A double network structure anti-jamming algorithm based on deep reinforcement learning

Yifan Wang¹, Xin Liu¹*, Mei Wang²

¹Guilin University of Technology, College of Information Science and Engineering, Guilin, China
²Guilin University of Technology, Guilin, China

*Corresponding author: liuxin2017125@glut.edu.cn

Abstract. At this stage, when dealing with intelligent jammer with environmental cognition, most anti-jamming methods only consider how to maximize avoidance of it. However, the user’s previous waveform and frequency actions have been leaked. As the intelligent jammer continues to learn, the performance of anti-jamming will decline. Existing literature has proposed the idea of hidden anti-jamming. The user uses rational communication action-making to prevent the jammer from obtaining the information of the user so that the jammer cannot target the user for jamming, but its model is relatively simple and cannot cope with complex jamming types. This paper designs a kind of decision related judgment module between jammer and user based on Generative Adversarial Networks (GAN), which uses the environment state received by the user to reversely infer the user's decision whether to avoid jammer’s sense. Specifically, the neural network is used to fit the environment state under known user information, and calculate the loss between it and the real environment state to evaluate whether the user's decision is to avoid jammer’s sense. Then, this paper proposes a deep reinforcement learning algorithm with double-network structure, which can deal with various types of complex jamming while ensuring anti-jamming performance. The final simulation results show that the improved deep reinforcement learning algorithm proposed in this paper can improve the anti-jamming performance, achieve hidden anti-jamming, and have a wider range of applications.

Keywords: Environmental Cognition, Hidden Anti-Jamming, Generative Adversarial Networks, Double-Network Structure, Deep Reinforcement Learning

1. Introduction

With the advancement of technology, the development of wireless networks is getting faster and faster, and the proportion in people’s daily life and military communications is increasing year by year[1]-[3]. The wireless network brings many conveniences to people's life and works because of the openness of its transmission medium, but this also makes it extremely vulnerable to interference attacks[4]-[6].

At present, the main research direction of anti-jamming scholars is how to avoid jamming
effectively, that is, how to maximize communication without jamming on the same frequency. The general idea of anti-jamming technology based on avoid idea is that users use intelligent algorithms to reasonably change their frequencies to avoid jamming according to the receive spectrum state. Anti-jamming technologies based on intelligent algorithms such as genetic algorithm, particle swarm algorithm, and artificial bee colony algorithm[7]-[11] are mostly used to find undisturbed frequency transmission signals through search. However, due to the large limitations of this type of algorithm, it cannot be satisfied in an environment with an intelligent jammer. In order to cope with intelligent jammer with learning ability, reinforcement learning algorithms have been widely used and achieved relatively ideal results[12]-[16]. The core idea is to set a reward function to allow users to learn the rules of jammer’s decision-making to predict the frequency of jammer at the next moment and avoid them in advance. However, reinforcement learning technology faces the problem of dimensional disasters in a complex environment, which causes slow learning speed and restricts application scenarios. Anti-jamming methods based on deep reinforcement learning have become a popular research direction today. However, at this stage, the idea of anti-jamming methods based on deep reinforcement learning is mainly to avoid jamming, and its core ideas are roughly the same as those of reinforcement learning[17]-[21].

In summary, the anti-jamming technology based on the idea of avoidance at this stage is constantly improving the intelligence of the algorithm to counter jammer, and can even accurately predict the jamming frequency at the next moment to avoid it in advance. However, the past communication user's signal waveforms and frequency decision information may have been exposed. As the jammer continues to learn the above information, the anti-jamming performance will decrease gradually. The existing anti-jamming technology based on the hidden anti-jamming idea can avoid the leakage of user information when dealing with the jammer. Once the jamming type becomes complicated, the performance will be reduced[22]. In this regard, this paper adopts the hidden anti-jamming idea and proposes a deep reinforcement learning anti-jamming algorithm with a double-network structure. First of all, this paper designs a GAN network-based decision-related judgment module between jammer and user, which uses the GAN to fit the environmental state under known user information, and evaluates whether the user’s information is based on the loss between it and the real environmental state. Then, this paper designs a deep reinforcement learning framework, which can not only improve the user's anti-jamming performance but also avoid the user's information leakage. Finally, this paper proposes a deep reinforcement learning algorithm with a double network structure(ADRLDN), which can not only improve the anti-jamming performance but also deal with various types of complex jamming, which is widely used in actual scenarios.

