Dynamic Channel Allocation for Multi-UAVs: A Deep Reinforcement Learning Approach

†Xianglong Zhou, †*Yun Lin, †Ya Tu, ‡Shiwen Mao, and †Zheng Dou
†College of Information and Communication Engineering, Harbin Engineering University, Harbin, P.R.China
‡Department of Electrical and Computer Engineering, Auburn University Auburn, AL, U.S.A.
*Corresponding Author: linyun_phd@hrbeu.edu.cn

***Abstract*—It has been recognized that fixed spectrum and channel allocation will lead to waste of spectrum resources when multiple agents communicate at the same time. Dynamic allocation of channels is proposed to maximize the utilization of spectrum resources. In the environment of multiple unmanned aerial vehicles (UAVs), it is necessary to ensure that each UAV can communicate successfully without interfering with other UAVs. Dynamic allocation of channels plays an important role in such systems. In this paper, we propose a dynamic channel allocation scheme based on deep reinforcement learning for multi-UAV systems. A slotted time system is used by all the UAVs. Different from the traditional method, the occupancy of each channel is scanned first in each time slot. Then a channel will be selected for data transmission, with feedback from the environment when the transmission is over. The proposed channel allocation scheme incorporates a long short-term memory (LSTM) into the deep reinforcement learning framework, to better learn from the past experience and better adapt to the the highly dynamic environment in a multi-UAV system. The experimental results show that compared with the traditional reinforcement learning method (Q-learning and Deep Q Network (DQN)), the proposed method achieves faster convergence and better performance with respect to average collision rate, average reward, and average successful communication rate.**

***Index Terms*—Multi-Unmanned Aerial Vehicles (UAV), Dynamic Channel Allocation, Deep Reinforcement Learning, Long short-term memory (LSTM).**

I. Introduction

With the proliferation of communication devices, the communication environment has nowadays become increasing complex. The limited spectrum resources make the number of channels highly limited. In the traditional fixed channel allocation mode, the channels cannot be used optimally, and the spectrum channel efficiency is usually low. To this end, dynamic channel allocation (DCA) is proposed to solve such problems [1].

Unmanned aerial vehicles (UAVs) play an important role in disaster relief, air search and rescue, and express delivery [2]. In a system consisting of multiple UAVs, successful communications and reasonable task assignment are particularly important [3]. Dynamic channel allocation is mainly to solve the problem of optimal communications under limited available channels for a group of cooperative UAVs [4], [5]. In addition, the UAV usually has high mobility, so the channel allocation algorithm is required to gain the optimal allocation strategy in a short time [6].

As a research hotspot in the field of machine learning, reinforcement learning aims to acquire the environment model by self-adaptive learning without the need for training, and to achieve the optimal action strategy when facing a time-varying environment [7]. The basic idea of reinforcement learning is to maximize the cumulative reward values of agents from the environment, so as to obtain the optimal strategy to complete a task [8]. As the tasks in the real world become more and more complex, simple reinforcement learning may not be adequate. The integration of deep learning and reinforcement learning offers a more effective solution, i.e., deep Q-network (DQN), which incorporates a deep neural network to estimate the Q value and the relationship between the predicted value and the real value to find the optimal strategy, with the advantage of mitigating the curse of dimensionality problem as data gets larger [9], [10].

Dynamic channel access can maximize the use of spectrum resourcesby minimizing interference, which is similar to dynamic spectrum access (DSA). In [11], the authors propose a DSA algorithm for LTE cellular systems. It combines distributed enhanced learning and standardized inter-cell interference coordination signaling in the LTE downlink. In [12], the authors propose a channel spectrum access algorithm based on online synchronous Q-learning to avoid channel congestion in cognitive radio networks. In the research of multi-agent allocation, a non-cooperative shared spectrum allocation algorithm is proposed for multi-user and multi-channel cognitive radio systems [13]. A special kind of recursive neural network using reservoir computing improves the slow convergence of Q-learning and the learning efficiency [14].

