DQN-Based Shaped Reward Function Mould for UAV Emergency Communication

Chenhao Ye , Wei Zhu * , Shiluo Guo , Jinyin Bai

Posted Date: 14 August 2024

doi: 10.20944/preprints202408.0979.v1

Keywords: unmanned aerial vehicle UAV; deep Q-learning network DQN; reward shaping

Article

DQN-Based Shaped Reward Function Mould for UAV Emergency Communication

Chenhao Ye, Wei Zhu*, Shiluo Guo and Jinyin Bai

School of Information and Communication, National University of Defense Technology, Wuhan 430000, China; zhuwei929@hotmail.com (C.Y.); hg_guoshiluo@163.com (S.G.); bbaibai0612@163.com (J.B.)
* Correspondence: a379763802@163.com

Abstract: Unmanned aerial vehicle UAV has become an important tool for emergency communication. When disaster happens, the roads in the disaster area always are seriously damaged, and theground infrastructure communication facilities are damaged. Under emergency conditions, UAV can go into key areas as communication nodes to provide communication services for users in the area.But it's difficult for UAV to cover the entire disaster area and the efficient rescue time will be delayed. Therefore, how to make the UAV accurately perceive the regional situation and deployed in the right position is worth studying. In this paper, a virtual simulation environment is established reasonably and deep reinforcement learning algorithm is used to train UAV agents. At the same time, a series of problems such as sparse reward and long training time are still not solved in the development of reinforcement learning algorithms,and the reward function is designed to improve training efficiency. First of all, we set a specific mountain emergency communication scenario, and combined with the specific application of the UAV to carry out virtual simulation, and build a virtual environment. Furthermore, an additional shaped reward function is designed to address the sparse reward problem. Through the improvement of deep Q-learning network DQN algorithm and the reward design based on potential function, the final evaluation index is improved, and the effectiveness of the algorithm is verified. The experimental results demonstrate our work can effectively shorten the training time and increase convergencerate.

Keywords: unmanned aerial vehicle UAV; deep Q-learning network DQN; reward shaping

1. Introduction

Emergency communication refers to a situation when the original communication system is seriously damaged, the emergency department so as to better coordinate different departments and deal with emergencies efficiently. Due to the vast territory of our country and the frequent natural disasters in some areas, it is very important to
establish a scientific and effective emergency communication system. The movementand location of base stations are also limited by roads, ground fluctuation and vegetation, etc., and the setting of base stationsneeds to meet certain topographic and geomorphic conditions [1]. Therefore in most cases, due to terrain, resources and other constraints, largescale UAV networking emergency support cannot be carried out. At this time, it is very important to carry out dynamic planning for a single UAV ora group of UAV in order to maximize theprotection of users in demand.

After the post disaster rescue work is launched, commanders need a series of ba-sic information such as personnel location and key areas as decision-making support. At this time, the UAV can be a node, and it needs to traverse all the sensors in the disaster area and send the information back to the command post in time [2]. Underemergency conditions, command post nodesare very important for the rapid integrationof information reported at all levels.However, it is still difficult for the command post guarantee node to receive the situation information of the disaster area in time and make precise decisions. Therefore, it is worth thinking about how to maximize theefficiency of a single UAV so that it can guarantee as many users as possible. This paper establishes a virtual simulation

 doi:10.20944/preprints202408.0979.v1

environment and uses deep reinforcement learning to enable agents to explore solutions independently, providing reference for intelligent planning of UAV.

The UAV node has the characteristicsof strong mobility and is not affected by terrain environment. In the military context,the UAV play an important role in battlefield reconnaissance, surveillance, border patrol, positioning strikes and other military operations [3]. Yin Changsheng [4] used deep neural networks to carry out topologyplanning of emergency communication networks, and Chen Haoran [5] used DQN algorithm to carry out hierarchical design of emergency communication networks. Jiangbin Lyu[6] have proposed a method for the optimal deployment location of UAV. However, the above research still can not play the best support efficiency of a single UAV. In this paper, a UAV system is trained by adding DQN algorithm with reward shaping. The simulation environment is taken as the input. Finally, the trained agent UAV can automatically change the position, and the efficiency is better verified.

