Optimizing Reward Function Weights and Enhancing Control Mechanisms for Bipedal Robots Using LSTM and Attention Mechanisms

Lingzhi Cui , Tianqi Deng , Lihua MA , Wenhao He*

Institute of Automation, Chinese Academy of Sciences;
School of Artificial Intelligence, University of Chinese Academy of Sciences;
Cuilingzhi2021@ia.ac.cn

Abstract. This paper introduces an optimized control approach for bipedal robots, merging Bayesian optimization for reward function weights and a novel neural network structure combining LSTM and Transformer-based attention. Bayesian optimization enhances training stability and efficiency, while the hybrid network captures temporal patterns and long-range dependencies, outperforming traditional architectures in reward stability and performance. Simulated evaluations show our model's superior robustness against challenges like varying ground friction and external disturbances. Future work will focus on domain randomization and real-world robot fine-tuning to bridge the simulation-reality gap.

Keywords: Bipedal Robots, Bayesian Optimization, LSTM, Transformer-based Attention, Neural Network Architecture, Robustness, Simulated Evaluation, Domain Randomization, Reward Function Optimization..

1 INTRODUCTION

The control of bipedal robot gait presents highly complex dynamic nonlinear characteristics, posing significant challenges. Traditional control strategies often restrict themselves to local linearized dynamic analysis and reduced-order abstractions to produce feasible solutions. However, these methods often overlook many key details such as torque limitations, joint constraints, and other nonlinearities that must be excluded from control computations. Deep Reinforcement Learning (DRL) offers a promising model-free approach, fully exploiting the dynamic nature of bipedal robots.

In engineering practice, Proximal Policy Optimization (PPO) has become a mature, stable, and reliable mainstream reinforcement learning algorithm, widely used in addressing bipedal robot gait problems. The key to the PPO algorithm lies in the design of its reward function, particularly the determination of weights such as joint error, centroid error, spring error, and orientation error. A well-established baseline can make training more efficient, while directly copying existing parameters may lead to

* Corresponding author

extremely poor training results (slow convergence, low final convergence), especially when there are differences in the trunk length, material, motor, spring material, and load between our robot and the original Cassie version. This method in our work can be applied to other similar strong baseline search tasks. To overcome this challenge, this paper employs the Bayesian optimization algorithm's Expected Improvement (EI) strategy to perform optimal parameter search, identifying the optimal parameters for the reward function.

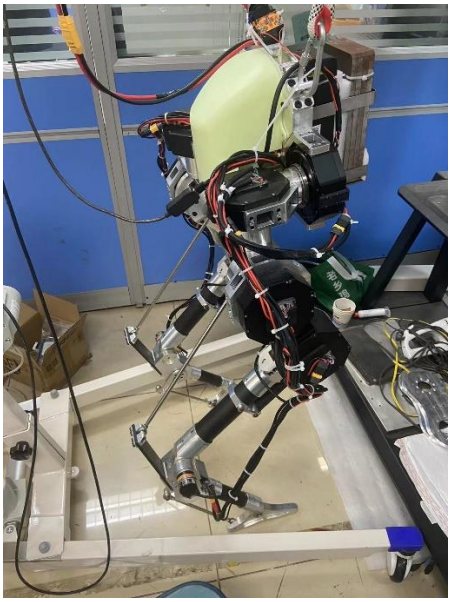

Fig. 1. The biped robot equipment that has been built in the laboratory.

The primary contributions of this paper focus on two equally significant aspects: firstly, the proposal of a low-cost reinforcement learning reward weight parameter optimization method; secondly, an in-depth exploration in the design of the actor network within PPO, innovatively proposing an efficient training strategy network of LSTM with attention mechanism. Through a comparison of feed-forward networks, feed-forward networks with attention mechanisms, LSTM networks, and LSTM networks with attention mechanisms in both training and simulation experiment scenarios, we conclude that the combination of LSTM and attention mechanism is an optimal solution, and we provide a thorough discussion of the underlying advantages of this network. Figure 1 shows the biped robot platform completed by our laboratory for actual machine verification, which is equipped with our own customized motor featuring high torque and high precision.