In order to improve the network performance of multi-hop cognitive radio networks, three schemes are proposed in [15]: The goal of the reinforcement learning method with average Q value is to select the route with the highest Q value and the shortest available channel time, respectively. Two reinforcement learning methods are introduced for sensing and access, respectively. Compared with traditional reinforcement learning methods, the number of sensors used for sensing is reduced, and the throughput and energy efficiency are both improved [16]. The combination of reinforcement learning algorithm and game theory can help to better solve the problems in complex environments. In [17], the authors propose a spectrum access algorithm, which can better adapt to the channel allocation of the multi-agents cooperative work. For

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

UAVs, the algorithms should have fast computing power to ensure that the optimal strategy can be quickly computed.

In this paper, we propose a dynamic channel allocation algorithm for UAVs. The algorithm is focused on channel allocation for multiple UAVs, and aims to achieve fast computation of competitive channel allocations for multiple UAVs in highly dynamic environment.

The main contributions of this paper are:

- In the proposed approach, a slotted time system is used by all the UAVs. Different from the traditional method, the occupancy of each channel is scanned first in each time slot. Then a channel will be selected for data transmission, with feedback from the environment when the transmission is over.
- The proposed channel allocation scheme incorporates a long short-term memory (LSTM) into the deep reinforcement learning framework, to better learn from the past experience and better adapt to the the highly dynamic environment in a multi-UAV system.

The experimental results show that compared with the traditional reinforcement learning method (Q-learning and Deep Q Network (DQN)), the proposed method achieves faster convergence and better performance with respect to average collision rate, average reward, and average successful communication rate.

In the rest of this paper, we introduce the system model and assumptions in Section II, and the proposed dynamic channel allocation scheme in Section III. We evaluate the performance of the proposed scheme in Section IV, and concludes this paper in Section V.

II. System Model and Assumptions

We consider the typical application scenario where multiple UAVs are deployed, as shown in Fig. 1. Without loss of generality, we assume that the channels are available for the UAVs only, i.e., the primary users are not involved (due to effective spectrum sensing or policy-based spectrum sharing). We focus on the dynamic channel allocation strategy among the UAVs in this paper. To be consistent with the practical situation, we consider the case where the number of UAVs, N, is larger than that of available channels, C. The UAVs need to send information to the receiving devices. To simulate the practical situation, we also assume that the number of receiving devices is the same as that of available channels.

In this paper, we focus on the channel allocation strategy, and assume the channels are reliable (i.e., ignoring the channel fading/shadowing effect, which, if needed, can be easily modeled by an additional communication failure probability and a modification of (10)). We only focus on the down link of UAVs to allocation the channel. We assume that all the UAVs participate in the training process, which can better characterize the real situation. Because the UAV is a highly dynamic agent, transmission time and transmission efficiency are used to measure the quality of a strategy. We include the distance between UAV i and receiving terminal j in the reward function, which is given by

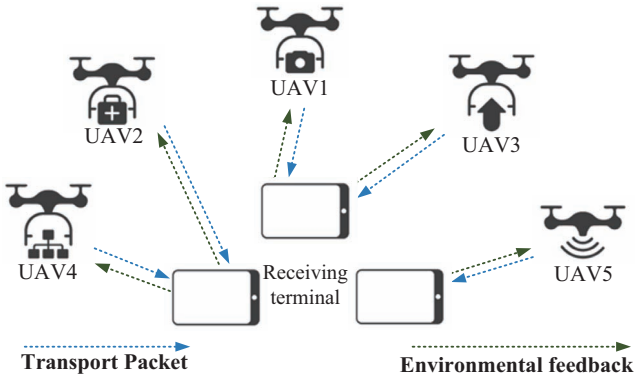

Fig. 1. System environment model: an example of 5 UAVs sharing 3 channels. The channel model is not introduced in the system model.