Aiming at the design of reward function, Wu Enda [7] designed the distance-based and subgoal-based heuristic reward function, and the experiment proved that it has a great effect on reducing training time. Dong Yunlong [8] proposed a theoretical optimization framework based on reward shaping drive to improve the training efficiency and stability of reinforcement learning according to the current low efficiency of reinforcement learning training. In this paper, reward shaping is applied to DQN algorithm, which improves the training efficiency of DQN algorithm to some extent. The main contribution of this paper is to continuously adjust parameters, optimize network structure, and improve agent performance. At the same time, the reward function based on potential function is added to improve the performance of the algorithm.

The work of this paper includes three aspects:

- Abstract modeling is carried out fortypical application scenarios of communica-tion equipment such as UAV, and environ-mental design is carried out based on Atari platform.
- A more effective action selection strategy is applied to replace the greedy strategy with the traditional DQN algorithm, and its performance is better than before.
- The reward function is redesigned to improve the efficiency and stability of the algorithm, and experimental verificationis carried out.

2. Materials and Methods

2.1 Reinforcement Learning

Reinforcement learning RL is mainly about abstracting the real world, transform-ing real problems into problems where agents take actions and strategies, and seek the best reward [9]. It does not require too much human prior knowledge, and relieson plenty computer computing power to carry out repeated experiments to improve the performance of agents.

RL itself can be modeled as a Markov decision process , where S represents the set of states, A represents the set of actions, T represents the transition probability, R represents reward reporting, and represents discount factors, which are used to represent short-term and long-term rewards.

The action value function represents the cumulative return of the agent from the initial state to the end of the turn by taking an action. It can be expressed as:

$$Q^*(s,a) = \prod^{max} E[R_t|s_t = s, a_t = a, \Pi], \quad (1)$$

At the same time, the optimal action value function may have more than one kind, the optimal action value function can be expressed as:

$$Q^*(s,a) = E_{s'-s}[r + \gamma \overset{max}{a'} Q(s',a')|s,a], \quad (2)$$

2.2 DQN

In the professional field the research on deep reinforcement learning algorithm is mainly divided into value function and strategy function. DQN algorithm is a deep reinforcement learning algorithm based on value function. By introducing the experience return mechanism, an experience pool is set

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2024 doi:10.20944/preprints202408.0979.v1

up to store training samples, and then a small batch of training samples is selected for iterative updating. This method can reduce the correlation between samples [10].

At the same time, the target network and the evaluation network are separated to improve the stability of the algorithm. The DQN algorithm sets up two neural networks, namely Target Network and Action Network. Firstly, the neural network is used to fit the Q value, and the F function is designed to reset the reward function according to the agent collision effect. Th parameter value of the network is updated in real time, and it can be assigned to the target network after N iterations. By separating two networks, the stability and convergence of the algorithm can be improved [11]. Its optimal value function can be expressed as:

$$Q^*(s,a) = {}^{max}_{\Pi} E(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s, a_t = a, \Pi), \tag{3}$$

The basic DQN algorithm uses the ε-greedy strategy to balance exploration and utilization to improve the agent's exploration of the environment. The strategy can be expressed as:

$$\Pi^* = \begin{cases} argmax V^{\Pi}(s) & 1-\varepsilon \\ rand(a) & \varepsilon \end{cases}, \tag{4}$$

The agent randomly selects the action in the action space with the probability of, and selects the known optimal action with the probability of 1-ε. The traditional DQN algorithm can solve the problem of dimensionality disaster when Q-learning algorithm deals with complex environment. However, the traditional DQN algorithm still cannot overcome the sparse reward problem, which requires a lot of extra training time.