2 RELATED WORK

In the realm of bipedal robots and their control mechanisms, the dynamism of technological advancements has consistently posed challenges and opportunities. These robots, characterized by their intricate designs and nuances in motion patterns, have witnessed a series of evolutions in their control methodologies. The field has intricately woven neural networks, optimization techniques, and sophisticated neural network structures to bridge the gap between theoretical prowess and practical efficiency. This section delves into the pivotal works that have laid the foundation and influenced our study.

2.1 Controlling Bipedal Robots Using Neural Networks

The control of bipedal robots using neural networks has been widely explored and applied. For instance, Zhaoming Xie, Glen Berseth, et al. demonstrated the use of deep reinforcement learning (DRL) for feedback control of the Cassie bipedal robot developed by Agility Robotics [1]. W. Thomas Miller, III delved into the use of cerebellar model arithmetic computers (CMAC) neural networks to address the dynamic balance challenges of bipedal walking [2]. Chuanyu Yang, Kai Yuan, Shuai Heng, and others proposed a novel learning framework that combines imitation learning, deep reinforcement learning, and control theory to achieve natural, dynamic, and robust human-styled walking for humanoid robots [3]. Furthermore, Tianyu Li, Hartmut Geyer, Christopher G. Atkeson, and Akshara Rai explored the application of deep reinforcement learning in high-fidelity simulators for learning control strategies and transferring them to real bipedal robots [4]. These studies underscore the significance of neural networks in addressing nonlinear and dynamic challenges in bipedal robot control.

2.2 Proximal Policy Optimization (PPO) in Robotics

Proximal Policy Optimization (PPO) has become the preferred algorithm in reinforcement learning, with its stability and efficiency leading to successful applications in many robotic tasks. PPO, introduced by Schulman et al. [5], enhances training robustness and stability by utilizing trust region methods and truncated reweighted objectives. The algorithm optimizes the objective function and ensures that policy updates do not lead to significant performance degradation. By balancing exploration and exploitation, PPO provides effective learning paths for various robotic tasks, including continuous control and complex decision-making problems.

Specifically, in the field of bipedal robot gait control, the PPO algorithm has been successfully applied in complex dynamic environments, achieving feasible real-time control strategies. Melo and Maximo [6] utilized PPO to learn humanoid robot running skills in the Soccer 3D environment. Their methodology, based on Deep Reinforcement Learning, learned running skills without any prior knowledge, using a neural network focused on robot dynamics. Their results significantly outperformed the previous state-of-the-art sprint velocities in the Soccer 3D domain, demonstrating improvements in sample efficiency and the ability to learn running skills in just a few hours.

However, the exploration of different actor network designs and the optimization of specific reward function parameters for bipedal robot control remain relatively unexplored. This research delves into these aspects, revealing novel contributions in actor network design and reward function optimization.

2.3 Reward Parameter Tuning in Reinforcement Learning

In Reinforcement Learning (RL), tuning reward parameters is crucial for the success of the learning process. Traditional methods often rely on tedious manual tuning or grid search, lacking efficiency and robustness. Recent years have seen the introduction of optimization techniques like Bayesian Optimization to automate the tuning process, showing promising results across various domains. Bayesian Optimization globally optimizes by constructing a probabilistic model of the objective function and uses specific acquisition functions (such as Expected Improvement) to balance exploration and exploitation. Particularly in complex hyperparameter spaces, such as reward weight selection in deep neural networks, Bayesian Optimization has proven to be a powerful tool.