The main contributions of this paper are as follows:

• Design a GAN network-based user and jammer decision-making correlation judgment module. Use the GAN to fit the environment state when the user's decision information is known, and calculate the loss between it and the real environment state, it is judged whether the user information is obtained by the jammer. From an overall perspective, if a jammer can sense the user's decisions, targeted jamming will be implemented at the next moment, so the environmental state will inevitably change accordingly. Conversely, if the jammer does not sense the user's decisions, the jamming implemented at the next moment will be more random, and the environment state will definitely deviate from the environment state under the known user's decision. According to this, the user can analyze the hidden effect of his own decision.

• Propose a deep reinforcement learning algorithm with a double network structure. Through the decision-making correlation judgment module, the loss between the real environment state and the environment state when the user is known to make a decision is calculated, and it is used as the direction of the user's learning. It is also can ensure that the user's decision is not captured by the jammer to achieve hidden anti-jamming. Even if the jammer intelligence is improved in the later stage, it is impossible to formulate an effective jamming strategy.

The rest part of this paper is presented as follows. In Section 2, the anti-jamming system model is given. After that, the deep reinforcement learning algorithm with a double network structure is based on
the hidden strategy in Section 3. Besides, the analysis of simulation results and the conclusion are given in Section 4 and Section 5 respectively.

2. Anti-jamming system model

![Anti-jamming system model](image)

As shown in the system model in Fig. 1, this article considers a scenario where a user (including a transmitter and a receiver) and a jammer are confronted in the presence of an environmental signal with a bandwidth of $B$. Due to geographical location and other reasons, the information of the user and the jammer sensing environmental signals is not the same. As shown in Fig. 2, the jammer transmits jamming signals at specific frequency points after learning according to the sensed environmental signal $E(f)$ and user signal $U(f)$. Jammer to achieve the purpose of jamming with the user, the transmit power $p_j$ of the jamming signal $J(f)$ must be relatively large and will be received by the user. The transmit power $p_u$ of the user signal $U(f)$ does not need to be large. Because of the presence of environmental signals, there is a weak point in the jammer’s sense. At the same time, if the frequency point $f'_t$ of the user signal coincides with the frequency point $f'_e$ of the environmental signal, the user's signal will be covered by the environmental signal, and the information of the user can be sensed from the jammer. The frequency $f'_e$ and power $p_e$ of the environmental signal $E(f)$ are unknown to the user, but some rules can be found from the changes in the environmental state.

![Work process of Jammer and user](image)

The starting and ending frequencies of the communication band of the two parties are respectively $f_s$ and $f_e$. The user can select frequency $f'_t = [f'_s, f'_e]$ as the center frequency, transmit the signal with $p_u = \int_{f'_s}^{f'_e} U(f) df$ as the transmit power, and $b_u$ as the spectral bandwidth of the user signal, so the user’s signal $U_t(f) = U(f'_t, b_u, p_u)$. The jammer can also freely choose frequency $f'_j$ as the center frequency to transmit jamming signals in the communication frequency band $[f'_s, f'_e]$, and its spectrum bandwidth can be changed according to the sensed state $b_j \in [b'_n, b'_x]$, where $b_n, b_x$ is the minimum and maximum bandwidth of the jamming signal, and $J_j(f) = J(f'_j, b_j, p_j)$ is the jamming signal. $E_t(f)$
represents the environmental signal perceived by the user at a time \( t \), and \( n(f) \) represents noise.