The following is the frame of DQN, as shown in Figure 1.

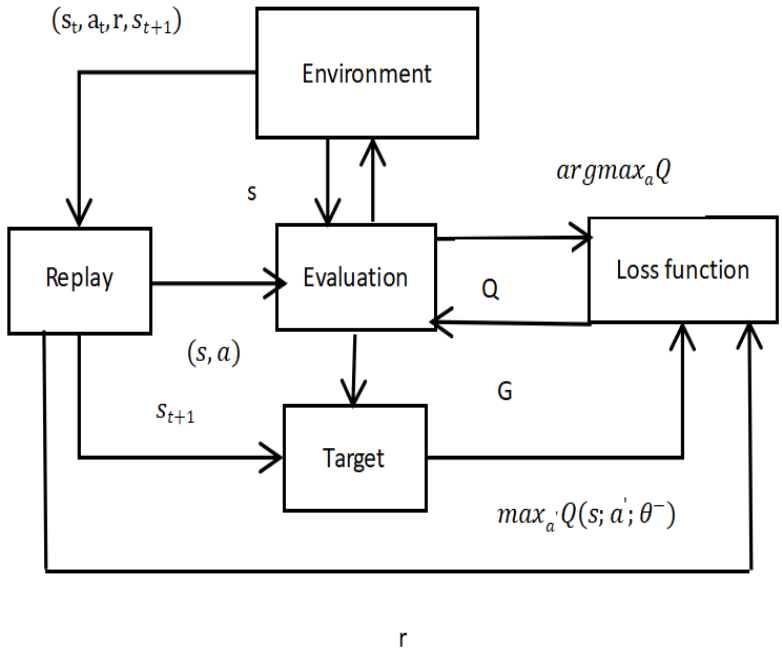

Figure 1. DQN.

2.3. System Model

This paper considers the UAV in mountain emergency communication scenario. This paper assumes that there are N users in unknown location, 1 UAV communication node and several relay nodes in a mountain area. At some occasions, due to the impact of landslides, all base stations in the area were seriously damaged and could not be used normally. Due to the narrow terrain and other factors, only 1 UAV can be used as a communication node, and the rest of the relay nodes can still be used normally. At this point, the UAV needs to establish a route connecting all users and communicate key information to each node.And we set both sides as relay nodes, and the information will be returned directly after being transmitted to the backbone node. When the information reaches

 doi:10.20944/preprints202408.0979.v1

the node,it is rewarded accordingly, and the user automatically exits the network. The simple test platform simulates time as a discrete time step, and the agent performs each action at a constant time step.

The disaster area is defined as a 25*43 grid, in which all the objects to be observed by the UAV. The UAV moves at a constant speed, moving 3 grids per step. In order to facilitate learning, it is assumed that the UAV agent does not change the starting position. At the initial moment, the UAV agent can first select the first observation target arbitrarily and start movement.After completing the observation round for all objects, the UAV agent will get the corresponding reward. If the observation cannot be completed within the specified time,the round ends and the UAV agent cannot receive the reward.

This article assumes:

- The paper does not consider attenuation and other factors in the transmission process, simplifying the system model.
- The relay node directly transmits the signal downward after receiving it, and there is no delay in the middle.
- The UAV flies at a constant slow speed and only moves in two-dimensional space.

3. Improvement and Experiment

3.1. Improvement of Algorithm

The simple reinforcement learning algorithm has high computational cost, which often requires millions of training data or more, and the training cost is relatively large. At the same time, it still faces the problems of poor training effect and lengthy training time. Above all, the main reason is the sparse reward problem. As for the agent, the reward can only be obtained in the final state of the game, while the intermediate state cannot be learned [12].

In the discrete space, the current stage focuses on the algorithm level, while the reward function is designed from the perspective of general theory. And the reward function of DQN also mainly depends on the environment, which has a large room for improvement. Q function is represented by convolutional layer [13].

