale processing to extract high-level and abstract features from the input state. The output of the Inception module, represented as a feature vector, is injected into the final layers of the PPO network to improve decision-making accuracy and accelerate convergence. This architecture enhances the agent's ability to generalize to new or previously unseen environments, thereby enabling it to maintain satisfactory performance even in scenarios that it has not encountered during training.

In summary, the model architecture proposed in this study is a hybrid structure that integrates the Proximal Policy Optimization (PPO) algorithm with an Inception network as a parallel perceptual–emotional pathway. The primary objective of this integration is to enhance the performance of the learning agent through multi-scale processing of the environmental state and the injection of high-level features into the main decision-making pathway. The proposed Emotional Deep Reinforcement Learning (EDRL) actor model can likewise be viewed as a decision-making system with two outputs: a logical output and an emotional output. In this model, the first output corresponds to the main PPO network, which serves as the logical agent and is responsible for learning the control policy. The second output corresponds to the Inception network, which operates as a perceptual module and extracts multi-scale, high-level features through a shorter and specialized processing pathway.

- Methodology

In reinforcement learning, the reward function is one of the most critical components, guiding the agent toward optimal decision-making. This function directly influences the agent's learning behavior and performance by quantifying the desirability of a particular action or state. In this study, the reward function is designed with emphasis on achieving satisfactory quadrotor control performance in tracking the desired trajectory. This component penalizes the positional error between the quadrotor and the desired reference trajectory, thereby encouraging the agent to follow the reference path accurately. To achieve these objectives, the overall reward is decomposed into several components that reflect key controllable characteristics of the system. Given that the objective of the problem is defined as trajectory tracking by the quadrotor, the reward function is formulated based on reward of position error (r_{pos}), reward of velocity error (r_{vel}), reward of acceleration error (r_{acc}), and reward of heading error (r_{ψ}). Finally, the total reward r is calculated as the weighted sum of the aforementioned components:

$$r = 0.6r_{pos} + 0.1r_{vel} + 0.1r_{acc} + 0.2r_{\psi} \quad (14)$$

The learning problem was formulated within an episodic reinforcement learning framework, in which each simulation cycle was considered an independent episode. An episode terminates under one of two conditions: (1) successful achievement of the target, or (2) Early Termination (ET) [1], which occurs when a significant

deviation from the reference trajectory or a loss of flight stability is detected. If neither of these conditions is satisfied, the episode continues for a predefined simulation duration, allowing the agent to complete the entire maneuver. The early-termination mechanism ensures that the agent interacts with each state by generating appropriate control outputs, thereby reducing exploration of unproductive regions of the state space and contributing to faster policy convergence.

In this study, each training episode was set to 10 s of simulation time and consisted of 1000 discrete time steps. The cumulative reward during training exhibited a continuous increasing trend, indicating the policy's learning capability and the improvement of training performance over time. Compared with previous studies on quadrotor control using deep reinforcement learning, such as Hwangbo et al. (approximately 2.15 million time steps) [4], Sharifi et al. (approximately 1.2 million time steps) [5], and Pi et al. (10 million time steps) [6], the proposed method demonstrates a substantial improvement in sample efficiency.

To evaluate the effectiveness of the proposed multilayer deep reinforcement learning controller, its performance was compared with that of a conventional deep reinforcement learning approach. First, to compare the behavior of the proposed algorithms, the conventional deep reinforcement learning algorithm was implemented as the baseline method, and its training process was investigated. Using this algorithm, the agent reached the optimal reward after 2948 episodes. Subsequently, it was observed that, using the proposed Emotional Deep Reinforcement Learning algorithm, the agent reached the optimal reward after 1869 episodes and successfully learned to follow the desired trajectory. This substantial reduction in the number of training episodes demonstrates the improved convergence characteristics of the proposed control algorithm, reflecting enhanced learning efficiency and the ability to acquire near-optimal policies with fewer interactions. Furthermore, the incorporation of the Inception network accelerated the training process by approximately 37% compared with the baseline model.

• Results and Discussion

In this study, an Emotional Deep Reinforcement Learning framework integrated with an Inception network was employed for the quadrotor trajectory-tracking task, where adaptability to dynamic and uncertain environments is a critical requirement. Upon completion of the training process for the proposed methods, simulation experiments were conducted using MATLAB to evaluate the performance of the controller. Overall, the proposed Emotional Deep Reinforcement Learning approach demonstrates satisfactory learning efficiency and adaptability, making it suitable for quadrotor navigation and trajectory-tracking tasks. The integration of the Inception network significantly reduced the number of trainable parameters, resulting in faster convergence during the learning process.

Table 1. Performance of the Deep Reinforcement Learning Algorithms

Algorithm	Rise Time	Settling Time	Overshoot	Steady State Error
Emotional Deep Reinforcement Learning	3.1	0	0.10	0.1
Deep Reinforcement Learning	3.3	0.2	0.1	0.1

• Conclusion

In this study, an Emotional Deep Reinforcement Learning control framework based on an Inception network was proposed for quadrotor control. Inspired by the two logical and emotional pathways of the human brain, the proposed architecture was designed such that the logical pathway is modeled by the Proximal Policy Optimization network, while the emotional pathway is represented by the Inception network, which extracts features from intermediate layers through a shortened processing pathway. The simulation results demonstrate that the proposed controller achieves faster learning compared with conventional Deep Reinforcement Learning methods. Based on the simulation results, the proposed Emotional Deep Reinforcement Learning algorithm reached the optimal policy approximately 37% faster, in terms of the

number of required training episodes. Overall, the proposed approach, through the integration of emotional intelligence concepts with reinforcement learning, provides an effective step toward the development of intelligent controllers with adaptive behavior and rapid decision-making capabilities.

- References

- [1] P. Norouzi, H. Shahbazi, K. Torabi, Comparative analysis of a multi-layer deep reinforcement learning controller and an optimized neural network controller utilizing growing particle swarm optimization for quadrotor control, *International Journal of Machine Learning and Cybernetics*, 16(11) (2025) 9007-9033.
- [2] P. Norouzi, H. Shahbazi, K. Torabi, Optimizing quadrotor navigation through emotional deep reinforcement learning: leveraging emotional rewards and states for enhanced training efficiency, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 48(2) (2026) 115.
- [3] P. Norouzi, H. Shahbazi, K. Torabi, A novel deep reinforcement learning framework with human activation function to control quadrotor: P. Norouzi et al, *International Journal of Dynamics and Control*, 14(2) (2026) 69.
- [4] J. Hwangbo, I. Sa, R. Siegwart, M. Hutter, Control of a quadrotor with reinforcement learning, *IEEE Robotics and Automation Letters*, 2(4) (2017) 2096-2103.
- [5] D. Sharifi, K. Torabi, M.I. Rahaghi, H. Shahbazi, Designing a reinforcement learning-based robust controller resisting external forces for a octarotor with new structure, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 45(3) (2023) 141.
- [6] C.-H. Pi, K.-C. Hu, S. Cheng, I.-C. Wu, Low-level autonomous control and tracking of quadrotor using reinforcement learning, *Control Engineering Practice*, 95 (2020) 104222.

طراحی و شبیه‌سازی کنترلر یادگیری تقویتی عمیق عاطفی مبتنی بر شبکه القا برای

کنترل کوادروتور

پیمان نوروزی، حامد شهبازی*، کیوان ترابی

دانشکده فنی و مهندسی، دانشگاه اصفهان، اصفهان، ایران

* shahbazi@eng.ui.ac.ir

چکیده

یادگیری تقویتی عمیق به عنوان یکی از رویکردهای پیشرفته در حوزه یادگیری ماشین، طی سال‌های اخیر به طور گسترده در کنترل سیستم‌های غیرخطی و چندمتغیره، به ویژه در ربات‌های پرنده و کوادروتورها، مورد توجه قرار گرفته است. با وجود پیشرفت‌های چشمگیر، الگوریتم‌های متداول یادگیری تقویتی عمیق همچنان با موانعی از قبیل ناپایداری در فرآیند همگرایی آموزشی، نیاز به توان محاسباتی بالا و حساسیت شدید به پیکربندی شبکه‌های عصبی روبه‌رو هستند. این چالش‌ها موجب کاهش کارایی و محدودیت در تعمیم‌پذیری الگوریتم‌ها در شرایط واقعی می‌شود. در این پژوهش، یک کنترل‌کننده یادگیری تقویتی عمیق عاطفی مبتنی بر ادغام شبکه القا معرفی و با کنترل‌کننده یادگیری تقویتی عمیق مرسوم برای مسئله تعقیب مسیر در کوادروتور مقایسه می‌گردد. معماری پیشنهادی با الهام از ساختار دوسویه مغز انسان شامل مسیر منطقی و مسیر هیجانی طراحی شده است که امکان ایجاد یک مدل تصمیم‌گیری دوسویه را فراهم می‌آورد. این معماری با بهره‌گیری از شبکه‌های موازی، علاوه بر افزایش سرعت واکنش عامل و کاهش بار شناختی، موجب بهبود قابلیت تعمیم در مواجهه با محیط‌های ناشناخته نیز می‌شود. در طراحی تابع پاداش، تلاش شده است تا با حفظ سادگی، عامل توانایی یادگیری تعقیب مسیر را به دست آورد. نتایج شبیه‌سازی عملکرد خوب کنترل‌کننده پیشنهادی در زمینه تعقیب مسیر پرواز و کارایی فرایند یادگیری را نشان می‌دهد.

کلمات کلیدی

یادگیری تقویتی عمیق، یادگیری تقویتی عمیق عاطفی، یادگیری تقویتی عمیق چندلایه، کوادروتور، شبکه‌های عصبی القا

در سال‌های اخیر، کوادروتورها به‌عنوان یکی از پرکاربردترین انواع ربات‌های پرنده، توجه قابل‌توجهی را در حوزه‌های علمی و صنعتی به خود جلب کرده‌اند. ویژگی‌هایی همچون قابلیت برخاست و فرود عمودی، مانورپذیری بالا، طراحی ساده و توانایی انجام مأموریت‌های متنوع مانند نقشه‌برداری، پایش محیطی، جست‌وجو و نجات و حمل محموله‌های سبک، این ربات‌ها را برای کاربردهای نظامی، غیرنظامی و تجاری بسیار جذاب ساخته است. با این حال، کنترل دقیق کوادروتورها به دلیل ماهیت ذاتی غیرخطی، ناپایدار و چندمتغیره این سیستم‌ها همچنان یک چالش اساسی محسوب می‌شود. یکی از مسائل کلیدی، وجود عدم قطعیت‌ها در مدل دینامیکی است که می‌تواند ناشی از اغتشاشات محیطی، تغییر بار، نویز خارجی یا خطاهای مدل‌سازی باشد. این عدم قطعیت‌ها می‌توانند منجر به رفتار غیرقابل پیش‌بینی و ناپایداری در عملکرد سیستم شوند.