The user, which is disposed at the receiving end, continuously senses the whole communication band. Considering the coexistence of the user signal and the jammer signal, the PSD of the signal at the receiving end can be expressed as:

\[
R_t(f) = g_u U_t(f) + g_J J_t(f) + g_E E_t(f) + n(f)
\]  

(1)

The discrete spectrum sample value is defined as \( s_{nf} = 10 \log_{10} \left[ \int_{f_{n-1}}^{f_n} R_t(f + f_j) df \right] \), where \( \Delta f \) is the resolution of the spectrum analysis. The user determines the transmission frequency based on the spectrum vector \( s_t = \{s_1, s_2, \ldots, s_N\} \). Considering that the decision of the opponents is related to the environmental state for a long time in the past, this article defines the environmental state as \( S_t = \{s_t, s_{t-1}, \ldots, s_{t-T}\} \) and \( T \) represents the length of historical states of backtracking. The user makes a learning decision according to the environment state \( s_t \) and transmits it to the transmitter through the control link. The user’s decision-making only involves frequency change \( a_t^i = f_t^i \in [f_s, f_e] \). Jammer’s decision-making \( a_t^j \) can not only change the frequency but also change its signal bandwidth according to the spectrum state, namely \( a_t^j \in [f_t^j, b_j] \), where \( f_t^j, f_j, b_j \in [b_s, b_e] \).

3. Deep reinforcement learning algorithm with dual network structure

Whether the user’s decision is sensed by the jammer is unknown to the user, but some information can be inferred from the change of the environmental state. The jammer will implement targeted jamming based on the user’s decision sequence, and the decision made by the jamming will affect the instantaneous environment state sensed by the user at the next moment. Therefore, this paper designs a GAN network-based decision-related judgment module between jammer and user, which uses a neural network to fit the environment state under known user decisions, calculates the loss between it and the real environment state and then evaluates whether the user’s decision avoids jammer’s sense.

The instantaneous environment state \( s_t^i \) received by the user is the change that occurs after the jammer makes the action \( a_{t-1}^j \), so there must be a certain functional relationship between the two, as shown in formula (2):

\[
s_t^i = \Psi(a_{t-1}^j)
\]  

(2)

Jammer making the action \( a_{t-1}^j \) is targeted jamming based on the user's previous action sequence \( \overline{a_{t-1}} = (a_{t-1}^i, a_{t-2}^i, \ldots, a_{t-T}^i) \). Therefore, the jammer’s action \( a_t^j \) must have a certain relationship with the user's previous decision sequence \( \overline{a_{t-1}} = (a_{t-1}^i, a_{t-2}^i, \ldots, a_{t-T}^i) \), as shown in formula (3):

\[
a_t^j = \varphi(\overline{a_{t-1}})
\]  

(3)

Although we cannot clearly know the specific calculation methods of these two functions \( \psi(x), \varphi(x) \), combining formula (2) and formula (3), we can get a conclusion that there also have a certain functional relationship in instantaneous environment state \( s_t^i \) received by the user and the user’s action sequence \( \overline{a_{t-1}} \), as shown in formula (4):

\[
s_t^i = \Xi(\overline{a_{t-1}})
\]  

(4)

In this article, it is necessary to compare the fitted environmental state with the real environmental state, and it is necessary to ensure both the generation of the network and the evaluation of the effect of
the network. Generative Adversarial Networks satisfies these two points, so this article uses Generative Adversarial Networks as the decision-making correlation module for user and jammer. If the loss between the instantaneous environment state $s_i^N$ fitted by the network and the instantaneous environment state $s_i^i$ received by the user is small and can eventually converge, then it can be determined that the jammer has captured the user’s information. Conversely, if the loss between the two does not decrease or converge, it can be inferred that the jammer does not obtain the user’s information.

The specific decision-making process of the opposing parties is shown in Figure 3. At the time $t$, the jammer is based on the sensed environmental state $S_j^j$, and the action $a_j^j$ is learned by the jammer decisions and learning module. According to the received environment state $S_i^i$, the user learns from the anti-jamming decisions and learning module and makes an action $a_i^i$. At the same time, the user’s action sequence $a_{t-1}^i$ is input into jammer and user decision-making correlation judgment module, and finally, the instantaneous environment state $s_i^N$ obtained by network fitting and the instantaneous user perception is calculated as the loss $L_i^i(\theta^i)$ between environmental states $s_i^i$. In order to ensure the fitting effect of the network, $L_i^i(\theta^i)$ reversely updates the parameters in the judgment module of the correlation between user and jammer decisions after each iteration $\theta^i$.