The training efficiency, convergence speed and stability of RL can be improved by reasonable reward function design. RL requires a lot of training resources for training, so if the experiment is repeated on the reward function, it will bring additional reward. And reward shaping refers to t-he training process of acceleration reinforcement learning by adding potential energy function according to the corresponding mathematical form on the basis of the original reward function. However, in many practical cases, the agent will receive little or no reward, which will lead to learning failure and ineffective exploration [14].

The shaping function refers to the sparse reward environment and the reward environment with action evaluation [15]. There is a reward shaping module and a control module that outputs a series of actions to maximize the reward signal [16]. Under the condition of constant external rewards, the designs of internal rewards areequivalent to adding prior human knowledge. But at the same time, strategy invariance is very important for reward shaping.

The following is the frame of reward shaping, as shown in Figure 2.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2024 doi:10.20944/preprints202408.0979.v1

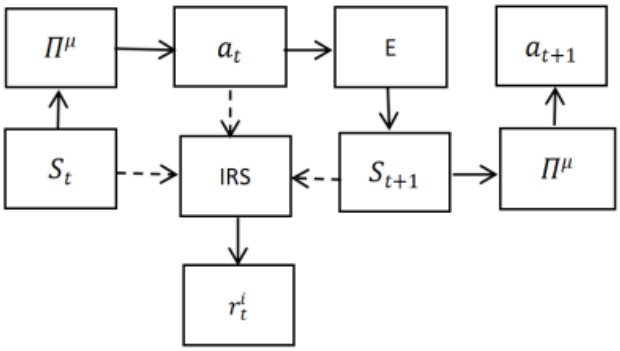

Figure 2. Reward shaping.

The following formula can prove that potential based reward function shaping does not change the original strategy：

$$R^{'}(s,a,s') = R(s,a,s') + F(s'), \tag{5}$$

$$F(s^{'}) = \varphi(s') - \gamma\varphi(s), \tag{6}$$

The shaping function based on potential energy is that the optimal strategy of reinforcement learning remains unchanged after the dynamic potential energy is introduced. And based on DQN algorithm, this paper introduces a reward shaping mechanism. At the same time, we classify the actions of the agent and set the progress towards the goal as a positive reward, and the rest as a negative or no reward. When the agent takes the next action according to the action value function, it will preferentially select the action that develops towards the goal progress.

The reward function based on dynamic potential energy can be expressed as:

$$F(s,t,s',t') = y\varphi(s',t') - \varphi(s,t), \tag{7}$$

And the formula proves that the Q value of the shaping function is independent of the action selection, and the optimal strategy remains unchanged:

$$\begin{gathered} Q_{i,\varphi}^{*}(s,a) = \sum_{s'} Pr(s'|s,a)U_{i,\varphi}(s') = \sum_{s'} Pr(s'|s,a)(U_i(s') - \varphi(s,t)) = \\ \sum_{s'} Pr(s'|s,a)U_i(s') - \sum_{s'} Pr(s'|s,a)\,\varphi(s,t) = Q_i^{*}(s,a) - \varphi(s,t), \end{gathered} \tag{8}$$

Where F represents the additional reward function, which is the state value after the action is taken at the next moment,and the formula indicates that both F and will change dynamically with time.

This paper adds the reward shaping on DQN. And the following is the code, as Table 1 shows.

Table 1. DQN with reward shaping

Algorithm 1
1.Initialize replay memory to capacity
2.Initialize action-value function Q with random weights
3.Initialize target action-value function with weights
4.For episode=1, X do
5. Update the position of UAV
6. For step=1, M do
7. Select a action with softmax policy
8. Set the reward based on the step
9. Implement and move on to the next state
10. Store transition

 doi:10.20944/preprints202408.0979.v1

11. Set
12. Gradient descent policy to loss function to improve
13. Every K step reset
14. End for
15. End for

3.2. Experiment

Our experiments were performed on a computer with an Intel Core i7-8700K CPU and an NVIDIA GeForce RTX 3060. The experiments were run on the Windows system, while the machine learning-related components made use of PyTorch. In this paper, the algorithm was designed and simulated based on the environment of Atari from Open AI. We used Open AI's Atari platform to build a new experimental environment. Only one UAV in the experiment was able to move left and right at different speeds and was able to explore different areas.