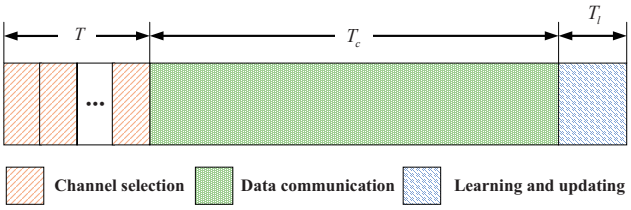

Fig. 2. Time slot model: channel scanning is carried out in the initial T, information transmission is carried out in T_c, and feedback information is received in T_l.

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{1}$$

Where $i = 0, 1, ..., N$ and $j = 0, 1, ..., C$, (x_i, y_i) represents the position of UAV i, and (x_j, y_j) represents the position of receiving terminal j. The reward function will be defined in the next section.

In addition, we assume that each UAV operates in a slotted-time manner, as shown in Fig. 2. In each time slot, the channel state is scanned first in a period of T. The scanning mode we adopt is *traversal*, which can avoid collision as much as possible. If there are free channels after scanning, the UAV will transmit information in the next duration of T_c, and receive state feedback in the third period of T_l till the end of the time slot [18].

III. Dynamic Channel Allocation Based on Reinforcement Learning

In order to improve the convergence speed of the deep reinforcement learning algorithm, we introduce a long-term and short-term memory network to preserve the historical information of action and environmental feedback. Through learning the historical information, we can build a more rapid model of the environment to adapt to the highly dynamic characteristics of the UAV cluster environment.

A. Deep Reinforcement Learning

In the multi-UAV environment, it is challenging to obtain a large number of data samples for training, and the training process is time consuming and computtion intensive. To this end,

reinforcement learning provides an effective solution. With the reinforcement learning design, the UAVs interact with the environment without any prior knowledge. Then the UAVs will learn how to use the channel correctly and efficiently by the reward mechanism of reinforcement learning.

Q-learning is a special case of value-based updating in reinforcement learning. It constantly updates a Q-table by first choosing an action and then to obtain a reward from the environment. The Q-table can provide the optimal action strategy. However, as the data size gets larger, the curse of dimensionality makes it challenging to generate and retrieve data from the Q-table [19].

DQN is a method to address this problem. This algorithm can obtain Q values by training a neural network [14]. We need the correct Q value for actions a_i. The Q value is the same as that in Q-learning, which is defined as $Q(S)$. We also need an estimate of Q to update the neural network, which is predicted by the neural network as $Q(s, a_i)$. Then we select the action that has the maximum estimated Q value in exchange for a reward from the environment. The Q value of the state can be computed as

$$Q(s\_) = R + \gamma \cdot \max Q(s\_, a_i), \tag{2}$$

where R is the reward of the a_{i+1}, $\gamma \in [0, 1]$ is the discount rate. The neural network parameters are updated as follows.

$$NN_{new} \leftarrow NN_{old} + \alpha(Q(s) - Q(s\_)), \tag{3}$$

where $\alpha \in [0, 1]$ is the learning rate.

DQN has a memory library to learn from the previous experience. When the DQN is updated, we can randomly select some previous experiences from the library to learn from. Since the historical data samples are randomly extract from the library, the correlation between past experience samples is disrupted and the neural network updates can be more efficient.

B. Long Short-term Memory (LSTM)

Long short-term memory (LSTM) belong to the class of time recursive neural networks (RNN). It is effective to address the problems of vanishing gradient and exploding gradient in the process of long-sequence training [20]. The hidden layer of the original RNN has only one state, which is very sensitive to short-term input, and the gradient disappears or explodes due to the exponential function. LSTM was proposed by Hochreiter and Schmidhuber to solve this problem by adding a state c to the RNN to preserve long-term memory [21]. The LSTM network is illustrated in Fig. 3.

