Deep Reinforcement Learning Based Resource Allocation for LoRaWAN

Aohan Li
Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan
aohanli@ieee.org

2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall) | 978-1-6654-5468-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VTC2022-Fall57202.2022.10012698

***Abstract*—It is predicted that the number of Internet of Things (IoT) devices will be more than 75 billion by 2025, where a large portion of IoT devices will be long-range (LoRa) powered by batteries. The battery lifetime limitations and the spectrum shortage have been the main problems in realizing LoRa wide area network (LoRAWAN) for the devices in hard-to-reach areas. The dynamic spectrum access technique has gained tremendous research interest as a promising paradigm due to its outstanding performance in improving spectrum efficiency. How to realize intelligent resource allocation (RA) to avoid collisions among IoT devices with low energy consumption is an important problem in LoRaWAN. However, either synchronization and prior information estimation, such as channel state information (CSI), are required, or the energy consumption of LoRa devices is not considered in related work, which may decrease the energy efficiency of the LoRa devices. In addition, the necessary prior information may be challenging to obtain in future networks. To address these issues, we propose a deep Q learning-based RA (DQLRA) method for LoRaWAN. In our proposed method, the gateway (GW) trains the deep neural network (DNN) only based on the transmission state, i.e., transmission failure or success, and the corresponding device number of each LoRa device. Then, each LoRa device can make decisions based on its device number and ACK or NACK information using the trained DNN. Synchronization and prior information estimation are not required in our proposed method, which may improve the energy efficiency of IoT devices. Simulation results show that the proposed method can achieve the optimal frame success rate (FSR) in most scenarios.**

***Index Terms*—LoRaWAN, Resource Management, Deep Reinforcement Learning, Energy Efficiency**

I. Introduction

According to statistic [1], there will be more than 75 billion Internet of Things (IoT) connected devices in use by 2025. A large portion of the IoT devices will be scattered over a wide geographic area and powered by battery [2]. Lower power wide area network (LPWAN) technologies have emerged to support such devices. The most popular LPWAN techniques are long-range wide area network (LoRaWAN), Sigfox, and NB-IoT [3]. Among these LPWAN techniques, LoRaWAN has attracted particular interest due to its ultra-low-power consumption, and long-range communication [4]. However, battery lifetime limitations have been one of the main problems in realizing the vision of a global LoRaWAN for the devices located in hard-to-reach areas [5]. In addition, the accompanying shortage of spectrum resources has also attracted widespread attention. Dynamic spectrum access techniques have gained tremendous research interest as promising paradigms due to their outstanding performance in improving spectrum efficiency [6]. However, how to realize intelligent resource allocation (RA) to support better communication performance with low energy consumption for LoRa devices is a significant problem that needs to be solved [8].

There are several works on RA for LoRaWAN [6], [9]. In [9], a joint optimization problem that optimizes channel and dynamic power allocation is formulated for uplink energy harvesting LoRaWAN. A matching theory-based channel allocation algorithm and a Markov decision process (MDP) based power allocation algorithm are proposed to solve the formulated problem. Numerical results demonstrate that the proposed algorithm can achieve near-optimal performance. However, real-time channel state information (CSI) is required. It is difficult, or even impossible, to obtain such accurate information in future networks due to the system's large-scale, ultra-dense, and massive heterogeneity of the system [10]. In addition, local optimal solutions rather than global ones will be obtained due to the non-convex optimization problem.

Deep reinforcement learning (DRL) has been widely used to address the RA problems since DRL enables network controllers to solve complex network optimization problems [11]–[13]. In [6], a novel system model for downlink LoRa wireless networks powered by both the energy harvesting source and the grid is investigated. The grid energy source compensates for the randomness and intermittency of the harvested energy. The minimize problem of grid energy cost is formulated. A deep reinforcement learning (DRL) based method is proposed to assign spectrum and energy resources. However, the energy consumption of the LoRa devices is not well considered. In addition, channel coefficient estimation is necessary to perform the proposed method, which may bring additional computation and energy consumption. To improve spectrum and energy efficiency with low computational complexity for LoRa devices, we propose a deep Q learning-based RA (DQLRA) method to allocate spectrum and spreading factor (SF) for LoRa devices. The SF is another crucial factor for LoRa devices, which determines the chirp rate. The time on air and the ability to resist noise are related to the value of the SF [7].