روش‌های کنترلی موجود برای کوادروتورها را می‌توان به طور کلی به دو دسته مبتنی بر مدل و بدون مدل تقسیم کرد. کنترل‌کننده‌های مبتنی بر مدل، مانند PID و LQR، به مدل‌های دینامیکی دقیق متکی هستند و در محیط‌های پایدار و قابل پیش‌بینی عملکرد خوبی دارند. در مقابل، کنترل‌کننده‌های بدون مدل، به‌ویژه آن‌هایی که بر پایه یادگیری ماشین عمل می‌کنند، در محیط‌های پویا و نامطمئن، سازگاری بیشتری ارائه می‌دهند. در این میان، یادگیری تقویتی عمیق^۱ به دلیل توانایی یادگیری از تعامل با محیط در کنار بهره‌گیری از شبکه‌های عصبی عمیق، به‌عنوان یک راهکار نویدبخش مطرح شده است. تاکنون چندین الگوریتم یادگیری تقویتی عمیق برای وظایف کنترلی توسعه یافته‌اند، از جمله شبکه Q عمیق (DQN)^۲، بهینه‌سازی خط‌مشی جانبی (PPO)^۳، بهینه‌سازی خط‌مشی ناحیه اعتماد (TRPO)^۴ و معماری‌های بازیگر-منتقد^۵. الگوریتم DQN، به‌عنوان یکی از نخستین الگوریتم‌های موفق DRL، از شبکه‌های عصبی برای تقریب توابع ارزش استفاده می‌کند، اما با مشکلاتی نظیر ناپایداری در آموزش و همبستگی داده‌ها مواجه است. برای رفع این محدودیت‌ها، نسخه‌های بهبودیافته‌ای همچون Double-DQN پیشنهاد شده‌اند که به ترتیب به کاهش برآورد بیش از حد ارزش و جداسازی بهتر ارزش حالت و مزیت اقدام کمک می‌کنند. با این وجود، این رویکردها همچنان نیازمند تنظیم دقیق و پیچیده پارامترها هستند. در سوی دیگر، الگوریتم‌های مبتنی بر خط‌مشی مانند PPO و TRPO برای بهبود پایداری و کارایی یادگیری معرفی شده‌اند. به‌ویژه، PPO که به دلیل سادگی و عملکرد مقاوم خود در محیط‌های پویا، از محبوبیت بالایی برخوردار است و با ایجاد تعادل میان کاوش و بهره‌برداری، یادگیری را بهبود می‌بخشد. TRPO نیز تلاش می‌کند با اعمال محدودیت‌های سخت‌گیرانه بر به‌روزرسانی خط‌مشی، پایداری بیشتری ایجاد کند، هرچند از نظر محاسباتی سنگین‌تر است.

یادگیری تقویتی عمیق به‌عنوان یک حوزه نوظهور در یادگیری ماشین، ترکیبی از یادگیری تقویتی کلاسیک و شبکه‌های عصبی عمیق است و در سال‌های اخیر توجه زیادی را به خود جلب کرده است. یادگیری تقویتی عمیق کاربردهای گسترده‌ای در حل مسائل پیچیده تصمیم‌گیری بلندمدت مانند کنترل ربات، بازی‌های ویدئویی و سامانه‌های خودمختار پیدا کرده است. یکی از پیشرفت‌های بنیادین در یادگیری تقویتی، معرفی الگوریتم Q-Learning در سال ۱۹۹۲ بود که به‌عنوان یک روش بدون مدل برای یافتن خط‌مشی بهینه در فرآیندهای تصمیم‌گیری مارکوف شناخته می‌شود [۱]. با پیشرفت سخت‌افزارهای محاسباتی و افزایش توان پردازشی، استفاده از شبکه‌های عصبی عمیق در یادگیری تقویتی گسترش یافته است. نقطه عطف در این حوزه، معرفی شبکه DQN در مرجع [۲] بود که با استفاده از بازپخش تجربه و شبکه هدف، پایداری آموزش را بهبود بخشید که عملکردی فراتر از انسان در بازی‌های آتاری^۶ به دست آورد و نقطه عطف مهمی به‌شمار می‌رفت. برای بهبود بیشتر DQN، نسخه‌هایی همچون Double-DQN که مشکل برآورد بیش از حد مقادیر Q را برطرف می‌کند، و ارزش حالت را از مزیت اقدام جدا می‌سازد، ارائه شده‌اند [۳، ۴]. در کنار روش‌های مبتنی بر ارزش، الگوریتم‌های مبتنی بر خط‌مشی نیز به طور گسترده توسعه یافته‌اند. الگوریتم بهینه‌سازی PPO به دلیل سهولت پیاده‌سازی و

¹ Deep Reinforcement Learning (DRL)

² Deep Q-Network

³ Proximal Policy Optimization

⁴ Trust Region Policy Optimization

⁵ Actor-Critic

⁶ Atari

عملکرد پایدار خود در محیط‌های پویا، محبوبیت زیادی یافته است [۵]. الگوریتم بهینه‌سازی خط‌مشی TRPO با اعمال محدودیت‌های سخت‌گیرانه بر به‌روزرسانی خط‌مشی، پایداری آموزش را بیش‌تر بهبود می‌بخشد [۶]. همچنین روش‌های بازیگر-منتقد که از دو شبکه جداگانه برای یادگیری خط‌مشی (بازیگر) و ارزیابی تابع ارزش (منتقد) استفاده می‌کنند، با ترکیب مزایای رویکردهای مبتنی بر ارزش و خط‌مشی، عملکرد قدرتمندی نشان داده‌اند [۷]. یادگیری تقویتی عمیق برای کوادروتورها پتانسیل دستیابی به مانورهای پروازی پایدار و چابک را نشان داده است. با این حال، هنوز چالش‌هایی در پیاده‌سازی یادگیری عمیق همچون فرآیندهای یادگیری زمان‌بر، ابعاد بالا در فضای حالت، الزامات محاسباتی زیاد و عدم توانایی تعمیم به محیط‌های جدید برای ربات‌ها وجود دارد [۸]. کار بر روی این چالش‌ها مورد توجه محققان قرار گرفته و یکی از راهکارها استفاده از هوش زیستی انسانی بر پایه‌ی هوش منطقی و عاطفی است که در این مقاله به آن پرداخته شده است. در ادامه به پیشینه استفاده از هوش عاطفی در محاسبات پرداخته شده است. اولین گام به سمت این رویکرد در مرجع [۹] با ارائه اولین رویکرد ریاضی برای توصیف رفتار فرآیند یادگیری عاطفی مغز^۱ انجام شده است. ایده به کارگیری این الگوریتم یادگیری در ناحیه کنترل خودکار در [۱۰] ارائه شده است. پیوند BEL با طراحی سیستم‌های کنترلی منجر به ایجاد مفهومی به نام کنترل هوشمند مبتنی بر یادگیری هیجانی مغز^۲ می‌باشد [۱۱] و [۱۲] و [۱۳] که فرآیند یادگیری عاطفی را در سیستم لیمبیک مغز انسان شبیه‌سازی می‌کند [۱۵] و [۱۶]. این الگوریتم به‌طور گسترده‌تری در کاربردهای مختلف مهندسی کنترل، از جمله موتورهای الکتریکی، سیستم‌های سروو، کنترل حرکت، کاهش تلاش کنترلی، سیستم‌های قدرت و غیره مورد استفاده قرار گرفته است [۱۷]. ظرفیت یادگیری، حداقل پیچیدگی محاسباتی و مهم‌تر از همه عدم نیاز به دانش قبلی از دینامیک سیستم، BELBIC را به یک کنترل‌کننده متمایز نسبت به کنترل‌کننده‌های هوشمند مورد بحث قبلی، مانند ANN^۳ و FLC^۴ تبدیل می‌کند [۱۸].

در [۱۹] مقایسه بین الگوریتم کنترلی تناسبی-انتگرال‌گیر-مشتق‌گیر با کنترلر هوش عاطفی تنظیم‌شده با الگوریتم ازدحام ذرات، برتری کنترلر طراحی‌شده در تعقیب ورودی مطلوب با تلاش کنترلی کمتر برای مدل مبدل DC-DC را نتیجه داده است. همچنین کنترلر طراحی‌شده توانایی مقابله با اغتشاشات را داشته و نتایج خوبی را از خود نشان داده است. در [۲۰] با توجه به اینکه سلول‌های سوختی سیستم‌های پیچیده غیرخطی و متغیر با زمان هستند، چگونگی کنترل و تثبیت مؤثر ولتاژ خروجی آن‌ها به چالش اصلی برای کاربرد گسترده آن‌ها تبدیل می‌شود. برای بهبود کارایی عملکرد و کنترل ردیابی توان و جلوگیری از نقض محدودیت‌های کنترلی در این مقاله یک کنترل‌کننده تناسبی-انتگرال‌گیر-مشتق‌گیر تطبیقی مبتنی بر یادگیری عمیق طراحی‌شده است که مقدار جریان هیدروژن را در مقدار مرجع حفظ می‌کند و ولتاژ خروجی را تثبیت می‌کند. در [۲۱] محققین به طراحی یک تشخیص‌دهنده شی بصری با الهام از یادگیری عاطفی مغز انسان پرداخته و ساختار شبکه عاطفی زیستی را پیاده‌سازی کرده‌اند. در این مدل لایه‌های آمیگدال^۵، اوربیتوفرونتال کورتکس^۶ و سنسوری کورتکس^۷ شبیه‌سازی شده‌اند که از ویژگی‌های اصلی مدل پیشنهادی، سرعت پاسخ‌دهی بیشتر و پیچیدگی کمتر مدل می‌باشد. در [۲۲] یک الگوریتم یادگیری برای مدل محاسباتی مبتنی بر امیگدال و اوربیتوفرونتال کورتکس به روش نظارت‌شده با در نظر گرفتن نرخ فروپاشی امیگدال که پایه عصبی زیستی دارد، پیشنهاد شده است. نتایج روش پیشنهادی با روش‌های پرسپترون چندلایه، سیستم استنتاج عصبی فازی تطبیقی^۸ و عصبی فازی خطی محلی (LINF)^۹ مقایسه شده است. دقت بالاتر در نقاط بحرانی، پیچیدگی محاسباتی کمتر و آموزش تطبیقی از ویژگی‌های مدل پیشنهادی می‌باشد. در [۲۳] یک مدل ریاضی هوش عاطفی برای مسائل طبقه‌بندی با استفاده از تابع فعال‌سازی^{۱۰} تانسینگ ارائه شده و نتایج مورد مقایسه قرار گرفته‌اند. در [۲۴]، روشی ارائه شده است که یادگیری تقویتی توسط احساسات هدایت می‌شود و باعث می‌شود احساسات در کنترل