There are three inputs to LSTM: the present input value of the network, the previous output value, and the previous unit state. There are two outputs from LSTM: the present output value and the present unit state. In this paper, LSTM is mainly used to preserve historical observation data. Specifically, we use LSTM to solve the problem of circular neural network's weak ability to process long-term memory information. In the process of reinforcement learning, more historical information can help the UAV to learn the characteristics of the environment more quickly. LSTM networks can also predict environmental feedback from future UAV actions. This can speed up the convergence of the algorithm.

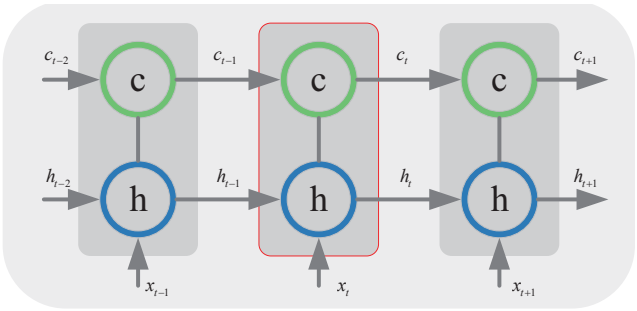

Fig. 3. The LSTM network structure diagram: input value of the current time x_t, output value of the previous time h_t, and unit state of the previous time c_t.

C. Algorithm Structure

Next we describe the UAV dynamic channel allocation scheme. In the multi-UAV environment, we first define the key elements of reinforcement learning, i.e., agent, action, state, reward function, and strategy. Channel allocation strategy is determined by the deep Q-value network, and the agent will gradually provide the optimal strategy through continuous interaction with the environment. In the multi-UAV context, the elements are defined as follows.

We define each individual UAV as an *agent*. There are two types of actions for each agent, i.e., access or no access. Then, according to the number of channels, the action form of UAV is defined as

$$a_t^n \in \{0, 1, ..., C\}, n \in \{0, 1, ..., N\}, \tag{4}$$

where C is the number of channels, which is the same as the amount of actions. a_t^n means the action of UAV n, $a_t^n = c$ means the UAV n access the channel c, $c \in \{0, 1, ..., C\}$. If $c = 0$, the UAV does not access a channel.

Since the drone operation is highly dynamic, we use the distance from the drone to the receiving terminal as the reward value from the environmental feedback. When two or more drones access the same channel, a *collision* occurs, at which point the reward value is a negative constant $-C$(In this paper, we set $-C = -1$). When the UAV does not access any channel, the reward function value is 0. In other general cases, the reward value is defined as $1 - \log_2\left(d_{ij}\right)$, where d_{ij} represents the distance from the UAV to the receiving terminal (given in (1)). The reward function is expressed as follows.

$$r = \begin{cases} -C, & \text{interference} \\ 0, & \text{no action} \\ 1 - \log_2\left(d_{ij}\right), & \text{otherwise.} \end{cases} \tag{5}$$

The structure we adopt for the algorithm an integration of LSTM and DQN, so the iterative updates of the Q value are

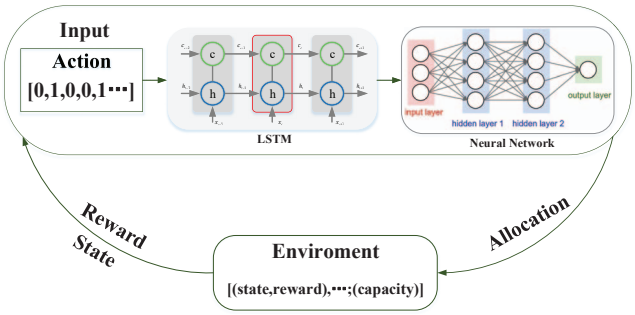

Fig. 4. The algorithmic framework, where LSTM is integrated with DQN to store the historical action and environment status of the UAV in the LSTM, and the output of the LSTM is used as input to the DQN; The environmental feedback includes channel occupancy, collision, and information transmission success.