Jia Wu et al. [7] emphasized the importance of hyperparameters in machine learning algorithms and proposed an efficient hyperparameter optimization algorithm based on Bayesian optimization. They utilized Gaussian processes to establish a relationship between the performance of machine learning models and their hyperparameters, turning the hyperparameter tuning problem into an optimization problem. Their experiments demonstrated the effectiveness of their method in optimizing widely used machine learning models.
Vu Nguyen [8] highlighted the flexibility and power of Bayesian optimization (BO) for hyper-parameter tuning and global optimization of expensive black box functions. His research summarized the recent advances in Bayesian optimization methodologies and their applications.
Aaron Klein et al. [9] introduced a generative model for the validation error as a function of training set size, which is learned during the optimization process. Their Bayesian optimization procedure, FABOLAS, models loss and training time as a function of dataset size and automatically trades off high information gain about the global optimum against computational cost. Their experiments showed that FABOLAS significantly outperforms other state-of-the-art Bayesian optimization methods in terms of speed.
However, the application of these methods to the unique challenges of bipedal robot control remains an open question. This study advances this line of research by employing Bayesian Optimization with the Expected Improvement (EI) strategy, targeting specific reward weights for bipedal robot control..

2.4 Neural Network Structure Improvement: Transformer, LSTM, and LSTM+Transformer

Traditional neural network architectures may still face limitations when handling temporal dependencies in sequential tasks. For instance, even though feed-forward neural networks and LSTM networks have been extensively applied to the gait control of bipedal robots, they might still encounter challenges with complex temporal dependencies and dynamic nonlinear problems.

LSTM, with its long short-term memory units, captures long-term dependencies in time series and has been validated and applied in multiple robot control projects [12,13,14]. Attention mechanisms have been employed across various domains, especially in Natural Language Processing (NLP) , enhancing the performance of models handling sequential data by focusing on pertinent parts of the input. The Transformer architecture, introduced by Vaswani et al. [10], captures long-distance dependencies through self-attention, offering a fresh perspective for sequential tasks. Notably, Bednarek et al. [11] utilized the Transformer architecture for haptic terrain classification in legged robots, demonstrating its efficacy in this domain.

In recent years, the combined architecture of LSTM and Transformer has been validated in text classification tasks, showcasing promising performance [15,16,17]. This combined approach has been widely validated and applied across various domains.

In this study, for the first time, we apply the combined approach of LSTM and Transformer to bipedal robot control. This integration not only inherits the strengths of both but also provides a novel perspective for bipedal robot control, marking our primary innovation.

3 METHOD

3.1 Introduction to Bayesian Optimization

Bayesian optimization is a global optimization method based on probabilistic models, suitable for problems with high-dimensional parameter spaces and costly evaluation functions. The foundational concept here is the Gaussian Process (GP). A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. It is fully defined by its mean function $m(x)$ and covariance function $k(x, x')$. A Gaussian process is denoted as:

$$f(x) \sim \mathcal{GP}\big(m(x), k(x, x')\big) \tag{1}$$

Where $m(x)$ is the mean function., $k(x, x')$ is the covariance function or kernel.

The method primarily employs Gaussian process regression to establish the probabilistic model of the target function. Given a set of observations with associated noisy function values $D = \{(xi, yi)\}$, the posterior over functions can be computed using the Bayes theorem.

The next step is to utilize specific acquisition functions to determine where to sample next. One of the most popular acquisition functions is the Expected Improvement (EI) which is given by:

$$EI(x) = \left(\mu(x) - f(x^+ - \xi)\Phi(Z) + \sigma(x)\phi(Z)\right) \quad (2)$$

Where:

- $\mu(x)$is the predicted mean at point x.
- $\sigma(x)$is the predicted standard deviation.
- $f(x^+)$is the best observed value so far.

- Φ and ϕ are the CDF and PDF of the standard normal distribution respectively.
- $Z = \frac{\mu(x) - f(x^+) - \xi}{\sigma(x)}$ if $\sigma(x) > 0$ and 0 otherwise.
- ξ is a small positive value to ensure exploration.