¹ Brain Emotional Learning (BEL)

² Brain Emotional Learning Based Intelligent Controller (BELBIC)

³ Artificial neural network

⁴ Fuzzy logic controller

⁵ Amygdala

⁶ Orbito-frontal Cortex

⁷ Sensory Cortex

⁸ Artificial Neural Fuzzy intelligence system (ANFIS)

⁹ Linear intelligence Neural Fuzzy

¹⁰ Activation function

یک سیستم دخیل شود و با بازخوردهای عاطفی به یادگیری تقویتی کمک کند. در این روش، احساسات به عنوان یک سیگنال پاداش است و عامل یاد می گیرد که به گونه ای رفتار کند تا به احساسات مطلوب برسد و از احساسات بد دوری کند. در [۲۵]، روشی جهت تصمیم گیری در شرایط اضطراری و مبتنی بر هوش هیجانی انسان ارائه شده است که می تواند ویژگی های شخصیتی یک عامل، رفتارهای عاطفی آن و رویدادهای بیرونی که بر آن تأثیر می گذارد را ترکیب کند و یک تصمیم گیری با رویکردی جدید در شبیه سازی نجات روبوکاپ انجام دهد.

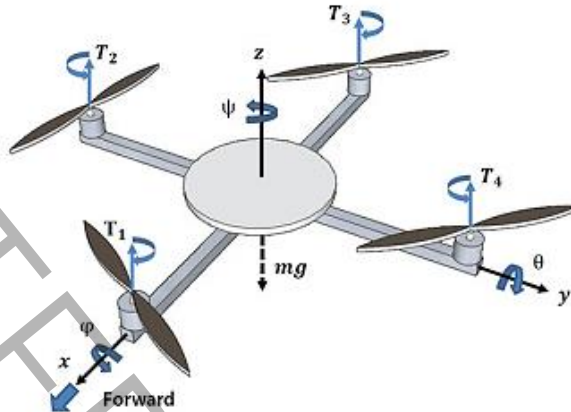
هدف کلی الگوریتم یادگیری، تعیین یک سیاست بهینه است که پاداش تجمعی مورد انتظار را در طول زمان به حداکثر برساند. اجرای حرکات مختلف با کوادروتور چالش های قابل توجهی را به همراه دارد، حتی برای خلبانان انسانی با تجربه که به تمرین گسترده ای برای تسلط بر این حرکات نیاز دارند. با این حال، پیشرفت های اخیر در هوش مصنوعی قابلیت شگفت انگیز آن را در تعامل با محیط و انجام وظایف پیچیده نشان داده است. با وجود پیشرفت های چشمگیر در زمینه هوش مصنوعی، توسعه کنترل کننده های مبتنی بر یادگیری ماشین که قادر به انجام مانورهای چابک در شرایط نامطمئن باشند، همچنان یک معما مهم در تحقیقات باقی مانده است. علاوه بر این، آموزش چنین شبکه هایی به زمان قابل توجه و تخصص زیادی برای پیاده سازی نیاز دارد که معمولاً نیازمند فرآیندهای دقیق تنظیم مجدد است. شبکه های عصبی پیچشی (CNNs)^۱ به عنوان ابزار اصلی برای پردازش داده های دیداری و مکانی مطرح شده اند. این شبکه ها با استفاده از فیلترهای محلی، الگوهای مکانی را از داده های ورودی استخراج می کنند. با افزایش عمق CNN ها، توانایی آن ها در شناسایی ویژگی های سطح بالا بهبود می یابد؛ با این حال، این امر چالش هایی همچون افزایش پیچیدگی محاسباتی، محو شدن گرادین ها و اشباع شبکه را نیز به همراه دارد. برای رفع این مشکلات و ارتقای کارایی معماری های عمیق، طراحی های پیشرفته تری مانند معماری Inception ارائه شده اند. این معماری ها از ساختارهای ماژولار بهره می برند که امکان استخراج ویژگی در مقیاس های چندانگانه را فراهم کرده و در عین حال، تعداد پارامترها و هزینه های محاسباتی را کاهش می دهند. معماری Inception نخستین بار توسط Szegedy و همکاران در مقاله ی [۲۶] معرفی شد. این معماری در مدل شناخته شده ی GoogLeNet پیاده سازی گردید که عملکرد برجسته ای را در رقابت ImageNet به دست آورد. به دلیل طراحی نوآورانه اش، معماری Inception به عنوان یک پیشرفت چشمگیر در توسعه ی شبکه های عمیق شناخته می شود. با وجود پیشرفت های چشمگیر در الگوریتم های یادگیری تقویتی عمیق، چالش هایی همچون پیچیدگی ساختار شبکه های عمیق، حساسیت به مقاردهای اولیه و تنظیم ابرپارامترها همچنان پابرجاست. این مشکلات معمولاً منجر به افزایش زمان آموزش، هزینه محاسباتی بالا و ناپایداری در فرایند یادگیری می شوند. الهام گیری از سازوکار تصمیم گیری دومسیره ی مغز انسان شامل مسیر منطقی (قشر پیش پیشانی) و مسیر هیجانی (آمیگدالا و دستگاه لیمبیک) می تواند به طراحی سامانه های یادگیرنده ای منجر شود که ضمن تحلیل منطقی، قادر به اتخاذ تصمیم های سریع و واکنشی نیز باشند. این ساختار زیستی باعث افزایش سرعت واکنش، کاهش بار شناختی و بهبود قابلیت تعمیم در موقعیت های ناآشنا می شود. در راستای رفع چالش های فوق، این پژوهش یک چارچوب کنترل یادگیری تقویتی عمیق عاطفی مبتنی بر شبکه ی القا^۲ را برای کنترل کوادروتور در محیط های ناشناخته معرفی می کند. در این ساختار، شبکه یادگیری تقویتی عمیق به عنوان مسیر منطقی و شبکه القا به عنوان مسیر هیجانی عمل می کنند و به صورت موازی در فرآیند تصمیم گیری مشارکت دارند. معماری پیشنهادی با الهام از تعامل میان قشر پیش پیشانی و آمیگدالا در مغز انسان، موجب افزایش پایداری، سرعت یادگیری و قابلیت تعمیم عامل یادگیرنده می شود. شبکه ی القا به عنوان یک مسیر پردازشی موازی، ویژگی های چندمقیاسی و انتزاعی محیط را استخراج کرده و آن ها را به مسیر تصمیم گیرنده ی شبکه یادگیری تقویتی عمیق تزریق می کند. به طور خلاصه، در این پژوهش به معرفی و به کارگیری یک ساختار دومسیره ی الهام گرفته از مغز انسان، بر پایه ی ترکیب مسیرهای منطقی و هیجانی، برای ارتقای سرعت آموزش و تعمیم پذیری در کنترل کوادروتور پرداخته شده است.

^۱ Convolutional Neural Networks

^۲ Inception

۲- مدلسازی کوادروتور

کوادروتور یک وسیله پرنده با شش درجه آزادی حرکت باقابلیت پرواز عمودی و انجام مانورهای پیچیده را دارا می‌باشد. این سازه دارای ساختاری شبیه صلیب می‌باشد که چهار ملخ در گوشه‌های آن قرار دارد و با استفاده از تغییر سرعت ملخ‌ها می‌تواند حرکات و مانورهای مختلف را انجام بدهد. شکل ۱ پیکربندی چهارپره را نشان می‌دهد. در این آرایش، موتورهای ۱ و ۳ در جهت عقربه‌های ساعت و موتورهای ۲ و ۴ در خلاف جهت عقربه‌های ساعت می‌چرخند.



شکل ۱: ساختار کوادروتور

تنظیم یکنواخت سرعت همه روتورها نیروی بالابر را تغییر داده و امکان کنترل ارتفاع و برخاست عمودی سیستم را فراهم می‌کند. برای دستیابی به زاویه یاو (ψ) مورد نظر، سرعت روتورهای روبه‌روی هم تنظیم می‌شود؛ به این صورت که سرعت موتورهای ۱ و ۳ نسبت به موتورهای ۲ و ۴ افزایش یا کاهش می‌یابد. حرکت چهارپره در راستای محور X توسط زاویه پیچش (θ) حول محور Y کنترل می‌شود، در حالی که حرکت در راستای محور Y توسط زاویه رول (ϕ) حول محور X کنترل می‌گردد. روتورها نقش کلیدی در تعیین حرکت چهارپره ایفا می‌کنند و برای محاسبه توان و گشتاور لازم جهت کنترل زاویه و موقعیت، از زوایای اوایلر استفاده می‌شود. مدل دینامیکی کوادروتور به صورت روابط (۱) تا (۶) از مرجع [۲۷] گرفته شده است.

$$\ddot{x} = (\cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi) \frac{U_1}{m} \quad (۱)$$

$$\ddot{y} = (\cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi) \frac{U_1}{m} \quad (۲)$$

$$\ddot{z} = -g + (\cos(\phi) \cos(\theta)) \frac{U_1}{m} \quad (۳)$$

$$\ddot{\phi} = \dot{\phi} \dot{\psi} \left[\frac{I_{yy} - I_{zz}}{I_{xx}} \right] + \frac{J_r}{I_{xx}} \dot{\Omega}_r + \frac{1}{I_{xx}} U_2 \quad (۴)$$