The rest of the paper is organized as follows. Section II introduces the system model and problem formulation. Section III presents the proposed DQLRA method. Section IV demonstrates the simulation results. Finally, Section V concludes the paper.

978-1-6654-5468-1/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Southeast University. Downloaded on February 06,2023 at 14:08:11 UTC from IEEE Xplore. Restrictions apply.

II. System Model and Problem Formulation

We focus on the uplink transmission of a LoRaWAN consisting of one GW and N LoRa devices denoted by $\mathcal{N} = \{1, 2, ..., n, ..., N\}$. The LoRa devices are randomly located around the GW in a circle covered by the network. We assume that each LoRa device in the LoRaWAN has a rechargeable battery with finite capacity. It can harvest energy from the ambient environment, such as solar power, thermal energy, wind, and vibrations. Let B_T be the overall bandwidth owned by the GW, divided into M blocks with a size of B, denoted as $\mathcal{C} = \{C_1, C_2, ..., C_m, ..., C_M\}$. We consider that each channel can be accessed simultaneously by at most D devices with different SFs, which means that multiple frames can be transmitted in the LoRaWAN simultaneously, as long as the LoRa devices are with different SFs [6]. Denote $\mathcal{SF} = \{SF_1, SF_2, ..., SF_s, ..., SF_S\}$ as the SF set, where SF_s denotes the s-th SF and S is the number of SF values, which equals to 6 in LoRaWAN generally. The values of the elements in the SF set corresponds to 7, 8, 9, 10, 11, 12.

We use $\alpha_{m,s,n}(t)$ to indicate whether channel C_m and SF SF_s are assigned to LoRa device n at time t, that is:

$$\alpha_{m,s,n}(t) = \begin{cases} 1, & \text{when LoRa device } n \text{ uses } C_m \text{with SF } SF_s \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

Denote $r_{m,s,n}(t)$, which indicates if the transmission is successful for LoRa device n using channel C_m and SF_s at time t. Its expression is given as follows:

$$r_{m,s,n}(t) = \begin{cases} 1, \text{the transmission is successful.} \\ 0, \text{otherwise.} \end{cases} \tag{2}$$

Denote $l_{m,s,n}(t)$, which indicates if LoRa device n attempts to transmit messages using channel C_m with SF_s at time t, and it is given as follows:

$$l_{m,s,n}(t) = \begin{cases} 1, \text{transmission attempted} \\ 0, \text{otherwise.} \end{cases} \tag{3}$$

At time t, the FSR of the LoRaWAN can be given by

$$P(t) = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} r_{m,s,n}(t)}{\sum_{n=1}^{N} \sum_{m=1}^{M} l_{m,s,n}(t)}, \tag{4}$$

which provides the ratio of the successful and attempted transmission times at time t for the LoRaWAN.

This paper aims to maximize the FSR of the LoRaWAN by selecting the appropriate channel and SF. The optimization problem can be formulated as:

$$(\mathbf{P1}) \max_{\{m,s\}} \sum_{t=1}^{T} P(t) \tag{5a}$$

$$\text{s.t.} \quad C1: \sum_{n} \alpha_{m,n,j}(k) \leq D \quad \forall m \tag{5b}$$