In the experiment, baseline3 of pytorch was used as the experimental baseline, and the DQN algorithm was compared with the DQN algorithm adding reward shaping. The hyperparameter settings are shown in the following Table 2:

Table 2. Number of hyper parameter

Hyper parameter	Number
Seed	25
entry 2	data
Learning_rate	e-4
Grad_clipping_value	5
Replay_buffer_size	1000000
Batch_size	32
Gamma	0.98

In this experiment, 2000000 rounds of training are set for the UAV agent. The comparison results between the improved algorithm with the reward shaping and the baseline algorithm are shown in Figure 3.

 doi:10.20944/preprints202408.0979.v1

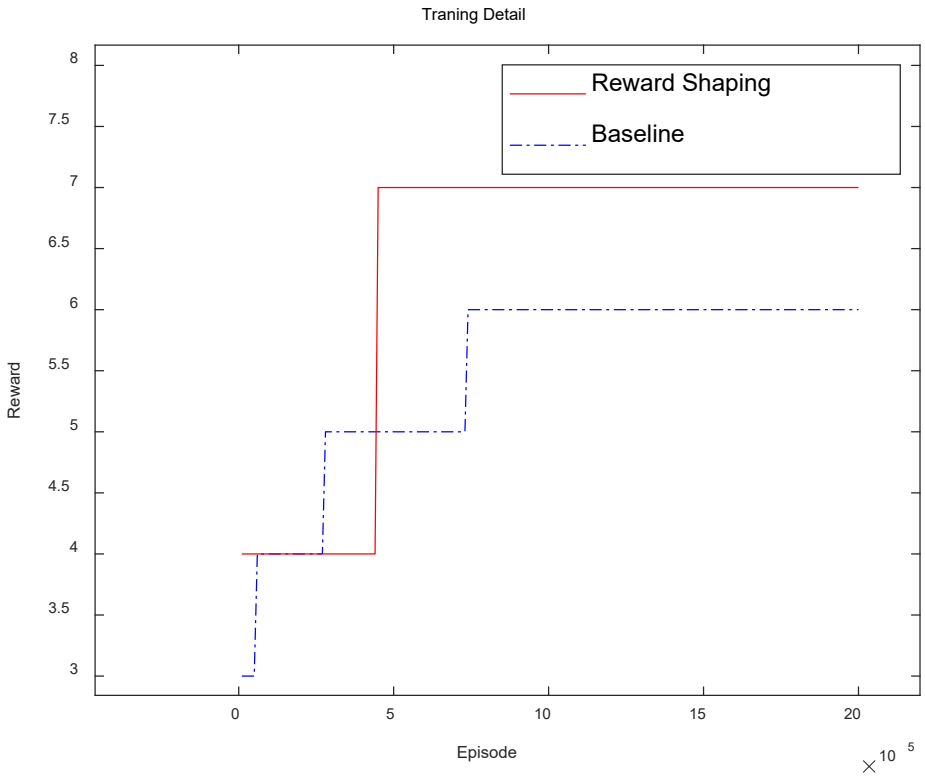

Figure 3. Training Reward.

As shown in Figure 3, after 200 thousand training sessions, the final DQN with reward function can obtain higher reward than before. This means that the improved algorithm is able to train the UAV agent faster, make the agent adapt to the new environment, and get the reward quicker. In summary, compared with the baseline algorithm and the traditional DQN algorithm, the final DQN with reward function has better performance and convergence,and can enable the UAV agent to obtain better scores in shorter training rounds.

