CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Data-driven Deep Reinforcement Learning for Online Flight Resource Allocation in UAV-aided Wireless Powered Sensor Networks

Kai Li*
Wei Ni
Harrison Kurunathan*
Falko Dressler

*CISTER Research Centre
CISTER-TR-220102

2022/05/16

Data-driven Deep Reinforcement Learning for Online Flight Resource Allocation in UAV-aided Wireless Powered Sensor Networks

Kai Li*, Wei Ni, Harrison Kurunathan*, Falko Dressler

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: kai@isep.ipp.pt, Wei.Ni@data61.csiro.au, jhk@isep.ipp.pt, dressler@ccs-labs.org

https://www.cister-labs.pt

Abstract

In wireless powered sensor networks (WPSN), data of ground sensors can be collected or relayed by an unmanned aerial vehicle (UAV) while the battery of the ground sensor can be charged via wireless power transfer. A key challenge of resource allocation in UAV-aided WPSN is to prevent battery drainage and buffer overflow of the ground sensors in the presence of highly dynamic lossy airborne channels which can result in packet reception errors. Moreover, state and action spaces of the resource allocation problem are large, which is hardly explored online. To address the challenges, a new data-driven deep reinforcement learning framework, DDRL-RA, is proposed to train flight resource allocation online so that the data packet loss is minimized. Due to time-varying airborne channels, DDRL-RA firstly leverages long short-term memory (LSTM) with pre-collected offline datasets for channel randomness predictions. Then, Deep Deterministic Policy Gradient (DDPG) is studied to control the flight trajectory of the UAV, and schedule the ground sensor to transmit data and harvest energy. To evaluate the performance of DDRL-RA, a UAV-ground sensor testbed is built, where real-world datasets of channel gains are collected. DDRL-RA is implemented on Tensorflow, and numerical results show that DDRL-RA achieves 19\% lower packet loss than other learning-based frameworks.

Data-driven Deep Reinforcement Learning for Online Flight Resource Allocation in UAV-aided Wireless Powered Sensor Networks

1st Kai Li
CISTER Research Centre
Porto, Portugal
kai@isep.ipp.pt

2nd Wei Ni
CSIRO
Sydney, Australia
wei.ni@data61.csiro.au

3rd Harrison Kurunathan
CISTER Research Centre
Porto, Portugal
jhk@isep.ipp.pt

4th Falko Dressler
School of Electrical Engineering and Computer Science
TU Berlin
Berlin, Germany
dressler@tkn.tu-berlin.de

***Abstract*—In wireless powered sensor networks (WPSN), data of ground sensors can be collected or relayed by an unmanned aerial vehicle (UAV) while the battery of the ground sensor can be charged via wireless power transfer. A key challenge of resource allocation in UAV-aided WPSN is to prevent battery drainage and buffer overflow of the ground sensors in the presence of highly dynamic lossy airborne channels which can result in packet reception errors. Moreover, state and action spaces of the resource allocation problem are large, which is hardly explored online. To address the challenges, a new data-driven deep reinforcement learning framework, DDRL-RA, is proposed to train flight resource allocation online so that the data packet loss is minimized. Due to time-varying airborne channels, DDRL-RA firstly leverages long short-term memory (LSTM) with pre-collected offline datasets for channel randomness predictions. Then, Deep Deterministic Policy Gradient (DDPG) is studied to control the flight trajectory of the UAV, and schedule the ground sensor to transmit data and harvest energy. To evaluate the performance of DDRL-RA, a UAV-ground sensor testbed is built, where real-world datasets of channel gains are collected. DDRL-RA is implemented on Tensorflow, and numerical results show that DDRL-RA achieves 19% lower packet loss than other learning-based frameworks.**

***Index Terms*—UAV, WPSN, Deep reinforcement learning, LSTM, Wireless power transfer**

I. Introduction

Wireless powered sensor networks (WPSN) are deployed to sustainably monitor surroundings [1] or charge electric vehicles [2]. The data of distributed ground sensors that are deployed in harsh areas can be collected by using unmanned aerial vehicles (UAVs). The UAV can also charge the ground sensors remotely via wireless power transfer (WPT) [3] to power sensory data generation and wireless transmission. At different altitudes, the UAV is able to communicate with ground sensors. Thanks to line-of-sight (LoS) communications, transmit rate of WPSN and the UAV is highly improved [4].

Fig. 1 depicts a UAV-aided WPSN for precision agriculture, where ground sensors are deployed on a remote farmland for sensing the environment, e.g., acid precipitation, and ambient temperature and humidity [5], [6]. When the UAV approaches a ground sensor, the ground sensor equipped with a WPT receiver is charged by the UAV [7]. When the UAV flies away from the ground sensor, the ground sensor uses the harvested energy that is stored in the battery for the sensing operation. Moreover, the UAV can fly along its flight trajectory, and schedule the data transmission of the sensors [8], [9]. The transmit data queue of the sensor can be used to buffer sensory data that is transmitted to the UAV when the UAV is not around.