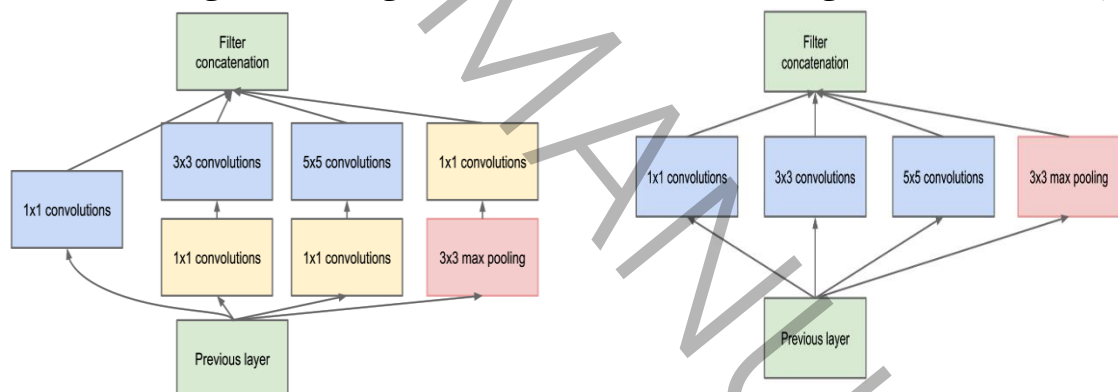
$$\ddot{\theta} = \dot{\phi} \dot{\psi} \left[\frac{I_{zz} - I_{xx}}{I_{yy}} \right] - \frac{J_r}{I_{yy}} \dot{\Omega}_r + \frac{1}{I_{yy}} U_3 \quad (۵)$$

$$\ddot{\psi} = \dot{\phi} \dot{\psi} \left[\frac{I_{xx} - I_{yy}}{I_{zz}} \right] + \frac{1}{I_{zz}} U_4 \quad (۶)$$

در این روابط، موقعیت‌های مرکز جرم در ساختار به ترتیب با x ، y و z نشان داده شده‌اند و زوایای اوایلر متناظر به ترتیب با ϕ ، θ و ψ نمایش داده می‌شوند. متغیرهای I_x ، I_y و I_z به ترتیب جرم و ممان‌های اینرسی را نشان می‌دهند و l فاصله بین روتورها و مرکز جرم را مشخص می‌کند. همچنین، Ω سرعت زاویه‌ای و J_r ممان اینرسی روتورها را بیان می‌کند. علاوه بر این، U_1 ، U_2 ، U_3 و U_4 ورودی‌های سیستم هستند که مسئول تولید نیروی تجمعی در راستای محور z و گشتاور چرخشی حول محور y می‌باشند. این ورودی‌ها همچنین در ایجاد چرخش حول محورهای x و z توسط روتورها نقش دارند.

۳- یادگیری تقویتی عمیق عاطفی مبتنی بر شبکه القا^۱

یادگیری تقویتی یک روش اساسی در زمینه سیستم‌های هوشمند است که بر اساس روابط علی عمل می‌کند. در این رویکرد، عامل هوشمند با محیط تعامل دارد و منتظر نتایج اعمال خود می‌ماند که ممکن است به صورت پاداش یا تنبیه بروز کند. نتیجه‌ای که پاداش دارد، نشان‌دهنده یک عمل مطلوب است و عامل را به هدفش در محیط نزدیک‌تر می‌کند، در حالی که نتیجه تنبیهی نشان‌دهنده یک عمل نامطلوب است و عامل را از هدفش دورتر می‌سازد. بنابراین، عامل باید تشخیص دهد که چگونه می‌تواند پاداش‌های بیشتری کسب کند و در نهایت به هدف خود دست یابد. هدف اصلی یادگیری تقویتی، آموزش یک عامل برای انجام یک وظیفه در یک محیط نامعین است. عامل بازخورد و پاداش‌هایی از محیط دریافت می‌کند و با اقدام‌های مناسب پاسخ می‌دهد و این پاداش‌ها را به عنوان معیاری برای اثربخشی عمل در دستیابی به هدف نهایی در نظر می‌گیرد. یادگیری تقویتی دو جزء اساسی دارد: سیاست و الگوریتم‌های یادگیری. سیاست، اقدام‌ها را بر اساس مشاهدات محیطی تعیین می‌کند که معمولاً به عنوان یک تابع تقریب پارامتری یا یک شبکه عصبی عمیق نمایش داده می‌شود. به طور همزمان، الگوریتم یادگیری به طور مداوم پارامترهای سیاست را بر اساس اقدامات، مشاهدات و پاداش‌ها تنظیم می‌کند. هدف اصلی Inception افزایش عمق و عرض شبکه بدون افزایش چشمگیر در تعداد پارامترها یا بار محاسباتی می‌باشد. این هدف از طریق به کارگیری یک بلوک ماژولار چندشاخه محقق شده است. هر بلوک Inception شامل چند مسیر موازی برای استخراج ویژگی‌ها از داده‌ی ورودی است. این مسیرها به طور همزمان از کانولوشن‌هایی با اندازه هسته‌های مختلف استفاده می‌کنند تا ویژگی‌ها را در مقیاس‌های گوناگون شناسایی کنند. در پایان بلوک، خروجی تمامی شاخه‌ها در امتداد محور کانال (بعد عمق) به یکدیگر متصل می‌شوند و خروجی نهایی ماژول Inception را تشکیل می‌دهند (شکل ۲). این طراحی موازی به شبکه اجازه می‌دهد تا در یک لایه، اطلاعات را در اندازه‌های مختلف پردازش کند و در نتیجه توانایی آن در شناسایی الگوهای جزئی و دقیق در کنار الگوهای انتزاعی تقویت شود. علاوه بر این، با به کارگیری روش‌های کاهش ابعاد همچون کانولوشن، معماری Inception به طور چشمگیری هزینه‌ی محاسباتی را کاهش داده و در عین حال، قدرت بازنمایی شبکه را حفظ می‌کند.

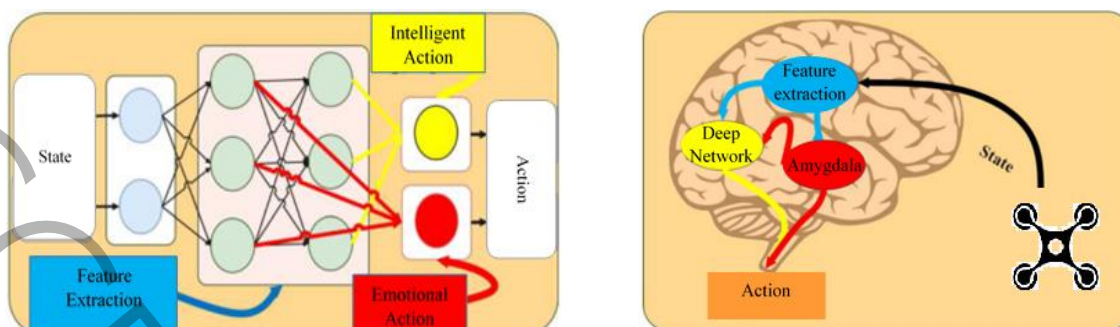


شکل ۲: الف) ماژول Inception، نسخه ابتدایی ب) ماژول Inception همراه با کاهش بُعد [۲۶]

در مغز انسان، تصمیم‌گیری صرفاً یک فرآیند منطقی و تحلیلی نیست؛ بلکه سامانه‌های عاطفی نیز نقشی حیاتی ایفا می‌کنند. هوش هیجانی^۲ به توانایی مغز در درک، تفسیر، پردازش و واکنش به محرک‌های هیجانی اشاره دارد. این قابلیت به انسان اجازه می‌دهد به‌ویژه در موقعیت‌های پیچیده، مبهم یا بحرانی، به سرعت، به‌طور مؤثر و سازگارانه واکنش نشان دهد. این ساختار دو مسیره‌ی مغز، شامل مسیر منطقی و مسیر هیجانی، یک مدل تصمیم‌گیری دوسویه را تشکیل می‌دهد (شکل ۳). در این مدل، مغز می‌تواند به‌طور هم‌زمان تحلیل منطقی دقیق انجام دهد و همچنین بر اساس الگوهای هیجانی و تجربه‌های آموخته‌شده، تصمیم‌های سریع اتخاذ کند. این سازوکار باعث افزایش سرعت واکنش، کاهش بار شناختی و بهبود تعمیم در موقعیت‌های ناآشنا می‌شود.

^۱ Inception

^۲ Emotional Intelligence



شکل ۳: چارچوب یادگیری تقویتی عمیق عاطفی مبتنی بر Inception و فرآیند تصمیم‌گیری انسانی