The Expected Improvement strategy aims to choose the next evaluation point by maximizing the expected improvement over the current best solution.

Application of Bayesian Optimization. In our work, Bayesian optimization was employed to adjust the weight parameters of the reward function.

Specifically, we used the Expected Improvement strategy as our acquisition function and defined appropriate bounds for each parameter. Through multiple iterations, the algorithm continuously selects new points in the parameter space for evaluation, thereby identifying the optimal reward function parameters.

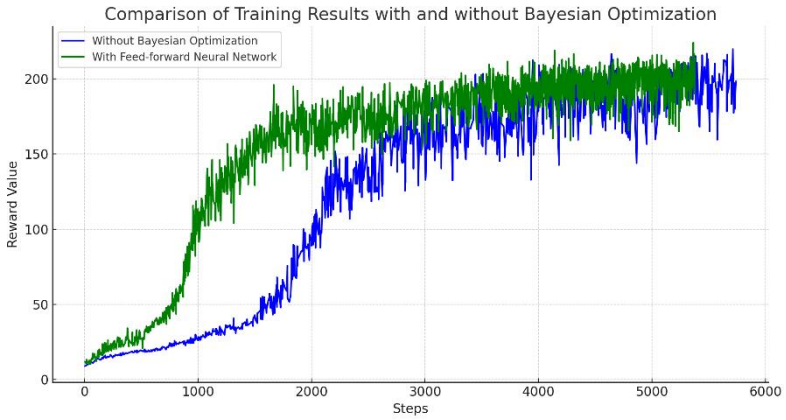

Fig. 2. Comparison of Reward Trajectories: Feedforward Neural Network vs. Model Before Bayesian Optimization

Experimental Results. As evident from the Fig, 2, the reward function parameters refined through Bayesian optimization exhibited higher stability and efficacy throughout the training process. Specifically:

Initial Convergence Speed: The growth rate of rewards in the initial few thousand steps is visibly faster with the optimized parameters compared to the original parameters. This implies that the reward parameters selected via Bayesian optimization can guide the agent to find an effective strategy more rapidly.

Stability: After approximately 5,000 steps, while the reward trends for both exhibit rises and falls, the reward curve of the optimized parameters is relatively more stable with minor fluctuations. This indicates that the optimized parameters offer greater robustness during the training process.

Final Reward: In the later stages of training, post the 20,000-step mark, the rewards obtained using the optimized parameters are distinctly higher than those achieved using the original parameters. This further validates that the reward parameters chosen by Bayesian optimization assist the agent in attaining superior final performance.

In conclusion, the reward function parameters adjusted through Bayesian optimization not only accelerated the learning process but also enhanced the stability and final performance of the training. These results underscore the efficacy and value of Bayesian optimization in automatically tuning hyperparameters in deep reinforcement learning tasks.

3.2 Network Architecture Design and Experiments

In recent robotic control tasks, the choice of neural network architecture plays a pivotal role in determining the performance and efficiency of the control policy. Our study focuses on evaluating and introducing advanced neural network architectures, especially incorporating attention mechanisms, into the design of the actor network in the Proximal Policy Optimization (PPO) algorithm.

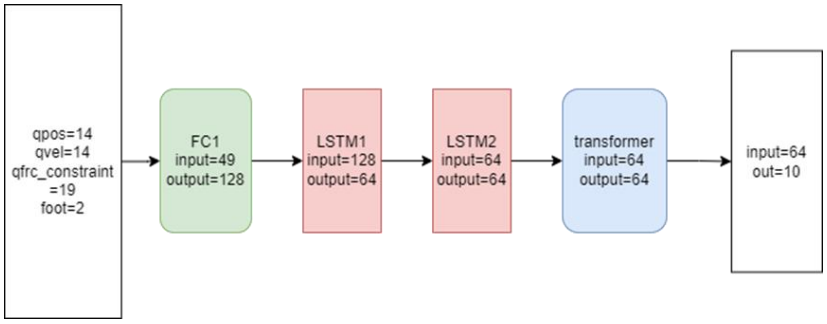

Fig. 3. Proposed Hybrid Network Architecture Combining LSTM and Transformer for Bipedal Robot Control.