similar to that in DQN, as

$$Q(s_t^n, a_t^n) \leftarrow \\ Q(s_t^n, a_t^n) + \alpha[r_{t+1}^n + \gamma \max_{a_{t+1}^n} Q(s_{t+1}^n, a_{t+1}^n) - Q(s_t^n, a_t^n)], \tag{6}$$

where $\alpha \in [0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discounted rate of reward. When Q value is updated, we choose the following action principles.

$$a_{t+1} = \begin{cases} \text{argmax}_a\, Q(s, a), & \text{if } 0 < \varepsilon \leq 0.9 \\ \text{Choose random action}, & \text{if } 0.9 < \varepsilon < 1. \end{cases} \tag{7}$$

In (7), $\varepsilon \in (0, 1)$ is a randomly generated number.

Environmental feedback includes channel occupancy status, collision of UAV access, and residual channel capacity. The proposed LSTM+DQN algorithm framework for dynamic channel allocation of the UAV fleet is illustrated in Fig. 4. The detailed procedure of the algorithm is presented in Algorithm 1.

Algorithm 1 The LSTM+DQN Algorithm

for episode $i = 1, 2, ..., M$ **do**
  for time-slot $t = 1, 2, ..., T$ **do**
    for each UAV in UAVs **do**
      Initialization of LSTM and DQN networks;
      Select actions according to (7);
      Obtain feedback from the environment on channel information and reward value r_{t+1};
      The state and action are input into the LSTM network, and the output of LSTM network is used as the input to the DQN network;
      The DQN computes the Q-estimation Q_{t+1}, and calculates Q-value according to the actual Q_t;
      Update the Q value according to (6) as:
      $Q_t \leftarrow Q_t + \alpha \left[r_{t+1} + \gamma \max_{a_{t+1}} Q_{t+1} - Q_t \right]$
    end for
  end for
end for
return channel allocation strategy

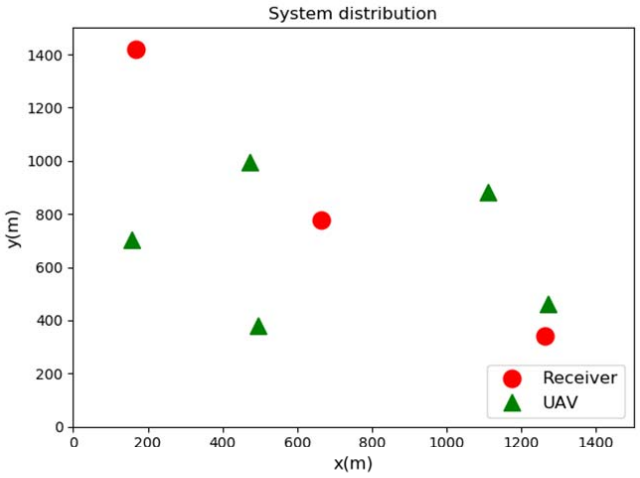

Fig. 5. Two-dimensional distributed view of UAVs and receiving terminals (5 UAVs are represented by green triangles and 3 receiving terminals are represented by red dots).

The average collision rate (ACR), the average reward (AR) and the average successful communication rate (ASCR) are used to evaluate the performance of the proposed algorithm. The ACR is defined as:

$$ACR = \frac{1}{M} \sum_{i=1}^{M} \frac{n_i}{C}, \tag{8}$$

where M is the total number of training steps, C is the number of channels, n_i is the number of collision channels during the i-th training step. The AR is defined as:

$$AR = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} r_j, \tag{9}$$

where N is the number of UAVs and r_j is the reward to the j-th UAV. The ASCR is defined as:

$$ASCR = \frac{1}{M} \sum_{i=1}^{M} \frac{k_i}{C}, \tag{10}$$

where k_i is the number of successful communication channels during the i-th training step.

IV. Simulation Evaluation and Discussions

A. Configuration

In this section, we present our simulation evaluation of the proposed dynamic channel allocation scheme for multi-USAV systems. Firstly, we simulate the distribution of UAVs and receiving terminals in the environment, and generate positions randomly in the network area. For the results reported in this section, there are five UAVs and three channels to be dynamically allocated to the UAVs in an area of 1.5km × 1.5km, as shown in Fig. 5. In each channel, there is a receiving terminal to communicate with the UAVs. This paper considers UAVs working in a uniform motion environment, without considering the impact of UAVs take-off and landing and acceleration.