4. Conclusions and Future Work

Aiming at the problem of using UAV as communication node in emergency situation, this paper constructs virtual environment through abstract modeling.And at the same time, based on DQN and combined with dynamic reward function shaping method, the reward function is set to improve.At the same time, the hyperparameters aremodified repeatedly according to the experimental results to improve the performanceof the algorithm. Experiments show that this algorithm can optimize the performanceof the original algorithm, shorten the training time, improve the training efficiency, and provide a reference for the autonomous location selection of UAV under emergency conditions.

A high-performance intelligent decision-making approach for wargames that still has some limitations were presented in this paper. This paper does not evaluate the merits and demerits of the reward function, and only verifies the positive effect of the reward function on improving the training efficiency, without verifying whether it will affect the final training effect of the agent. So the next step is to optimize the shaping method of dynamic reward function and make reasonable design for specific problems.

References

1. Liang, Zhang, F. Qiang, and A. Nirwan."3-D Drone-Base-Station Placement With In-Band Full-Duplex Communica-tions." IEEE Communications Letters PP(2018):1-1.
2. Wang Tianzhi. Research on UAV depl-oyment and path planning for Emergency Communication [D]. Beijing Univer-sity of Posts and Telecommunications,2022.

 doi:10.20944/preprints202408.0979.v1

3. YAN Luhong, Guo Wenpu, Xu Dongh-ui, Yang Haiyu. Application Research of Computers,2022, vol. 39 (1): 226-230,235.
4. Yin Changsheng, Yang Ruopeng, Zhu Wei, Zou Xiaofei. Emergency commun-ication network planning method basedon deep reinforcement learning [J]. Systems Engineering and Electronics,2020,vol. 42 (9): 2091-2097 .
5. Chen Haoran, Zhu Wei, Yu Sheng. E-mergency communication network plan-ning method based on deep reinforce-ment learning [J]. Command, Control and Simulation,2023, vol. 45 (1): 150-156.
6. Jiangbin Lyu;Yong Zeng;Rui Zhang;Te-ng Joon Lim.Placement Optimization ofUAV-Mounted Mobile Base Stations[J].IEEE Communications Letters,2017,Vol.21(3): 604-607.
7. Andrew Y. Ng;Daishi Harada;Stuart R-ussell.Policy invariance under reward t-ransformations: theory and application to reward shaping[A].International Conference on Machine Learning[C],1999.
8. Dong Yunlong. Research and Applicati-on of reinforcement learning based on Reward Shaping [D]. Huazhong Unive-rsity of Science and Technology,2022.
9. Yu Fei, Hao Jianguo, Zhang Zhongjie. Action exploration strategy in reinforc-ement learning based on action probab-ility [J]. Journal of Computer Applicat-ions and Software,2023, vol. 40 (5): 184-189,226.
10. Shi Hongyuan. Research on DQN (Deep Q-Network) algorithm in complex e-nvironment [D]. Nanjing University of Information Science and Technology,2023.
11. WU Jinjin. Research on Overestimationof value function for DQN [D]. Sooc-how University,2020.
12. Yang Weiyi, Bai Chenjia, CAI Chao, Zhao Yingnan, Liu Peng. Sparse rewardproblem in deep reinforcement learning[J]. Computer Science,2020, vol. 47 (3):182-191.
13. Liu Hongyan. Research on UAV com-munication trajectory Optimization bas-ed on deep reinforcement learning [D].Nanchang University,2023.
14. Li Qiru, Geng Xia. Robot path Planni-ng based on improved DQN Algorithm[J]. Computer Engineering,2023,(12): 111-120.
15. Yang D. Research on reward strategy techniques of deep reinforcement learni-ng for complex confrontation scenarios[D]. National University of Defense T-echnology,2020.
16. Niu Songdeng. Research on Student Motivation Based on Reinforcement L-earning [D]. University of Electronic Science and Technology of China,2022.