Network Architecture Design. Our proposed network architecture starts by processing the input state features through a fully connected layer. It then employs an LSTM layer to capture the temporal dependencies inherent in the sequential data. . Figure3 showcases our network architecture, detailing each layer and its connections. Following this, a Transformer layer is integrated to manage long-distance dependencies, leveraging the self-attention mechanism. The output layer subsequently generates a policy for each possible action and estimates the value of the current state.
The attention mechanism, inspired by its success in Natural Language Processing (NLP), has shown promise in capturing essential features from input data. In our model, we incorporate the attention mechanism after each LSTM layer, enabling the network to focus on specific temporal features that are more relevant to the current state.
This hybrid architecture, combining both LSTM and Transformer, aims to harness the strengths of both models: the LSTM's ability to remember short-term dependencies and the Transformer's capability to manage long-distance dependencies.:

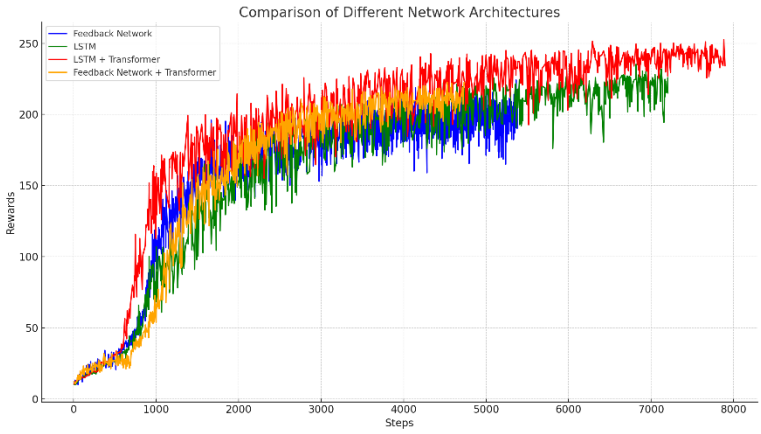

Fig. 4. Reward Trajectories Across Different Neural Network Architectures and Optimization Techniques

Experiments and Results. To validate the effectiveness of our proposed architectures, we carried out experiments comparing the performance of various models, as illustrated in Figure 5.

Feedback Network: The performance of the baseline feedforward network reached a reward of around 230 after 500k steps but showed considerable oscillations throughout the training process.

Feedback Network with Attention: With the integration of the attention mechanism, the reward increased to about 260 after the same number of steps, indicating the potential

of attention in capturing crucial features. However, oscillations persisted, although slightly reduced.

LSTM Network: Implementing the LSTM network resulted in a stable increase in rewards, reaching up to 300 after 500k steps with reduced oscillations, emphasizing the advantage of capturing short-term dependencies.

LSTM with Attention Mechanism: Combining LSTM with attention, the reward further surged to nearly 325 after 500k steps, with minimal oscillations. This consolidates the assertion that the attention mechanism, combined with LSTM's temporal feature-capturing ability, can significantly *enhance* the model's performance.

4 Experimental Results and Simulation Evaluation

In the experimental phase, we put our network models to the test to gauge their effectiveness in handling the control tasks of the Cassie bipedal robot. Various architectures were assessed: a straightforward feedforward network, feedforward with attention mechanism, LSTM, and the proposed LSTM combined with a transformer. The training process for each of these models was visualized in terms of reward metrics over training steps.