$$C2: \sum_{m} \alpha_{m,n,j}(k) \leq 1 \quad \forall n \tag{5c}$$

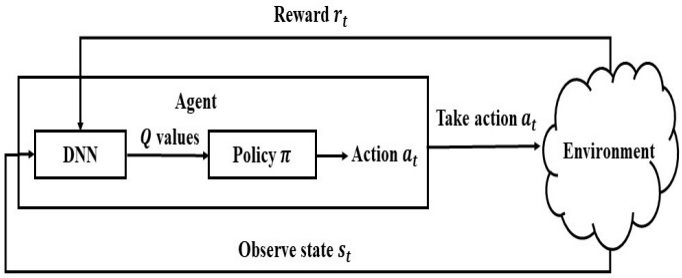

Fig. 1. The structure of the DRL techniques

where $C1$ indicates that at most D LoRa devices can use the same channel simultaneously, and $C2$ restricts that each LoRa device can be assigned only one channel simultaneously. T is the length of time for making decisions.

Although problem **P1** can be solved by the conventional optimization methods, the computational complexity is large in large scale LoRaWAN. In addition, the necessary prior information for optimization is often challenging to obtain in advance. Hence, we design a DQLRA scheme to reduce the computational complexity and avoid prior information requirements.

III. DQLRA SCHEME

In this section, we develop the DQLRA scheme based on the DQL technique. To well understand our proposed DQLRA scheme, the DQL technique is first introduced. Then, the DQLRA scheme is presented.

A. DQL Technique

DQL technique is one of the deep reinforcement learning (DRL) techniques. DRL is the area of machine learning that deals with sequential decision-making [5]. The DRL problem can be formalized as an agent that has to make decisions in an unknown environment to optimize a given notion of cumulative rewards [8]. The general structure of the DRL techniques is illustrated in Fig. 1. At each decision epoch t, the agent selects an action based on the Q values obtained by DNN using policy π. The DNN is being used as a function approximator of the Q function, which can take state information input as a vector and learn to map them to Q-value for all possible actions. The agent can obtain a reward while the environment turns to a new state by executing the selected action. Then, the DNN is updated based on the obtained reward, the q value chosen using policy π corresponding to the current state, and the maximum Q value corresponding to the next state. Fig. 2 shows the structure of the DNN. In Fig. 2, the Q-values for all possible actions at epoch t, i.e., $Q(S_t, A_1), ..., Q(S_t, A_l), ..., Q(S_t, A_L)$, can be mapped by the DNN from the state information at epoch t, i.e., S_t, where $A = \{A_1, ..., A_l, ..., A_L\}$ is the action set.

The main objective of DQL is to minimize the distance between $Q(s, a; w_t)$ and temporal difference target y_t^{tar}, where $Q(s, a, w_t)$ denotes the Q-value for the state s and action a

Authorized licensed use limited to: Southeast University. Downloaded on February 06,2023 at 14:08:11 UTC from IEEE Xplore. Restrictions apply.

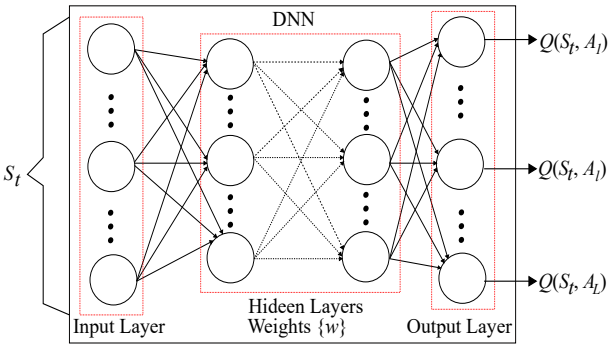

Fig. 2. Structure of the DNN

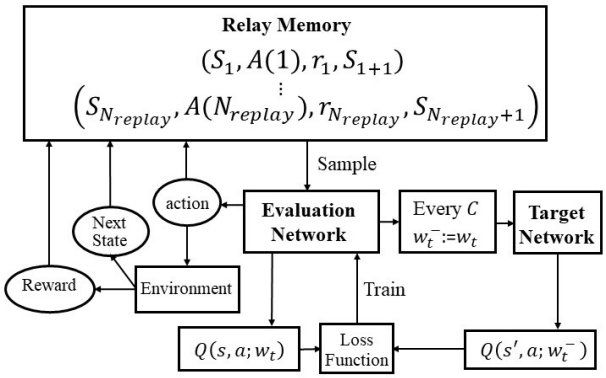

Fig. 3. The Sketch of the DQL

pair at the t^{th} iteration. The objective can be expressed as the minimization of the square loss function, as shown below.