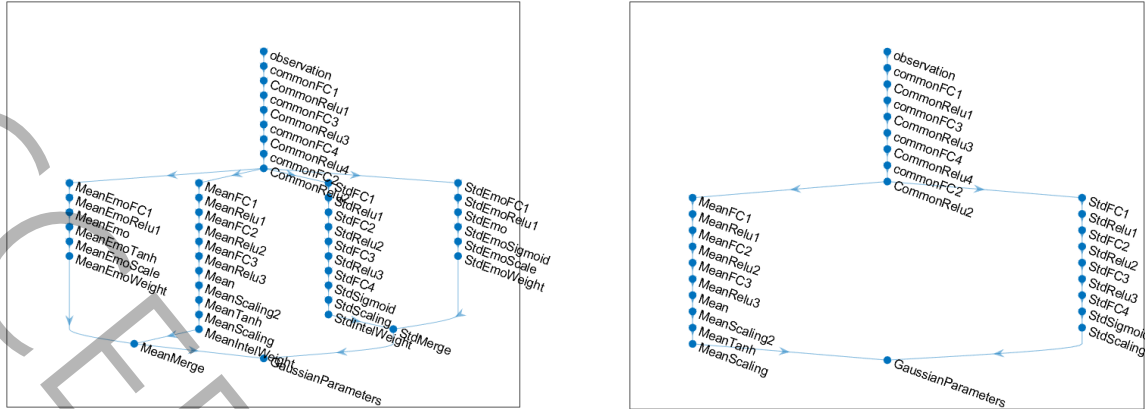
در معماری پیشنهادی، ورودی حالت محیط به‌طور هم‌زمان به شبکه‌های PPO و Inception ارسال می‌شود. شبکه Inception به‌عنوان یک ماژول پردازش ادراکی موازی در کنار شبکه‌ی اصلی PPO عمل می‌کند و ویژگی‌های عمیق و چندمقیاسی را استخراج می‌کند و در لایه‌های نهایی با خروجی‌های میانی شبکه PPO ادغام می‌شوند. خروجی نهایی برای انتخاب عمل، بر اساس این بازنمایی غنی و مشترک تولید می‌گردد. در مدل پیشنهادی شبکه PPO مسئول یادگیری و به‌روزرسانی خطمشی، شبکه Inception پردازش چندمقیاسی را برای استخراج ویژگی‌های سطح بالا و انتزاعی از حالت ورودی انجام می‌دهد و خروجی ماژول Inception، به‌صورت یک بردار ویژگی، به لایه‌های نهایی شبکه PPO تزریق می‌شود تا دقت تصمیم‌گیری و سرعت همگرایی بهبود یابد. این معماری توانایی عامل را در تعمیم به محیط‌های جدید یا ناشناخته افزایش داده و امکان حفظ عملکرد رضایت‌بخش را حتی در سناریوهایی که قبلاً با آن‌ها مواجه نشده است، فراهم می‌سازد. به‌طور خلاصه، معماری مدلی که در این مطالعه پیشنهاد شده است، یک ساختار ترکیبی است که الگوریتم PPO را با یک شبکه‌ی Inception به‌عنوان مسیر ادراکی-هیجانی موازی تلفیق می‌کند. هدف اصلی از این یکپارچه‌سازی، ارتقای عملکرد عامل یادگیرنده از طریق پردازش چندمقیاسی حالت محیط و تزریق ویژگی‌های سطح بالا به مسیر اصلی تصمیم‌گیری است. شکل ۴ نمایش کلی از معماری شبکه‌ی عملگر پیشنهادی با لایه‌های مورد استفاده برای ساختار شبکه شامل لایه‌های کاملاً متصل^۱، رلو^۲، تانژانت هیپربولیک^۳ و غیره را نشان می‌دهد. در این معماری، حالت‌ها^۴ پس از پیش‌پردازش اولیه در مدل عملگر سنتی گوسی به‌طور هم‌زمان به دو مسیر موازی تغذیه می‌شود تا یک خروجی منطقی را تولید کنند. از آنجاییکه فرایندهای گاوسی بر دو مفهوم اصلی میانگین و کوواریانس بنا شده است در مدل عملگر سنتی گوسی دو مسیر وجود دارد که یک خروجی منطقی را ارائه می‌دهد. مدل عملگر پیشنهادی یادگیری تقویتی عمیق عاطفی طراحی شده نیز به‌عنوان یک سامانه‌ی تصمیم‌گیری با دو خروجی منطقی و عاطفی نیز قابل مشاهده می‌باشد. در این مدل خروجی نخست مربوط به شبکه‌ی اصلی PPO است که نقش عامل منطقی را بر عهده دارد و مسئول یادگیری خطمشی کنترلی است. خروجی دوم مربوط به شبکه‌ی Inception می‌باشد که به‌عنوان یک ماژول ادراکی عمل کرده و از طریق یک مسیر پردازشی کوتاه‌تر و تخصصی، ویژگی‌های چندمقیاسی و سطح بالا را استخراج می‌کند.

¹ Fully Connected Layer

² Relu Layer

³ Tanh Layer

⁴ states



شکل ۴: الف) مدل عملگر سنتی گوسی^۱ (ب) مدل عملگر پیشنهادی یادگیری تقویتی عمیق عاطفی طراحی شده به عنوان یک سامانه‌ی تصمیم‌گیری دو مسیره الهام گرفته از مغز انسان

۳-۱- یادگیری تقویتی عمیق مبتنی بر PPO

در این تحقیق، الگوریتم یادگیری تقویتی عمیق مبتنی بر PPO به عنوان روش اصلی یادگیری تقویتی عمیق برای کنترل کوادروتور به کار گرفته شده است. PPO یک روش مبتنی بر گرادین سیاست است که مزایای بهینه‌سازی مستقیم سیاست را با برآورد ارزش ترکیب می‌کند و تعادلی مطلوب بین بهره‌وری نمونه و پایداری آموزش برقرار می‌سازد. این ویژگی باعث می‌شود PPO به‌ویژه برای مسائل کنترل پیوسته که نیازمند تطبیق سریع و پایداری هستند، مانند پرواز کوادروتور در محیط‌های دینامیک، بسیار موثر باشد. در یادگیری تقویتی، عامل^۲ با محیط از طریق انتخاب عمل a_t در حالت s_t تعامل می‌کند، پاداش اسکالری r_t دریافت کرده و سیاست خود $\pi(a|s)$ را برای بهینه‌سازی بازده تجمعی مورد انتظار به‌روزرسانی می‌کند (شکل ۲):

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) \right] \quad (7)$$

که در آن θ پارامترهای سیاست، $\tau = (s_0, a_0, s_1, a_1, \dots, s_T)$ یک مسیر، $r(s_0, a_0)$ پاداش و γ عامل تخفیف است. برای جلوگیری از به‌روزرسانی‌های بزرگ و ناپایدار سیاست، PPO تابع هدف جایگزین برش‌خورده‌ای^۳ معرفی می‌کند:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t) \right] \quad (8)$$

که در آن $r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}$ نسبت احتمال بین سیاست فعلی و سیاست قدیمی، A_t مزیت تخمینی و ϵ محدوده برش را کنترل می‌کند (برابر با ۰.۲ در نظر گرفته شده است) [۲۸]. این برش تضمین می‌کند که به‌روزرسانی‌ها در یک محدوده اعتماد باقی بمانند و پایداری حفظ شود. پارامترهای سیاست از طریق صعود گرادین تصادفی به‌روزرسانی می‌شوند:

$$\theta_{k+1} = \theta_k + \alpha \nabla_{\theta} L^{\text{CLIP}}(\theta_k) \quad (9)$$

که در آن α نرخ یادگیری است.

¹ Gaussian model

² Agent

³ Clipped Surrogate Objective

۳-۲- طراحی تابع پاداش

در یادگیری تقویتی، تابع پاداش یکی از حیاتی‌ترین اجزاء است که عامل را به سوی تصمیم‌گیری بهینه هدایت می‌کند. این تابع به‌طور مستقیم رفتار و کارایی یادگیری عامل را با کمی‌سازی میزان مطلوبیت یک عمل یا حالت خاص تحت تأثیر قرار می‌دهد. در این مقاله برای طراحی تابع پاداش، عملکرد مناسب کنترل کوادروتور در دنبال کردن مسیر مطلوب مورد توجه قرار گرفته است. این بخش خطای موقعیتی بین کوادروتور و مسیر مرجع مطلوب را جریمه می‌کند و موجب تشویق عامل به دنبال کردن دقیق مسیر می‌شود. برای دستیابی به این اهداف، پاداش کل به چند مؤلفه تقسیم شده است که ویژگی‌های کلیدی قابل کنترل را بازتاب می‌دهند. با توجه به هدف مسئله که بصورت تعقیب مسیر مطلوب توسط ربات تعریف شده است، تابع پاداش بر اساس خطای موقعیت (e_{pos})، خطای سرعت (e_{vel})، خطای شتاب (e_{acc}) و انحراف جهت‌یابی (e_{psi}) تعریف شده است. این مؤلفه‌ها به صورت توابع نمایی از مربع خطاهای نرمال‌شده بصورت معادلات (۱۰) و (۱۱) و (۱۲) محاسبه می‌شوند:

$$r_{pos} = -\exp\left[-\left(\frac{1}{c_p} \|e_{pos}\|\right)^2\right] \quad (10)$$

$$r_{vel} = -\exp\left[-\left(\frac{1}{c_v} \|e_{vel}\|\right)^2\right] \quad (11)$$

$$r_{acc} = -\exp\left[-\left(\frac{1}{c_a} \|e_{acc}\|\right)^2\right] \quad (12)$$

که در آن $c_p=0.5(m)$ ، $c_v=1(m/s)$ و $c_a=1(m/s^2)$ و بصورت ضرایب ثابت معرفی می‌شوند. برای تشویق به پایداری چرخشی، به ویژه حول محور Z (ψ)، مؤلفه جریمه انحراف ψ نیز در نظر گرفته شده است:

$$r_{\psi} = -\exp\left[-\left(\frac{1}{c_{\psi}} \|e_{\psi}\|\right)^2\right] \quad (13)$$

که در آن $c_{\psi}=5(\text{degree})$ است. در نهایت، پاداش کل r به صورت مجموع وزنی مؤلفه‌های فوق محاسبه می‌شود:

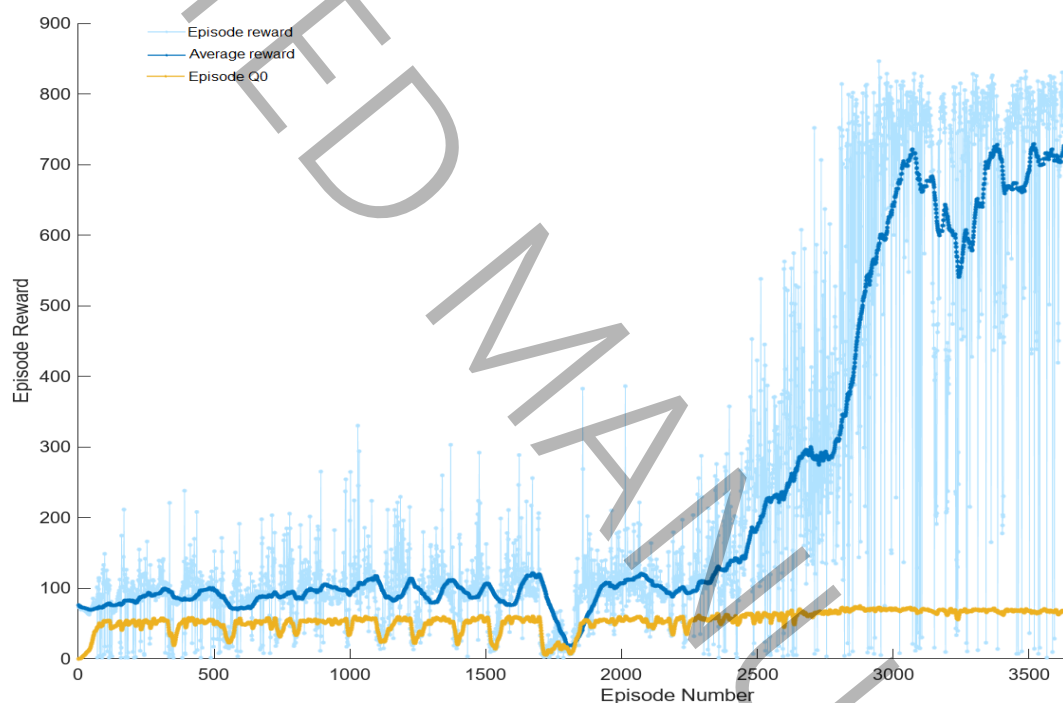
$$r = 0.6r_{pos} + 0.1r_{vel} + 0.1r_{acc} + 0.2r_{\psi} \quad (14)$$