4.1 Training Performance

The feedforward network provided a baseline for our experiment, achieving a reward of approximately 160 after 1.5 million steps. When the attention mechanism was integrated, the performance improved slightly, yielding a reward of around 170. The LSTM network achieved more consistent results with a reward of approximately 180. However, the LSTM combined with the transformer exhibited the best performance, achieving a reward of over 190, highlighting the benefits of combining temporal sequence processing with attention mechanisms..

4.2 Robustness in Simulatio

In the pursuit of a more comprehensive assessment of our leading model, we ventured into simulating intricate environments, presenting diverse challenges for the biped robot. These simulations aimed to emulate real-world scenarios where the robot might face unexpected terrains or disturbances.
To this end, we varied several factors to gauge the model's adaptability and resilience:
Ground Slope: This tests the robot's ability to walk on inclined surfaces without losing balance.
Ground Friction Coefficients: Alterations in static, dynamic, and rolling friction coefficients simulate different terrains, from icy landscapes to gravel-filled paths.
External Forces: By applying push and pull forces, we assessed the robot's stability against sudden external disturbances.

Weight Variations: By changing the robot's counterweight, we simulated scenarios where the robot might need to carry additional loads.
Figure 5 provides a visual representation of the biped robot being tested under these varying conditions in the Mujoco environment.
The results, as tabulated in Table 1, are telling. Our proposed strategy network structure notably surpasses traditional bipedal robot designs, displaying commendable performance even under the influence of multiple disturbances. Specifically:
The LSTM+Transformer architecture showcased remarkable resilience, especially in scenarios with steep ground slopes and high friction coefficients. Its ability to maintain stable locomotion under significant push or pull forces was also noteworthy.
Moreover, even when the robot's counterweight was varied, our model managed to adapt and sustain a stable gait, underscoring its versatility.
In essence, these simulation tests reiterate the robustness of our model, emphasizing its potential for real-world applications where unpredictable challenges might be commonplace.

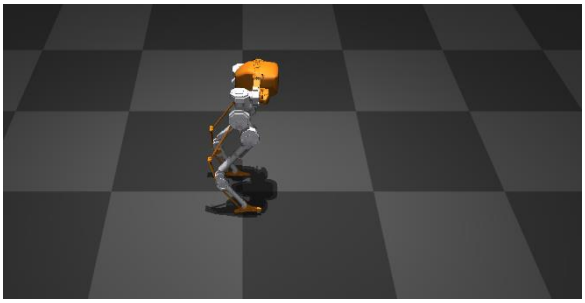

Fig.5. Test of Biped Robot in Simulation Environment

Table 1. Minimum disturbance for stable walking of bipedal robot for 200 seconds.

Network Architecture	Ground Inclination (in degrees)	Variation in Ground Static Friction Coefficient
feedback	3	0.5
feedback+transformer	4	0.4
lstm	4	0.7
Lstm+transformer	10	1

Network Architecture	Variation in Ground dynamic Friction Coefficient	Variation in rolling friction Coefficient
feedback	0.2	0.2
feedback+transformer	0.2	0.4
lstm	0.3	0.3
Lstm+transformer	0.5	0.5

Network Architecture	Pushes and Pulls(N)	Load(kg)

feedback	5	2
feedback+transformer	10	2
lstm	12	3
Lstm+transformer	15	4

5 CONCLUSION AND FUTURE WORK

In this study, we proposed an innovative approach for bipedal robot control by integrating LSTM and Transformer layers within Deep Reinforcement Learning. The methodology leverages the Proximal Policy Optimization algorithm, emphasizing the neural network policies associated with the dynamics of the robot. Our experimental results underscored the effectiveness of the proposed model, demonstrating both its robustness in handling complex temporal dependencies and its adaptability in addressing dynamic nonlinear challenges.

Key insights from our experiments highlighted:

- The combined architecture of LSTM and Transformer showed superior performance in capturing long-term dependencies and long-distance relationships, respectively.
- The adoption of Bayesian Optimization for reward function tuning significantly improved the efficacy of the training process. The method proved more efficient and robust compared to traditional methods, effectively navigating the intricate hyperparameter space.
- The use of attention mechanisms, particularly in the Transformer layers, enhanced the network's ability to focus on pertinent input segments, resulting in improved learning outcomes.