Besides, the user wants its action information not to be sensed by the jammer, therefore the threshold $\Phi_t$ is set. When $L_i^i(\theta^i)$ the value is less than or equal to the threshold $\Phi_t$, the user's action is considered to be obtained without jammer. On the contrary, when $L_i^i(\theta^i)$ the value is greater than the threshold $\Phi_t$, it is considered that the user's action has been obtained by the jammer. Define user return $r_i^i$ as:

$$r_i^i = \begin{cases} 1, & L_i^i(\theta^i) \leq \Phi_t \\ 0, & L_i^i(\theta^i) > \Phi_t \end{cases}$$ (5)

Considering that the expansion of the time dimension will cause the environment to become very huge, the convergence speed of reinforcement learning is too slow and the time is too long to be directly applied. In this paper, a convolutional neural network (CNN) is used to estimate the Q function. The convolutional neural network is convenient to capture the time-frequency law of jamming. At the same time, its effectiveness has been proved in the literature [14].

![Fig. 3. Flow chart of the decision-making process](image-url)
Algorithm 1: Deep reinforcement learning algorithm with double network structure

**Initialize** $D = \emptyset$, $i = 0$, network weights $\theta_0$ and $\theta^g_0$ with random values,

$\varepsilon=1$, Training = True.

**For** $t = 1, 2, \cdots , \infty$ **do**

If Training then

Select action $a_t$ via $\varepsilon$-greedy an algorithm or Select action $a_t = \arg \max_a Q(S_t, a; \theta)$

Generative Adversarial Networks

Input $a_{t-1}$, Output $s^N_t$

Compute $L^g_t(\theta^g) = E(s^N_t - s_t)^2$,

Compute $r_t = \begin{cases} 1, & L^g_t(\theta^g) \leq \Phi_t \\ 0, & L^g_t(\theta^g) > \Phi_t \end{cases}$

Compute $\nabla \theta L^g_t(\theta^g)$, update $\theta^g$ with gradient descent algorithm.

Execute $a_t$, record $r_t$ and sense $S_{t+1}$, store $e_t = \left(S_t, a_t, r_t, S_{t+1}\right)$, Store $e_t$ in $D$.

If $|D| > N/2$

Select $e_t$ from $D$ randomly,

Compute $\eta_t = r_t + \gamma \max_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \theta_{t-1})$,

Compute $L_i(\theta_t) = E\left(\eta_t - Q(S_t, a_t; \theta_t)\right)^2$,

Compute $\nabla \theta L_i(\theta_t)$, update $\theta$ with gradient descent algorithm, and $i = i + 1$,

$\varepsilon = \varepsilon - \Delta \varepsilon$

If $\varepsilon < 0$ then $\varepsilon = 0$

End If

End For

The Q function indicates that under an environmental state $S_t$, choosing action $a_t$ can obtain long-term rewards, and its definition is:

$$Q(S_t, a_t) = E\left(r_t + \gamma \max_{a_{t+1}} Q(S_{t+1}, a_{t+1}) | S_t, a_t\right) \quad (6)$$

Among them, $S_{t+1}$ represents the state at the next moment after the user chooses a decision $a_t$ in state $S_t$, $\gamma$ represents the reward discount factor, and $\max_{a_{t+1}} Q(S_{t+1}, a_{t+1})$ represents the maximum Q value that can be obtained by selecting a known action in experience under the environmental state $S_{t+1}$. Then we can learn and train according to the deep reinforcement learning algorithm. The specific algorithm is shown in Algorithm 1.

4. Simulation Results and Analysis

In this article, the user and jammer are set to fight in the frequency band of spectrum bandwidth
\( B = 20\text{MHz} \). The user’s signal bandwidth is \( b_u = 1\text{MHz} \). Jammer’s signal bandwidth \( b_j \in [b_n, b_x] \) can be changed according to the spectrum state. Among them, the minimum bandwidth is \( b_n = 1\text{MHz} \) and the maximum bandwidth is \( b_x = 3\text{MHz} \). The opposing parties conduct a spectrum detection once per \( 1\text{ms} \), and its center frequency can be changed once per \( 10\text{ms} \) according to the decision. The spectrum resolution \( \Delta f \) is set to \( 1\text{kHz} \), and the user retraces the time duration \( T = 100\text{ms} \). After many experiments, the threshold \( \Phi_f = 0.55 \) has the best effect, so set threshold \( \Phi_f = 0.55 \) and reward discount factor \( \gamma = 0.8 \).