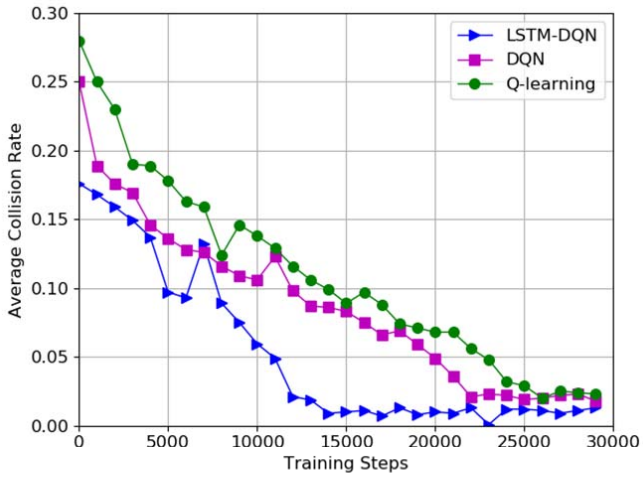

Fig. 6. The average collision rate achieved by Q-learning, DQN, and LSTM+DQN.

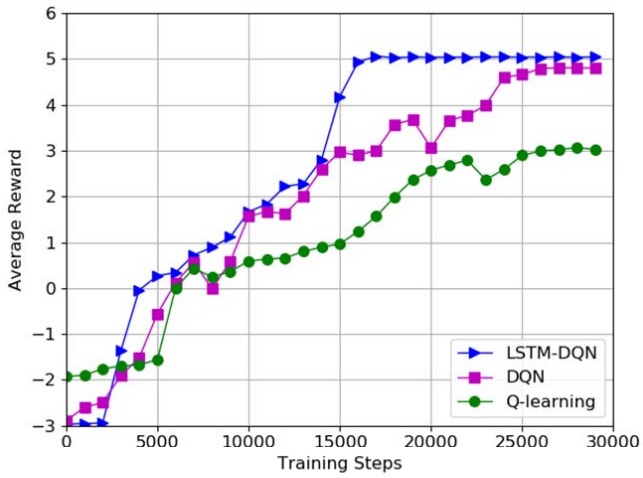

Fig. 7. The average reward achieved by Q-learning, DQN, and LSTM+DQN.

For the algorithm parameters, we choose to have 128 hidden layers for the LSTM network. The length of each stored historical sequence is 5. The number of hidden layers of the DQN neural network is also set to 128. The parameters of the LSTM+DQN, DQN, and Q-learning schemes are set as: the learning rate is $\alpha = 0.01$ and the discounted rate is $\gamma = 0.01$. The DQN and Q-learning algorithms are used as baselines to compare with our proposed LSTM+DQN algorithm.

B. Results and Discussions

To evaluate the performance of the three algorithms, during each iteration, we collect 2000 training samples to calculate the average collision rate (ACR), the average reward (AR) and the average successful communication rate (ASCR). The three evaluation index curves of the three algorithms are shown in Figs. 6, 7, and 8.

It can be seen from Fig. 6 that the ACR of the three algorithms are not much different at the beginning of the experiment. This is because the feedback information the agents get from the environment is not enough, while the algorithm is interacting with the environment. With the increased number of training, the UAVs have accumulated more and more information from the environment. From the figure we can see that the LSTM+DQN algorithm proposed in this paper converges after 15,000 training steps, which is the fastest among the three schemes. The other two algorithms start to converge after about 22,000 steps and 26,000 steps, respectively. The proposed algorithm converges for nearly 7000 steps faster than the traditional methods. This is because the LSTM network preserves more historical information, so that the neural network of the DQN can better predict the optimal channel access strategy. After reaching convergence, the average collision rate of the proposed method fluctuates between 0.01 and 0.025, while the collision rates of the other two algorithms are both between 0.02 and 0.05. The reason why the probability of collision will fluctuate is that there is a 10% possibility that the reinforcement learning will select random action to prevent getting stuck in a local optimum during the process.