۳-۳- فرآیند آموزش

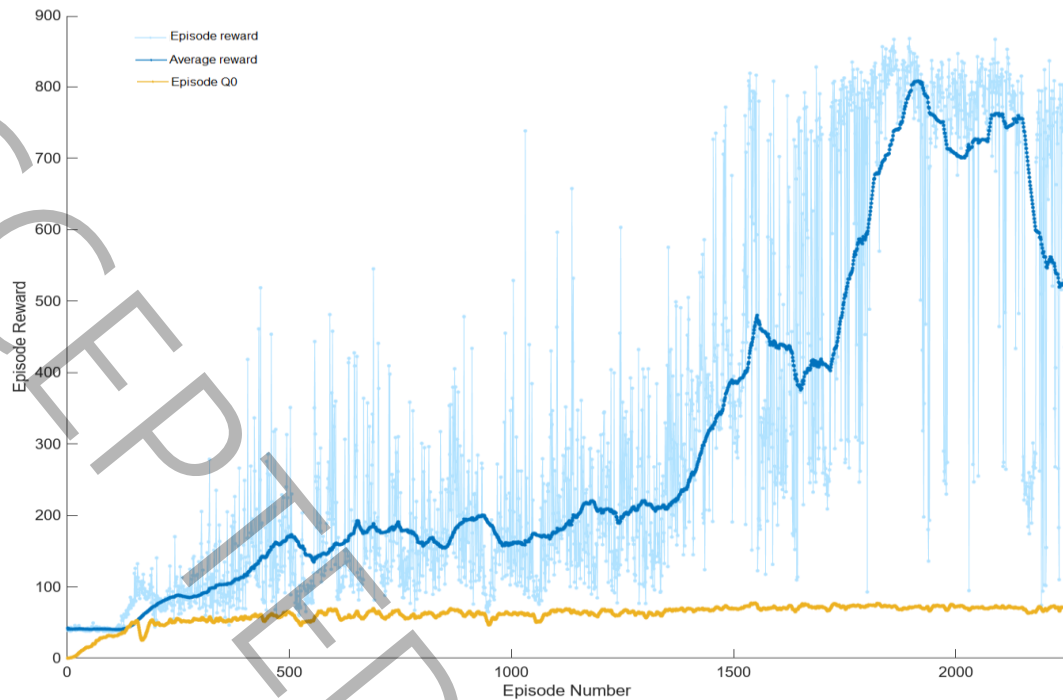
برای تسهیل آموزش مؤثر سیاست کنترل، مسیرهای مرجع با در نظر گرفتن محدودیت‌های فیزیکی ذاتی سیستم‌های کوادروتور، مانند محدودیت‌های سرعت و پایداری پرواز، استخراج شده‌اند. این مسیرها به‌عنوان راهنمای ساختاریافته برای کوادروتور عمل کردند و به آن امکان می‌دهند تا با کمینه کردن خطاهای دنبال کردن مسیر، به سمت موقعیت‌های هدف تعیین‌شده حرکت کند و سرعت را به‌صورت کنترل‌شده و بصورت واقع‌گرایانه تنظیم نماید. مسئله یادگیری در قالب یادگیری تقویتی اپیزودیک مدل‌سازی شده است به طوری که هر یک سیکل شبیه‌سازی به عنوان یک اپیزود مستقل در نظر گرفته شده است. اپیزود در یکی از دو حالت پایان می‌یابد: (۱) رسیدن موفقیت‌آمیز به هدف (۲) خاتمه زودهنگام^۱ (ET) که در صورت انحراف قابل توجه از مسیر مرجع یا از دست دادن پایداری پرواز رخ می‌دهد. در صورت عدم تحقق این شرایط، اپیزود برای مدت زمان معین شبیه‌سازی ادامه می‌یابد تا عامل بتواند کل مانور را کامل کند. طراحی خاتمه زودهنگام اطمینان حاصل می‌کند که عامل با تولید خروجی‌های کنترلی مناسب برای هر حالت تعامل داشته باشد، که منجر به کاهش کاوش در مناطق غیرسودمند برای فضای حالت شده و به تسریع همگرایی سیاست کمک می‌کند.

¹ Early Termination

در این مقاله برای هر اپیزود آموزشی ۱۰ ثانیه شبیه‌سازی در نظر گرفته شده است و شامل ۱۰۰۰ گام زمانی گسسته می‌باشد. پاداش تجمعی در طول آموزش روند افزایشی مداومی را نشان می‌دهد که نشانه توانایی یادگیری سیاست و بهبود عملکرد آموزشی در طول زمان است. در مقایسه با مطالعات پیشین کنترل کوادروتور با استفاده از یادگیری تقویتی عمیق مانند هوانگبو و همکاران (حدود ۲,۱۵ میلیون گام زمانی) [۲۹]، شریفی و همکاران (حدود ۱,۲ میلیون گام زمانی) [۳۰] و پی و همکاران (۱۰ میلیون گام زمانی) [۳۱] روش پیشنهادی بهبود چشمگیری در بهره‌وری نمونه‌ها نشان می‌دهد. برای ارزیابی اثربخشی کنترلر پیشنهادی یادگیری تقویتی عمیق چندلایه، عملکرد آن در برابر یادگیری تقویتی عمیق مرسوم سنجیده شده است. در گام اول جهت مقایسه رفتار الگوریتم‌های پیشنهادی، الگوریتم یادگیری تقویتی عمیق (به عنوان الگوریتم پایه) شبیه‌سازی شده و به بررسی روند آموزشی آن پرداخته شده است. با استفاده از این الگوریتم، عامل پس از ۲۹۴۸ اپیزود به پاداش بهینه دست پیدا می‌کند (شکل ۵). درگام بعد مشاهده می‌شود که عامل با استفاده از الگوریتم یادگیری تقویتی عمیق عاطفی پیشنهادی پس از ۱۸۶۹ اپیزود به پاداش بهینه دست می‌یابد و یاد می‌گیرد که مسیرمطلوب را طی کند (شکل ۶). این کاهش قابل توجه در تعداد اپیزودهای آموزشی، خواص همگرایی برتر الگوریتم کنترلر پیشنهادی را نشان می‌دهد که بازتاب‌دهنده افزایش کارایی یادگیری و توانایی دستیابی به سیاست‌های بهینه با تعاملات کمتر است که با به‌کارگیری شبکه القای پیشرفت آموزش به میزان ۳۷ درصد نسبت به مدل پایه تسریع می‌شود.



شکل ۵: پاداش هر اپیزود با چارچوب یادگیری تقویتی عمیق مرسوم



شکل ۶: پاداش هر اپیزود با چارچوب یادگیری تقویتی عمیق عاطفی

۴- شبیه‌سازی و نتایج تجربی

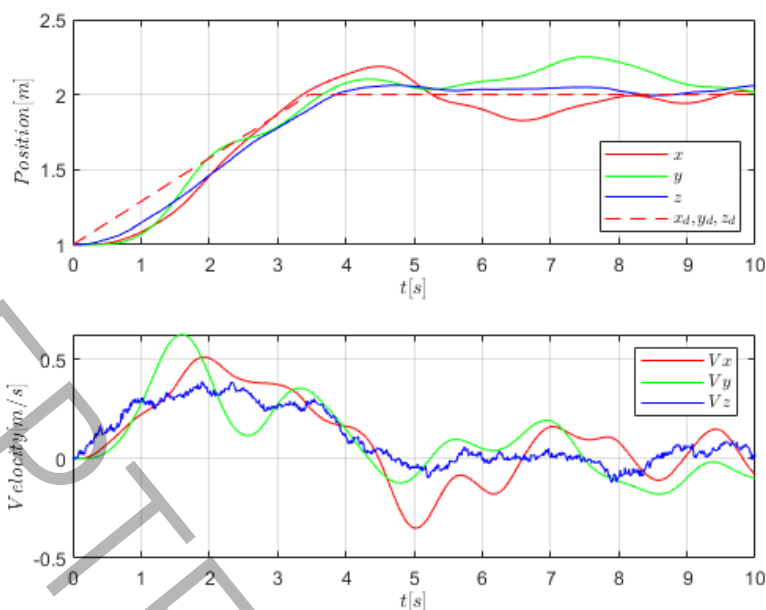
در این بخش به بررسی نتایج شبیه‌سازی کنترلر معرفی شده بر مدل دینامیکی کوادروتور پرداخته خواهد شد.