First of all, the experiment needs to verify the anti-jamming performance of the method proposed in this article. In the case of jammer based on reinforcement learning, when the user adopts the method proposed in this article and the deep reinforcement learning algorithm based on avoidance strategy (ADRLA), the normalized throughput of the system changes as shown in Figure 4. It can be seen from the figure that the method proposed in this paper can improve the throughput of the system when dealing with an intelligent jammer.

Fig. 4. Normalized throughput change graph

Then, in order to verify the effectiveness of the method proposed in this paper in terms of concealment, the method proposed in this paper and ADRLA are simulated in the same condition. In this experiment, the presence of environmental signals is considered. Jammer is based on a reinforcement learning algorithm and is targeted for jammer based on the sensed environmental state. It can be seen from Figure 5 that when dealing with an intelligent jammer, ADRLDN can reduce the probability of users being jamming by about 0.15 compared with ADRLA. It can be seen from Figure 6 that when the user adopts ADRLA, the loss value will continue to decrease with the number of iterations and eventually converge. When using the ADRLDN, the loss value does not decrease with the number of iterations but fluctuates randomly until the end of the algorithm. The result proves that when the user adopts the ADRLDN, the user’s information is not obtained by the jammer, which improves the anti-jamming performance and reduces the probability of the user being jamming.

Fig. 5. Probability of users being jamming
In order to verify the applicability of ADRLDN, comparative simulation experiments were carried out when dealing with intelligent jammer with different jamming strategies. In this experiment, the presence of environmental signals is also considered, and two different jamming strategies are considered. The first type of jamming only changes its frequency, the second type of jamming changes the signal bandwidth based on changing the frequency. The User uses three different methods which are mentioned above and an anti-jamming deep reinforcement learning algorithm based on hiding strategy(ADRLH) to conduct simulation experiments, that is and the experimental results are compared with the probability of users being jamming.

It can be seen from Fig. 7 that the method based on the hidden anti-jamming idea is obviously better than the anti-jamming method based on the avoiding idea in dealing with an intelligent jammer. As the jamming type becomes more complex, as shown in Fig. 8, the anti-jamming performance of ADRLH is very severe, and the ADRLDN proposed in this paper will not reduce the effect due to the complexity of the jammer. Experimental results show that ADRLDN can adapt to a more complex jammer and is closer to the actual scene in the application scenario. It is not difficult to imagine that ADRLDN can be more widely used in practical applications.
5. Conclusion
This paper adopts the hidden anti-jamming idea and proposes a deep reinforcement learning algorithm with a double-network structure. Firstly, based on the generative countermeasure network, the user and jammer decision-making correlation judgment module is designed, which can deduce whether the user's decision is sensed by jammer through the change of environment state. Specifically, it uses the GAN to fit the environment state when the known user's information, and calculate the loss value between it and the real environment state, and use the loss value to evaluate whether the user decides to avoid jammer's sense. It is conceivable that if the loss value is larger, the user gets less information obtained by the jammer. Combined with the deep reinforcement learning algorithm, a deep reinforcement learning algorithm with a double-network structure is proposed, which strives to counter more complex jamming models while ensuring that the user's information is not obtained. Finally, the simulation experiment results show that the deep reinforcement learning algorithm with the double-network structure proposed in this paper is not only superior in anti-jamming performance than the current anti-jamming method based on avoiding the idea, but also can adapt to more complex types of jamming.

Acknowledgment
This research work was supported by the National Natural Science Foundation of China under Grant 61961010, the Key Laboratory Found of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology) under Grant No. CRKL180203, the ‘Ba Gui Scholars’ program of the provincial government of Guangxi”.