$$L_t(w_t) = \mathbb{E}_{s,a,r,s'}[(y_t^{tar} - Q(s, a; w_t))^2]. \tag{6}$$

In DQL,

$$y_t^{tar} = [r + \gamma max_{a'} Q(s', a'; w_t^-)|S_t = s, A(t) = a], \tag{7}$$

where S_t and $A(t)$ denote the current state and action, respectively. s' and a' denote the state and action at the next epoch, respectively. r denotes immediate reward while γ denotes the discount factor of the future rewards. In DQL, $Q(s', a'; w_t^-)$ and $Q(s, a; w_t)$ are estimated separately by two different neural networks, which are often called target and evaluation networks. The sketch of the DQL is shown in Fig. 3. The weight parameters w_t and w_t^- belongs to evaluation network and target network, respectively. The parameters w_t of the evaluation network are updated every epoch based on loss function $L_t(w_t)$. The parameters w_t^- of the target network are updated every $C \in \mathbb{N}$ iterations with the following assignment: $w_t^- = w_t$. An important feature added to DQL is replay memory, which keeps all information that are the sets of tuples $< S_t, A(t), r_t, S_{t+1} >$ for the last $N_{replay} \in \mathbb{N}$ time steps termed as mini-batch. The experience information is collected following an ϵ-greedy policy, which will be presented later. The buffer of the replay memory is fixed. As the new experiences get stored, the old experiences get removed. The data for updating parameters of the evaluation network is selected randomly within the replay memory, which can help the network to learn from a variety of data instead of just formalizing decisions based on immediate experiences. In addition, one mini-batch update has less variance than a single tuple update. Consequently, it allows for a more significant update of the parameters, which has an efficient parallelization of the algorithm.

B. DQLRA Scheme

In the DQLRA scheme, the GW and each LoRa device are termed an agent during the training phase and test phase, respectively. The things related to the LoRaWAN are regarded as an environment. An action can be selected based on the Q values obtained by the DNN using the ϵ-greedy algorithm. A reward can be obtained while the environment turns to a new state by executing the selected action. The state of the DQLRA at time t is set as the LoRa device number and the corresponding indicator of the transmission state after performing the previously selected action, i.e., $r_{m,s,n}(t-1)$. The action set of the DQLRA method consists of the channel and SF sets. The policy for taking actions used in the DQLRA, i.e., ϵ-greedy algorithm, can be expressed as

$$A(k) = \begin{cases} \underset{A(k)\in A}{\arg\max}\, Q(S_k, A(k)), & \text{with probability } 1-\epsilon \\ \underset{A(k)\in A}{\text{random}}(A(k)), & \text{with probability } \epsilon \end{cases} \tag{8}$$

$A = \{A_1, A_2, ..., A_{MS}\}$ in the DQLRA scheme. The main idea of the ϵ-greedy algorithm is to take action resulting in the highest Q value with probability 1-ϵ while randomly selecting one action with probability ϵ. The RMSProp optimization algorithm updates the DNN in the DQLRA scheme. It utilizes the normalized gradients and is very robust. Additionally, it can deal with stochastic objectives very nicely, making it applicable to mini-batch learning [14]. The loss function for training the DNN in the DQLRA scheme is given by

$$Loss(w_k) = \mathbb{E}\left[\sum_{S_k, A(k)\in D} \left(y_k^{tar} - y_k^{eva}\right)^2\right], \tag{9}$$

where D is a data set that storages the state-action pairs. For the DQLRA scheme, y_k^{tar} and y_k^{eva} are defined as Eq. (7) and $Q(s, a; w_k)$. The reward function R at time t is defined as follows:

$$R(t) = \begin{cases} 1, & \text{if transmission is successful.} \\ -1, & \text{otherwise.} \end{cases} \tag{10}$$