Fig. 7 shows the AR of the UAVs. The AR value represents whether the distribution strategy we obtained is the optimal strategy. According to the experimental results in Fig. 7, the proposed LSTM+DQN algorithm obtains the highest reward value after convergence, which is 4% higher than that of the DQN algorithm. In addition, the AR of the Q-learning method is about 40% lower than the other two algorithms. This is because the first two algorithms have a memory library for learning historical information, which helps to find the optimal strategy. Since LSTM can store historical information for a long period of time, its AR converges about 8,000 steps earlier than the other two algorithms.

The ASCRs of the three schemes are presented in Fig. 8. A successful communication means that data sent by an UAV is received by the receiving terminal successfully. In this paper, both collision and non-access are considered to be communication failure. Therefore, the rate of successful communication is not equal to 1 minus the probability of collision. The reason is that even when there is no collision, the UAVs may still choose not to transmit information with a small probability. It can be seen from the curve that after convergence, the rate of successful communication of the proposed algorithm is about 5% higher than the other two traditional algorithms. Although the improvement is moderate in this particular scenario, it nevertheless demonstrates the advantage of our proposed algorithm. However, in a real world communication environment, the traffic demand of UAV communications could be relatively large, which is the aspect we will consider in our future work.

Based on the above results and discussions, we conclude that the proposed LSTM+DQN algorithm achieves a faster convergence speed and an superior channel allocation strategy than the two baseline schemes.

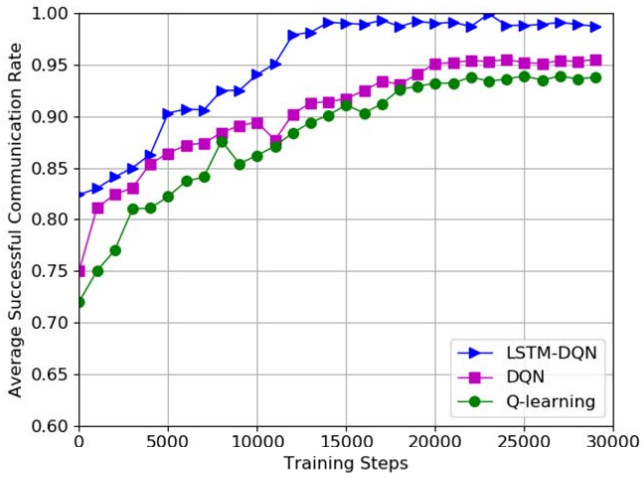

Fig. 8. The average successful communication rate achieved by Q-learning, DQN, and LSTM+DQN.

V. Conclusion

In this paper, we studied the dynamic channel allocation strategy for a multi-UAV system. To reduce the chance of collision, each UAV first scanned the channels before accessing the channel; it also randomly accessed the unoccupied channels. We proposed to enhance the DQN algorithm with an LSTM network to preserve more historical information, and the output of the LSTM is used as the input to the DQN, so that the UAVs can more effectively learn the features of the environment with more historical information. Experimental results showed that the proposed algorithm achieved a faster convergence speed, compared two the two baseline schemes, i.e., the DQN and Q-learning algorithms. In addition, our proposed algorithm also achieved a higher average reward, which was 40% higher than that of the Q-learning algorithm, indicating the proposed algorithm could find a more competitive channel allocation strategy. This paper considered a simulation environment without accurately modeling the UAV-receiving terminal channels. In our future work, we will systematically model the real environment and develop an implement our algorithm with off-the-shelf UAVs to evaluate the proposed scheme in a more realistic environment.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (61771154) and the Fundamental Research Funds for the Central Universities (HEUCFG201830). This paper is also funded by the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation. This wok is also sported in part by the NSF under Grant ECCS-1923163.