References
[1] K. Pelechrinis, M. Iliofotou and S. V. Krishnamurthy, "Denial of Service Attacks in Wireless Networks: The Case of Jammers," in IEEE Communications Surveys & Tutorials, vol. 13, no. 2, pp. 245-257, Second Quarter 2011
[2] A. Kavianpour and M. C. Anderson, "An Overview of Wireless Network Security," 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, 2017, pp. 306-309
[3] Y. Zou, J. Zhu, X. Wang and L. Hanzo, "A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends," in Proceedings of the IEEE, vol. 104, no. 9, pp. 1727-1765, Sept. 2016
[4] R. Dorus and P. Vinoth, "Mitigation of jamming attacks in wireless networks," 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, 2013, pp. 168-171
[5] B. Yu and L. Zhang, "An improved detection method for different types of jamming attacks in wireless networks," The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), Shanghai, 2014, pp. 553-558
[6] S. Sowmya and P. D. S. K. Malarchelvi, "A survey of jamming attack prevention techniques in wireless networks," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, 2014, pp. 1-4
[7] T. Haihong, L. Guisheng and Y. Jiang, "Space-borne antenna adaptive anti-jamming method based on gradient-genetic algorithm," in Journal of Systems Engineering and Electronics, vol. 18, no. 3, pp. 469-475, Sept. 2007
[8] Zhao Wang, Youming Li, Cuiru Zhao and Jiehui Chen, "An energy-efficient resource allocation scheme based on genetic simulated annealing algorithm," 2012 International Symposium on Communications and Information Technologies (ISCIT), Gold Coast, QLD, 2012, pp. 867-870
[9] Xianyang Hui, J. Chen and Yingtao Niu, "An initial population optimization-based particle swarm algorithm for anti-jamming communication decision engines," 11th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2015), Shanghai, 2015, pp. 1-6
[10] Xianyang Hui, Yingtao Niu and Mi Yang, "Decision method of anti-jamming communication
based on binary artificial bee colony algorithm," 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, 2015, pp. 987-991
[11] F. Ye, Z. Zhou, H. Tian, Q. Sun, Y. Li and T. Jiang, "Intelligent Anti-Jamming Decision Method Based on the Mutation Search Artificial Bee Colony Algorithm for Wireless Systems," 2019 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), Atlanta, GA, USA, 2019, pp. 27-28
[12] D. A. Alfenas, D. P. Shibata, J. Jose and M. R. P. Barretto, "Adaptive Markov Systems: Formulation and Framework," in IEEE Latin America Transactions, vol. 12, no. 7, pp. 1271-1277, Oct. 2014
[13] T. Chen, J. Liu, L. Xiao and L. Huang, "Anti-jamming transmissions with learning in heterogeneous cognitive radio networks," 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), New Orleans, LA, 2015, pp. 293-298
[14] X. Pei et al., "Joint Time-frequency Anti-jamming Communications: A Reinforcement Learning Approach," 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 2019, pp. 1-6
[15] L. Kong et al., "A Reinforcement Learning Approach for Dynamic Spectrum Anti-jamming in Fading Environment," 2018 IEEE 18th International Conference on Communication Technology (ICCT), Chongqing, 2018, pp. 51-58
[16] S. Liu et al., "Pattern-Aware Intelligent Anti-Jamming Communication: A Sequential Deep Reinforcement Learning Approach," in IEEE Access, vol. 7, pp. 169204-169216, 2019
[17] G. Han, L. Xiao and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2087-2091
[18] X. Liu, Y. Xu, Y. Cheng, Y. Li, L. Zhao and X. Zhang, "A heterogeneous information fusion deep reinforcement learning for intelligent frequency selection of HF communication," in China Communications, vol. 15, no. 9, pp. 73-84, Sept. 2018
[19] X. Liu, Y. Xu, L. Jia, Q. Wu and A. Anpalagan, "Anti-Jamming Communications Using Spectrum Waterfall: A Deep Reinforcement Learning Approach," in IEEE Communications Letters, vol. 22, no. 5, pp. 998-1001, May 2018
[20] W. Li, J. Wang, L. Li, G. Zhang, Z. Dang and S. Li, "Intelligent Anti-Jamming Communication with Continuous Action Decision for Ultra-Dense Network," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019
[21] Y. Li, Y. Xu, X. Wang, W. Li and W. Bai, "Power and Frequency Selection optimization in Anti-Jamming Communication: A Deep Reinforcement Learning Approach," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 815-820
[22] Y. Wang, X. Liu, M. Wang and Y. Yu, "A hidden anti-jamming method based on deep reinforcement learning", 2020, arXiv:2012.12448.