The number of packets in the queue of the ground sensor is distinctive from each other, due to time-varying data arrivals. When the UAV schedules a ground sensor which has a short queue to transmit data and WPT, the data buffer of the unscheduled sensors can be already full, and the newly arrived data can overflow the buffer. Moreover, packet transmission errors increase if a ground sensor experiencing a poor link quality is scheduled by the UAV, and the ground sensor can suffer from insufficient energy harvesting. In practice, the UAV is hardly to know the time-varying network dynamics, e.g., number of packets in the queue, WPT charging, and channel link qualities between the UAV and the ground sensor [10]. Additionally, the time-varying network dynamics

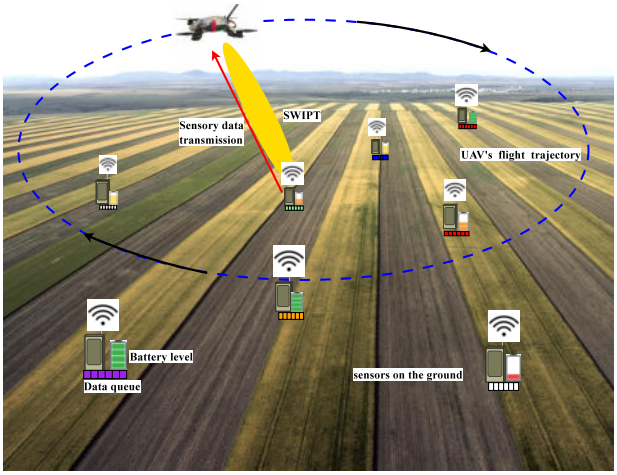


Fig. 1: UAV-aided WPSN for precision agriculture.

and waypoints of the UAV consists of a large number of real-time values, which results in an extremely large flight resource allocation space [11]. Therefore, it is difficult to jointly optimize the trajectory control and the scheduling of data transmission and WPT in a continuous domain, for minimizing buffer overflows and communication errors.

In this paper, a flight resource allocation, named as DDRL-RA, is proposed based on deep deterministic policy gradient (DDPG). DDRL-RA enables the UAV to learn the battery energy charged by WPT, and number of packets in the data queue of the sensors, and the link qualities. The UAV, i.e., trajectory planning and data transmission scheduling of ground sensors Since the trajectory of the UAV, battery energy and channel link qualities consist of a large number of real numbers, DDRL-RA needs to be trained in a continuous domain. We also investigate Long short-term memory (LSTM) with DDRL-RA for predicting the airborne channel qualities in terms of data transmission and WPT. The LSTM addresses the partial observability of the UAV on the states of the sensors, approximating the obscure states of unselected sensors at every instant for DDPG implementation.

In this paper, we present the literature review in Section II. The network model is studied in Section III. Section IV develops DDRL-RA to train the flight resource allocation at the UAV. In Section V, datasets of channel gains are collected from a UAV-ground sensor testbed, and the performance is evaluated. We conclude this paper in Section VI.

II. Related Work

In [12], the ground nodes' utilities are adjusted to determine a packet delivery and energy transfer policy for the UAV. Graph-based Markov Decision Process is used to formulate the problem. A mean-field approximation algorithm is studied to choose the best policy for each system state. The UAV's flight trajectory can be designed to increase the minimum received energy among all ground sensors given the maximum UAV flying speed limit [13]. The UAV-speed-constrained trajectory can be transformed into an equivalent UAV-speed-free problem, which is solved via Lagrange dual method. UAV-aided WPT is used to charge the ground sensors in [14]. Given different deployment of the ground sensors, the location of the UAV is designed to improve the sum harvested energy of the ground sensors, according to the power consumption of the UAV. In [15], the UAV's trajectory planning and ground sensors' scheduling scheme is studied to satisfy UAV's flying constraints from two WPT perspectives, i.e. the sum harvested energy of all ground sensors and the minimum received energy among all ground sensors. A resource allocation algorithm is presented to solve the problem by alternately adjusting wake-up scheduling of the ground sensor according to the UAV trajectory plan. A radio-map-based design approach is studied in [16], in which the UAV exploits the information of channel propagation environments for finding waypoints of the UAV. The objective is to increase the minimum energy transferred to all ground sensors over a particular charging duration. A basic two-ground sensor scenario is considered in [17]. The UAV's trajectory is designed to improve the amount of energy transferred to the two ground sensors during a given charging period. It shows that when the distance between the two sensors is smaller than a certain threshold, the boundary of the energy region is found when the UAV hovers above a fixed location between them.