The received reward is related to the selected actions. Two cases are presented in the chosen actions, which is clearly shown in Eq. (10). The reward is 1 if the chosen action performs the successful transmission. Otherwise, the reward is -1.

Authorized licensed use limited to: Southeast University. Downloaded on February 06,2023 at 14:08:11 UTC from IEEE Xplore. Restrictions apply.

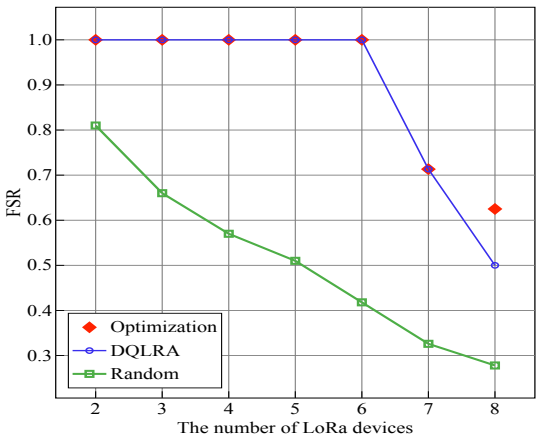

Fig. 4. Performance Evaluation in FSR

IV. Performance Evaluation

In this section, we evaluate the performance of our proposed DQLRA method and compare it to optimal solutions and random selection in terms of FSR. In the random selection method, channel and SF are selected randomly. In the simulation, the number of actions is set to 6. That is, the product of the number of channels and that of the SFs is 6. At this time, the LoRaWAN supports the communications of at most six users simultaneously. γ is set as 0.9. ϵ is set as 1 initially and decreases with the learning process. Specifically, its value decrements by 0.05 per step. RMSProp optimization algorithm with a learning rate of 0.01 is used to update the DNN in the simulation. The number of epochs is set to 5000. The learning process is stopped if the FSR is 1 for 100 times in a row, even if the number of learning times does not reach 5000.

Simulation results are shown in Fig. 4. The FSR of the proposed DQLRA and the random selection methods are the average FSR of 100 tests. Since the settings of the LoRaWAN in the simulation supports the communications from at most six users simultaneously, the optimal FSR is 1 when the number of LoRa devices is not larger than 6. When the number of LoRa devices is larger than 6, the collision between at least two LoRa devices happens. Hence, the optimal FSR are 0.71428 and 0.625 when the numbers of LoRa devices are 7 and 8, respectively. From Fig. 4, we can see that our proposed DQLRA method can achieve a much higher FSR than the random selection method. In addition, our proposed DQLRA algorithm can achieve the optimal FSR except for the case when the number of devices is 8. The reason may be that the system complexity increases as the number of devices increases. More than 5000 times of training is required to train the DNN. Therefore, it is necessary to consider increasing the number of learning times and adjusting other learning parameter settings.

V. Conclusion

In this paper, we proposed a DQLRA method for LoRaWAN. In our proposed scheme, GW trains the DNN based on whether the transmission is a success and the device number of the LoRa device. Then, each device decides its access channel and SF based on the ACK information and device number using the trained DNN. Prior information estimation and synchronous are not required compared to existing related work. Hence, the energy efficiency of IoT devices may be improved. Simulation results show that our proposed scheme can achieve optimal FSR in most scenarios.