References

[1] D.R. Enrico, R. Fantacci, and G. Giambene, "Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 2, pp. 397–405, Feb. 1995.

[2] A. Franchi, C. Secchi, M. Ryll, H. H. Buelthoff, and P. R. Giordano, "Shared control: Balancing autonomy and human assistance with a group of quadrotor UAVs," *IEEE Robotics and Automation Magazine*, vol. 19, no. 3, pp. 57–68, Sept. 2012.

[3] Y. Cai, F. R.Yu, J. Li, Y. Zhou, and L. Lamon, "Medium access control for Unmanned Aerial Vehicle (UAV) ad-hoc networks with full-duplex radios and multipacket reception capability," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 390–394, Aug. 2013.

[4] G.R. Ding, Q.H. Wu, L.Y Zhang, Y. Lin, T.A. Tsiftsis, and Y.D. Yao, "An amateur drone surveillance system based on cognitive Internet of Things," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 29–35, Jan. 2018.

[5] Z.Y. Feng, L. Ji, Q.X. Zhang, and W. Li, "Spectrum management for mmWave enabled UAV swarm networks: Challenges and opportunities," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 146–153, Jan. 2019.

[6] Z. Liu, Y. Chen, B. Liu, C. Cao, and X. Fu, "HAWK: An unmanned mini-helicopter-based aerial wireless kit for localization," *IEEE Transactions on Mobile Computing*, vol.13, no. 2, pp. 287–298, Nov. 2014.

[7] A. Nowé, and T. Brys, "A gentle introduction to reinforcement learning," In: Schockaert S., Senellart P. (eds) *Scalable Uncertainty Management*, pp.pp 18–32, SUM 2016. Lecture Notes in Computer Science, vol 9858. Springer, Cham.

[8] V. Francois-Lavet, P. Henderson, R. Islam, M.G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 11, no. 1, pp.3–4, Nov. 2018.

[9] D. Zhang, ,X. Han , and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, Sept. 2018.

[10] Z. Yang, K. Merrick, L. Jin, and H.A. Abbass, "Hierarchical deep reinforcement learning for continuous action control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5174–5184, Nov. 2018.

[11] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems," *IEEE Transactions on Mobile Computing*, vol. 15, no. 4, pp. 817–825, Apr. 2016.

[12] Z.B. Gao, B. Wen, L.F. Huang, C. Chen, Z.W. Su, "Q-Learning-based power control for LTE enterprise eemtocell networks," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2699–2707, Dec. 2017.

[13] J.J. Ni, M.H. Liu, L. Ren, S.X. Yang, "A multiagent Q-Learning-based optimal allocation approach for urban water resource management system," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 1, 204–214, Jan. 2014.

[14] H.H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing based approach," *IEEE Internet of Things Journal* (Early Access), pp. 1-1, Sept. 2018.

[15] A.R. Syed, K.L.A. Yau, J. Qadir, H. Mohamad, N. Ramli, and S.L. Keoh, "Route selection for multi-hop cognitive radio networks using reinforcement learning: An experimental study," *IEEE Access Journal*, vol. 4, no. 2, pp. 6304–6324, Sept. 2016.

[16] V. Raj, I. Dias, T. Tholeti, and S. Kalyani, "Spectrum access in cognitive radio using a two stage reinforcement learning approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 20–34, Jun. 2018.

[17] N. Oshri , and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310–323, Nov. 2018.

[18] J. Lunden, S.R. Kulkarni, V. Koivunen, and H.V. Poor, "Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 858–868, Apr. 2013.

[19] Y.T. Liu, Y.Y. Lin, S.L. Wu, C.H. Chuang, and C.T. Lin, "Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 347–360, Feb. 2017.

[20] C. Jitong and D.L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, June 2017.

[21] H. Sepp, and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Dec. 1997.