۴-۴- شبیه‌سازی کنترل‌کننده پیشنهادی

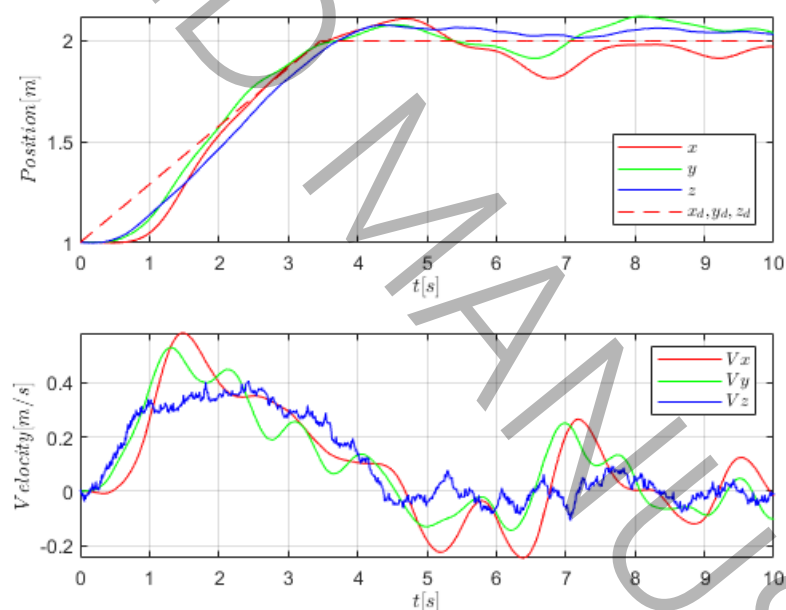
در این مقاله چارچوب یادگیری تقویتی عمیق عاطفی با ادغام شبکه Inception، برای وظیفه دنبال کردن مسیر کوادروتور به کار گرفته شد؛ جایی که تطبیق‌پذیری با محیط‌های پویا و نامطمئن یک الزام حیاتی است. پس از کامل شدن آموزش روش‌های پیشنهادی برای ارزیابی عملکرد کنترل‌کننده، آزمایش‌های شبیه‌سازی با استفاده از نرم‌افزار MATLAB انجام شده است که نتایج شبیه‌سازی در شکل ۷ و شکل ۸ در مسیر مطلوب ($x_d = y_d = z_d$, $x_d = (t/3.5) + x_0$, $x_d - x_0 \leq 1$) نشان داده شده است. در این شکل‌ها نمودار سرعت نشان‌دهنده تغییرات سرعت ربات در تعقیب مسیر می‌باشد. جدول ۱ شاخص‌های عملکرد الگوریتم‌های پیشنهادی را خلاصه کرده و بهبودهایی از نظر کاهش تعداد اپیزودهای آموزشی، کاهش خطای مسیر را برجسته می‌سازد. به طور کلی، رویکرد الگوریتم یادگیری تقویتی عمیق عاطفی پیشنهادی، کارایی یادگیری و تطبیق‌پذیری خوبی از خود نشان داده که آن را برای وظایف ناوبری کوادروتور مناسب می‌سازد. استفاده از این الگوریتم به طور قابل توجهی تعداد پارامترهای قابل آموزش را کاهش داد که منجر به همگرایی سریع‌تر در فرآیند یادگیری شده است.

جدول ۱: عملکرد پاسخ الگوریتم‌های یادگیری تقویتی عمیق

نوع الگوریتم	Rise Time	Settling Time	Overshoot	Steady State Error
یادگیری تقویتی عمیق عاطفی	۳,۱	۵	۰,۱۵	۰,۱
یادگیری تقویتی عمیق	۳,۳	۵,۲	۰,۱	۰,۱

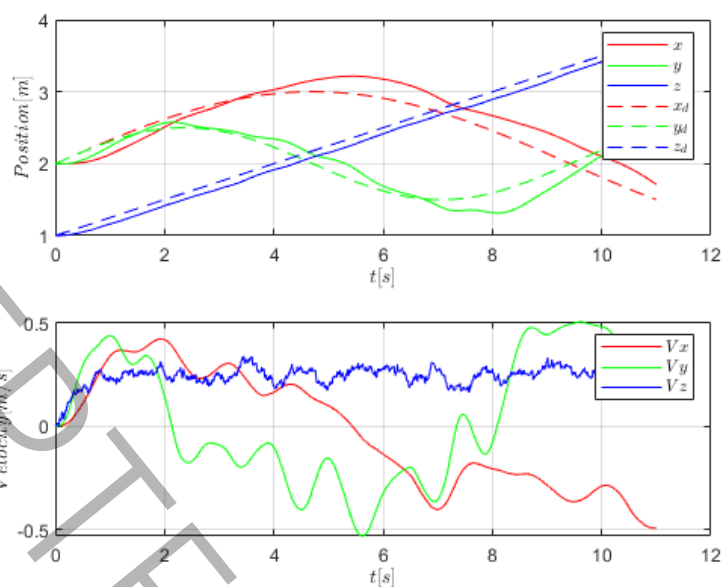


شکل ۷: دنبال کردن مسیر با استفاده از یادگیری تقویتی عمیق

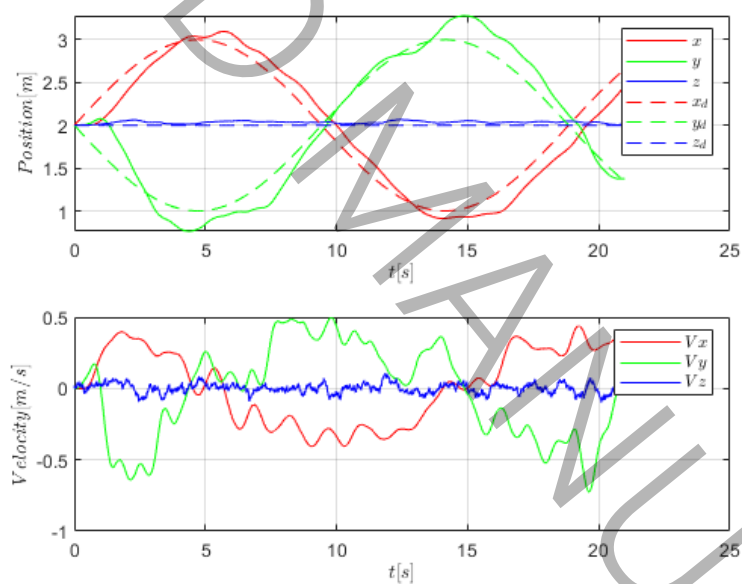


شکل ۸: دنبال کردن مسیر با استفاده از یادگیری تقویتی عمیق عاطفی

با توجه به نتایج شبیه‌سازی مشخص می‌شود که کنترلرهای ارائه شده به خوبی توانسته‌اند مسیر موردنظر را طی کنند. درحالی‌که زمان آموزش کمتری برای کنترلر طراحی شده یادگیری تقویتی عمیق عاطفی صرف شده است، کنترلر پیشنهادی عملکرد خوبی در کنترل ربات در مسیر موردنظر داشته‌است. برای بررسی بیشتر عملکرد الگوریتم یادگیری تقویتی عمیق عاطفی، در ادامه پاسخ کنترلر الگوریتم یادگیری تقویتی عمیق عاطفی بر مبنای حالت عاطفی به مانورهای مختلف مورد بررسی قرار گرفته است (شکل ۹ و شکل ۱۰) که نشان از توانمندی کوادروتور در تعقیب مسیر می‌باشد.



شکل ۹: ردیابی مسیر هلیکال با استفاده از یادگیری تقویتی عمیق عاطفی بر مبنای حالت عاطفی



شکل ۱۰: ردیابی مسیر سینوسی با استفاده از یادگیری تقویتی عمیق عاطفی بر مبنای حالت عاطفی

۵- اثبات پایداری

جهت اثبات پایداری به روش لیاپانوف و با در نظر گرفتن مدل دینامیکی کوادروتور در راستای Z به عنوان نمونه داریم

$$\ddot{z} = \frac{u_1}{m} (\cos \phi \cos \eta) - g \quad (15)$$

برای کنترل ارتفاع، ورودی کنترلی بصورت زیر تعریف شده است.

$$u_1 = \frac{mg + F_z}{\cos \phi \cos \eta} \quad (16)$$

با جایگذاری در دینامیک مدل خواهیم داشت

$$\ddot{z} = \frac{1}{m} F_z \quad (17)$$

با بازنویسی معادله قبل بصورت فضای حالت خواهیم داشت

$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = \frac{F_z}{m} \end{cases} \quad (18)$$

از طرفی با در نظر گرفتن ورودی کنترلی عملگر به صورت $F_z = \alpha_1 \tanh(\zeta_1 z_1 + \zeta_2 z_2)$ برای کنترل کننده یادگیری تقویتی عمیق به بررسی پایداری کنترل کننده پرداخته خواهد شد. به منظور اثبات پایداری کنترل کننده، نامزد تابع لیاپانوف (V) به صورت رابطه زیر در نظر گرفته می شود.

$$V = -\frac{\zeta_1 m}{2\alpha} z_2^2 + \ln(\cosh(\zeta_1 z_1 + \zeta_2 z_2)) \quad (19)$$

شرط پذیرش این تابع به عنوان تابع V به صورت $\zeta_1 < 0$ می باشد.

$$\text{if } \zeta_1 < 0 \Rightarrow V > 0 \quad (20)$$

مشتق این تابع به صورت زیر قابل محاسبه می باشد.

$$\dot{V} = -\frac{\zeta_1 m}{\alpha} z_2 \dot{z}_2 + \frac{(\zeta_1 \dot{z}_1 + \zeta_2 \dot{z}_2) \sinh(\zeta_1 z_1 + \zeta_2 z_2)}{\cosh(\zeta_1 z_1 + \zeta_2 z_2)} \quad (21)$$

با جایگذاری خواهیم داشت

$$\begin{aligned} \dot{V} &= -\zeta_1 z_2 \tanh(\zeta_1 z_1 + \zeta_2 z_2) \\ &+ \left(\zeta_1 z_2 + \frac{\alpha \zeta_2}{m} \tanh(\zeta_1 z_1 + \zeta_2 z_2) \right) \tanh(\zeta_1 z_1 + \zeta_2 z_2) = \frac{\alpha \zeta_2}{m} \tanh^2(\zeta_1 z_1 + \zeta_2 z_2) \end{aligned} \quad (22)$$