References

[1] "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025," Accessed: June. 13, 2022. [Online]. Available: https://www.statista.com/statistics/471264/ iot-number-of-connected-devices-worldwide/

[2] H. Jiang, D. Qu, J. Ding, Z. Wang, H. He and H. Chen,"Enabling LPWAN massive access: grant-free random access with massive MIMO," *IEEE Wireless Commun. Mag.*, DOI: 10.1109/MWC.102.2100276, May 2022.

[3] D. Magrin, M. Capuzzo, A. Zanella, L. Vangelista, and M. Zorzi,"Performance analysis of LoRaWAN in industrial scenarios," *IEEE Trans. Indus. Info.*, vol. 17, no. 9, pp. 6241-6250, Sept. 2021.

[4] C. G. Hidalgo, J. Haxhibeqiri, B. Moons, J. Hoebeke, T. Olivares, F. J. Ramires, and A. F. Caballero,"LoRaWAN scheduling: From concept to implementation," *IEEE Trans. Indus. Info.*, vol. 8, no. 16, pp. 12919-12933, Aug. 2021.

[5] C. Delgado, J. M. Sanz, C. Blondia, and J. Famaey,"Batteryless LoRaWAN communications using energy harvesting: modeling and characterization," *IEEE Internet of Things J.*, vol. 8, no. 4, pp. 2694-2711, Feb. 2021.

[6] R. Hamdi, E. Baccour, A. Erbad, M. Qaraqe and M. Hamdi,"LoRa-RL: Deep reinforcement learning for resource management in hybride energy LoRa wireless networks," *IEEE Internet Things J.*, vol. 9, no. 9, pp. 6458-6476, May 2022.

[7] A. Li, I. Urabe, M. Fujisawa, S. Hasegawa, H. Yasuda, S.-J. Kim, and M. Hasegawa, "A Lightweight Transmission Parameter Selection Scheme Using Reinforcement Learning for LoRaWAN," *arXiv preprint arXiv:2208.01824*, Aug. 2022

[8] Z. Shi, X. Xie, H. Lu, H. Yang, J. Cai and Z. Ding,"Deep reinforcement learning-based multidimensional resouce management for energy harvesting cognitive NOMA communications," *IEEE Trans. Commun.*, vol. 70, no. 5, pp.3110-3125, May 2022.

[9] X. Lin, Z. Qin, Y. Gao, and J. A. McCann,"Resource allocation in wireless powered IoT networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4935-4945, June 2019.

[10] A. Alwarafy, M. Abdallah, B. S. Ciftler, A. Al-fuqaha, and M. Hamdi,"The frontiers of deep reinforcement learning for resource management in future wireless HetNets: techniques, challenges, and research directions," *IEEE Open J. Comput. Soc.*, vol. 3, pp. 322-365, Feb. 2022.

[11] H-S. Lee, D-Y. Kim and J-W. Lee,"Radio and energy resource management in renewable energy-powered wireless networks with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, DOI: 10.1109/TWC.2022.3140731, Jan. 2022.

[12] S. Guo and X. Zhao,"Deep reinforcement learning optimal transmission algorithm for cognitive internet of things with RF energy harvesting," *IEEE Trans. Cognitive Commun. Netw.*, vol. 8, no. 2, pp. 1216-1227, June 2022.

[13] L. P. Qian, C. Yang, H. Han, Y. Wu and L. Meng,"Learning driven resource allocation and SIC ordering in EH relay aided NB-IoT networks," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2619-2623, Aug. 2021.

[14] Geoffrey Hinton, Neural Networks for Machine Learning, online course. https://www.coursera.org/learn/neural-networks/home/welcome [Accessed 1 July 2022]

[15] Q. Zhang and S. A. Kassam, "Finite-state markov model for rayleigh fading channels," *IEEE Internet Things J.*, Vol. 47, No. 11, PP. 1688-1692, Nov. 1999.

Authorized licensed use limited to: Southeast University. Downloaded on February 06,2023 at 14:08:11 UTC from IEEE Xplore. Restrictions apply.