Looking to the future, several avenues beckon exploration and enhancement. Implementing domain randomization stands as a priority, allowing the robot to encounter randomized disturbances during training, thereby fostering an improved model robustness. As we transition our model from simulation to real-world applications, the need for fine-tuning motor parameters becomes apparent. Here, the inclusion of imitation learning can be pivotal, minimizing discrepancies between simulation outputs and tangible real-world operations.

Moreover, the application of curriculum-based learning is an exciting prospect. By gradually introducing more complex tasks and disturbances during training, we anticipate our model to develop even more nuanced and advanced behaviors. This adaptive learning approach is expected to enable our bipedal robot to exhibit intricate navigation capabilities and other sophisticated skills.

In essence, while our current findings are promising, the horizon offers numerous opportunities to refine, expand, and optimize our methodologies, promising even more advanced and reliable bipedal robot control in the future.

References

1. Xie Z, Berseth G, Clary P, et al. Feedback control for cassie with deep reinforcement learning[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 1241-1246..
2. Miller W T. Real-time neural network control of a biped walking robot[J]. IEEE Control Systems Magazine, 1994, 14(1): 41-48.
3. Yang C, Yuan K, Heng S, et al. Learning natural locomotion behaviors for humanoid robots using human bias[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2610-2617..
4. Li T, Geyer H, Atkeson C G, et al. Using deep reinforcement learning to learn high-level policies on the atrias biped[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 263-269.
5. Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.
6. Melo L C, Máximo M R O A. Learning humanoid robot running skills through proximal policy optimization[C]//2019 Latin american robotics symposium (LARS), 2019 Brazilian symposium on robotics (SBR) and 2019 workshop on robotics in education (WRE). IEEE, 2019: 37-42.
7. Wu J, Chen X Y, Zhang H, et al. Hyperparameter optimization for machine learning models based on Bayesian optimization[J]. Journal of Electronic Science and Technology, 2019, 17(1): 26-40.
8. Nguyen V. Bayesian optimization for accelerating hyper-parameter tuning[C]//2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE). IEEE, 2019: 302-305.
9. Klein A, Falkner S, Bartels S, et al. Fast bayesian optimization of machine learning hyperparameters on large datasets[C]//Artificial intelligence and statistics. PMLR, 2017: 528-536.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
11. .Bednarek M, Łysakowski M, Bednarek J, et al. Fast haptic terrain classification for legged robots using transformer[C]//2021 European Conference on Mobile Robots (ECMR). IEEE, 2021: 1-7.
12. Çatalbaş B, Morgül Ö. Two-legged robot motion control with recurrent neural networks[J]. Journal of Intelligent & Robotic Systems, 2022, 104(4): 59.
13. Li T H S, Kuo P H, Cheng C H, et al. Sequential sensor fusion-based real-time LSTM gait pattern controller for biped robot[J]. IEEE Sensors Journal, 2020, 21(2): 2241-2255.
14. Su B, Gutierrez-Farewik E M. Gait trajectory and gait phase prediction based on an LSTM network[J]. Sensors, 2020, 20(24): 7127.
15. Li J, Xu Y, Shi H. Bidirectional LSTM with hierarchical attention for text classification[C]//2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2019: 456-459.
16. Tran D T, Iosifidis A, Kanniainen J, et al. Temporal attention-augmented bilinear network for financial time-series data analysis[J]. IEEE transactions on neural networks and learning systems, 2018, 30(5): 1407-1418.
17. Li Y, Zhu Z, Kong D, et al. EA-LSTM: Evolutionary attention-based LSTM for time series prediction[J]. Knowledge-Based Systems, 2019, 181: 104785.