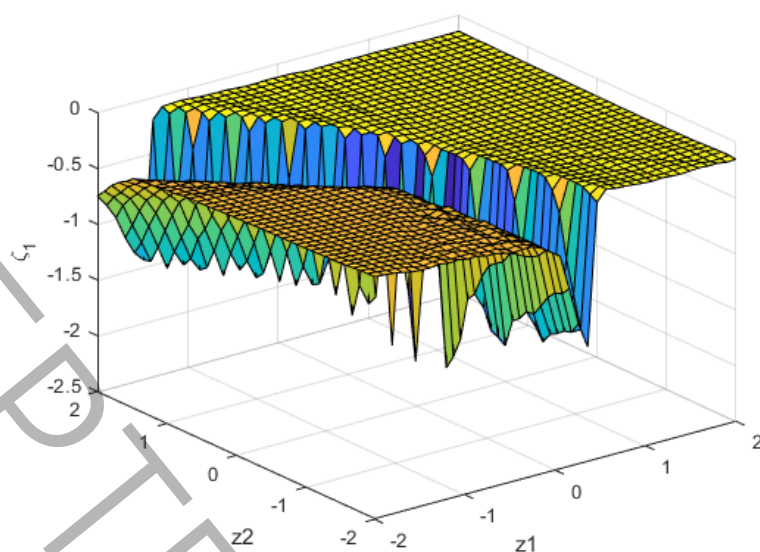
بنابراین شرط منفی معین شدن $\dot{V}$ به صورت $\zeta_2 < 0$ خواهد بود.

$$\text{If } \zeta_2 < 0 \Rightarrow \dot{V} < 0 \quad (23)$$

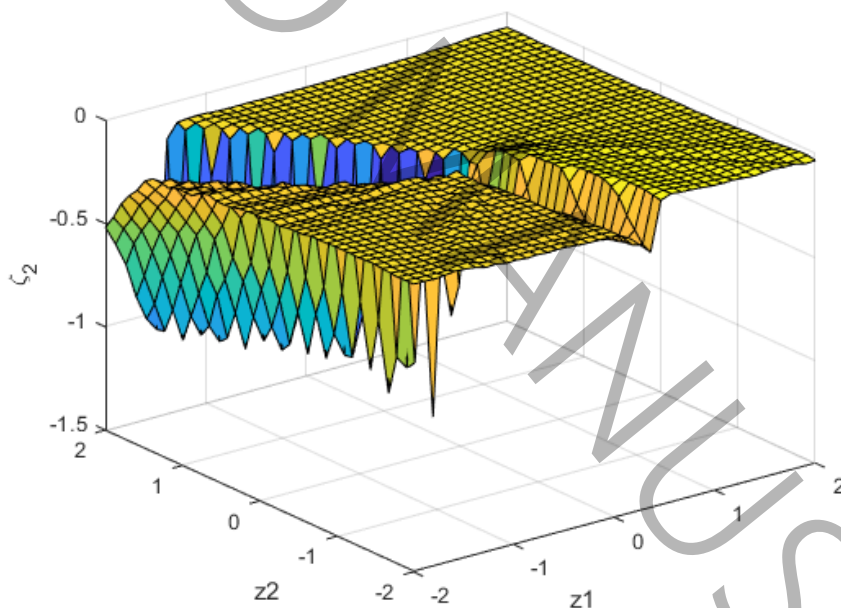
درنتیجه شرط لازم برای اثبات پایداری مطابق زیر می باشد.

$$\zeta_1 < 0 \ \& \ \zeta_2 < 0 \Rightarrow V > 0 \ \& \ \dot{V} < 0 \quad (24)$$

شکل ۱۱ و شکل ۱۲ مقادیر پارامترهای ζ_1 و ζ_2 را در شرایط مختلف ورودی ها نمایش می دهد. باتوجه به منفی بودن این ضرایب، بنابراین پایداری کنترل کننده قابل اثبات می باشد.



شکل ۱۱: مقادیر پارامترهای ۱



شکل ۱۲: مقادیر پارامترهای ۲

۶- نتیجه گیری

در این پژوهش، یک چارچوب کنترل کننده یادگیری تقویتی عمیق عاطفی مبتنی بر شبکه Inception برای کنترل کوادروتور ارائه شده است. این ساختار با الهام از دو مسیر منطقی و هیجانی مغز انسان طراحی شده، به طوری که مسیر منطقی توسط شبکه PPO و مسیر هیجانی بر مبنای شبکه جهت استخراج ویژگی های لایه های میانی با ایجاد مسیر کوتاه توسط شبکه Inception مدل سازی شده است. نتایج شبیه سازی نشان می دهد که کنترل کننده پیشنهادی سرعت آموزش بیشتری نسبت به روش های DRL مرسوم دارد. با توجه به نتایج شبیه سازی، الگوریتم یادگیری تقویتی عمیق عاطفی پیشنهادی ۳۷ درصد سریعتر به سیاست بهینه رسیده است. در

مجموع، رویکرد پیشنهادی با ترکیب هوش هیجانی و یادگیری تقویتی، گامی مؤثر در جهت توسعه کنترل‌کننده‌های هوشمند با رفتار سازگار و تصمیم‌گیری سریع را فراهم می‌آورد.

۷- منابع و مراجع

- [1] B. Jang, M. Kim, G. Harerimana, J.W. Kim, Q-learning algorithms: A comprehensive classification and applications, IEEE access, 7 (2019) 133653-133667.
- [2] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, Human-level control through deep reinforcement learning, nature, 518(7540) (2015) 529-533.
- [3] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in: Proceedings of the AAAI conference on artificial intelligence, 2016.
- [4] P. Norouzi, H. Shahbazi, K. Torabi, Comparative analysis of a multi-layer deep reinforcement learning controller and an optimized neural network controller utilizing growing particle swarm optimization for quadrotor control, International Journal of Machine Learning and Cybernetics, 16(11) (2025) 9007-9033.
- [5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347, (2017).
- [6] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in: International conference on machine learning, PMLR, 2015, pp. 1889-1897.
- [7] P. Norouzi, H. Shahbazi, K. Torabi, Optimizing quadrotor navigation through emotional deep reinforcement learning: leveraging emotional rewards and states for enhanced training efficiency, Journal of the Brazilian Society of Mechanical Sciences and Engineering, 48(2) (2026) 115.
- [8] Z. Jiang, A.F. Lynch, Quadrotor motion control using deep reinforcement learning, Journal of Unmanned Vehicle Systems, 9(4) (2021) 234-251.
- [9] J. Morén, C. Balkenius, A computational model of emotional learning in the amygdala, From animals to animats, 6 (2000) 115-124.
- [10] C. Lucas, D. Shahmirzadi, N. Sheikholeslami, Introducing BELBIC: brain emotional learning based intelligent controller, Intelligent Automation & Soft Computing, 10(1) (2004) 11-21.
- [11] H. Rouhani, M. Jalili, B.N. Araabi, W. Eppler, C. Lucas, Brain emotional learning based intelligent controller applied to neurofuzzy model of micro-heat exchanger, Expert Systems with Applications, 32(3) (2007) 911-918.
- [12] M.A. Rahman, R.M. Milasi, C. Lucas, B.N. Araabi, T.S. Radwan, Implementation of emotional controller for interior permanent-magnet synchronous motor drive, IEEE Transactions on Industry Applications, 44(5) (2008) 1466-1476.
- [13] S.A. Nahian, D.Q. Truong, K.K. Ahn, A self-tuning brain emotional learning-based intelligent controller for trajectory tracking of electrohydraulic actuator, Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 228(7) (2014) 461-475.
- [14] P. Norouzi, H. Shahbazi, K. Torabi, A novel deep reinforcement learning framework with human activation function to control quadrotor: P. Norouzi et al, International Journal of Dynamics and Control, 14(2) (2026) 69.
- [15] G. Pan, H. Matsuhisa, Y. Honda, Analytical model of a magnetorheological damper and its application to the vibration control, in: 2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies, IEEE, 2000, pp. 1850-1855.
- [16] B. Dalvi, Modelling and Testing of MR Dampers, Master's Thesis, Coventry University, Coventry, UK, (2015).

- [17] G. Yang, Large-scale magnetorheological fluid damper for vibration mitigation: modeling, testing and control, University of Notre Dame, 2002.
- [18] M.B. César, J. Gonçalves, J. Coelho, R.C. de Barros, Brain emotional learning based control of a SDOF structural system with a MR damper, in: *CONTROLO 2016: Proceedings of the 12th Portuguese Conference on Automatic Control*, Springer, 2017, pp. 547-557.
- [19] J.P. Coelho, M. Braz-César, J. Gonçalves, BELBIC Based Step-Down Controller Design Using PSO, in: *International Conference on Optimization, Learning Algorithms and Applications*, Springer, 2021, pp. 345-356.
- [20] J. Li, T. Yu, A novel data-driven controller for solid oxide fuel cell via deep reinforcement learning, *Journal of Cleaner Production*, 321 (2021) 128929.
- [21] E. Lotfi, S. Setayeshi, S. Taimory, A neural basis computational model of emotional brain for online visual object recognition, *Applied Artificial Intelligence*, 28(8) (2014) 814-834.
- [22] E. Lotfi, M.-R. Akbarzadeh-T, Adaptive brain emotional decayed learning for online prediction of geomagnetic activity indices, *Neurocomputing*, 126 (2014) 188-196.
- [23] E. Lotfi, Mathematical modeling of emotional brain for classification problems, *Proceedings of IAM*, 2(1) (2013) 60-71.
- [24] R.P. Marinier, J.E. Laird, Emotion-driven reinforcement learning, in: *Proceedings of the annual meeting of the cognitive science society*, 2008.
- [25] N.G. Aghaei, H. Shahbazi, P. Farzaneh, A. Abdolmaleki, A. Khorsandian, The structure of personality-based emotional decision making in robotic rescue agent, in: *2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, IEEE, 2008, pp. 1201-1206.
- [26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
- [27] E. Saif, İ. Eminoğlu, Modelling of quad-rotor dynamics and Hardware-in-the-Loop simulation, *The Journal of Engineering*, 2022(10) (2022) 937-950.
- [28] K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath, A brief survey of deep reinforcement learning, arXiv preprint arXiv:1708.05866, (2017).
- [29] J. Hwangbo, I. Sa, R. Siegwart, M. Hutter, Control of a quadrotor with reinforcement learning, *IEEE Robotics and Automation Letters*, 2(4) (2017) 2096-2103.
- [30] D. Sharifi, K. Torabi, M.I. Rahaghi, H. Shahbazi, Designing a reinforcement learning-based robust controller resisting external forces for a octarotor with new structure, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 45(3) (2023) 141.
- [31] C.-H. Pi, K.-C. Hu, S. Cheng, I.-C. Wu, Low-level autonomous control and tracking of quadrotor using reinforcement learning, *Control Engineering Practice*, 95 (2020) 104222.

ACCEPTED MANUSCRIPT