Some preliminary results of using a DDPG-based trajectory planning are presented in our recent work [18], where actions of the UAV are trained without the channel prediction. Different from previous works that only provided solutions based on known knowledge of the network, this paper focuses on a practical scenario where the UAV has no a-priori knowledge on the network state. A new data-driven deep reinforcement learning is developed to exploit DDPG with LSTM to train the trajectory and the scheduling of data collection and WPT.

III. Flight, Channel, and WPT Models

In this section, we study the flight, channel, and WPT models of the considered UAV-aided WPSN.

A. Flight model of the UAV

We denote the location of the UAV as $(x(t), y(t), z)$ on a Cartesian plane, and the altitude of the UAV maintains at z [19]. The patrol speed of the UAV is $v(t)$, and we have

$$V_{min} \leq |v(t)| \leq V_{max}, \tag{1}$$

where V_{min} and V_{max} represent the minimum and the maximum speeds, respectively. Moreover, the UAV can conduct $\Delta v(t)$ to accelerate the flight from $(x(t), y(t), z)$ to $(x(t+$

$1), y(t+1), z)$, where the timespan is Δt. As the angular velocity of the UAV is $\theta(t)/\Delta t$, we have

$$\Delta v(t) = \theta(t)/\Delta t \times c, \tag{2}$$

where $\theta(t) \in (0, 180°]$ and c is the distance between the circle centre and the position of the UAV. The acceleration/deceleration of the UAV fulfills

$$(V_{min} - V_{max}) \leq \Delta v(t) \leq (V_{max} - V_{min}), \tag{3}$$

where $\Delta v(t) < 0$ stands for the deceleration, and $\Delta v(t) \geq 0$ is for the acceleration.

Thus,

$$(V_{min} - V_{max}) \leq \theta(t)/\Delta t \times c \leq (V_{max} - V_{min}). \tag{4}$$

Given the maximum speed $V_{max} = 15$m/s for most of commercial UAVs, we assume $\theta(t) \in (0°, 15°]$, where $c = 1$m and $\Delta t = 1$s.

We consider the smooth turn mobility model of the UAV with aeronautics and practicality consideration. Fig. 2 illustrates the flight model of the UAV, where $\theta(t)$ is a turning angle at time t, and the coordinates of the circle centre at time t are $(x_o(t), y_o(t), z)$. Thus, we have [20]

$$\theta(t) = \arctan\left(\frac{y(t+1) - y_o(t)}{x(t+1) - x_o(t)}\right) - \arctan\left(\frac{y(t) - y_o(t)}{x(t) - x_o(t)}\right) \tag{5}$$

Particularly, it is assumed that the UAV does not move backward. The UAV flies along a trajectory, where the instantaneous heading of the UAV is adjusted online according to the proposed DDRL-RA framework. The details are provided in the next section.

B. UAV-ground channels

Given constant Sigmoid parameters a and b, the LoS probability of the UAV-ground channel is

$$\Pr_{\text{LoS}}(t) = \frac{1}{1 + a\exp(-b[\varphi_i(t) - a])}. \tag{6}$$

Let $\varphi_i(t)$ denote an elevation angle of sensor i [21]. The path loss of the data transmission to the UAV is

$$\begin{aligned} h_i(t) = \Pr_{\text{LoS}}(\varphi_i(t))(\eta_{\text{LoS}} - \eta_{\text{NLoS}}) + 20\log(\mathcal{R}\sec\varphi_i(t)) \\ + 20\log(f_c) + 20\log(4\pi/v_c) + \eta_{\text{NLoS}} \end{aligned} \tag{7}$$

where the radius of the communication range of the UAV is $\mathcal{R}$. The radio frequency is f_c, and v_c gives the speed of light. η_{LoS} is the excessive path loss of LoS, and η_{NLoS} is the non-LoS one, where their values can be set for different application scenarios [22].

The UAV and the ground sensor can carry out channel reciprocity to be aware of the complex coefficient of the reciprocal UAV-ground channel. We denote the data rate and transmit power of sensor i as $r_i(t)$ and $P_i(t)$, respectively. According to [23], we have

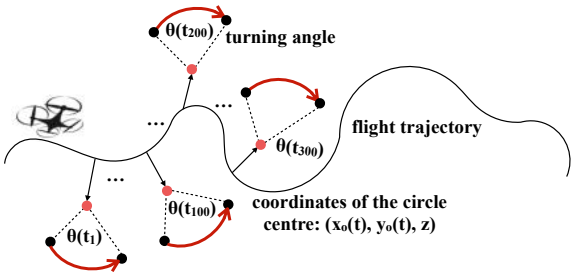


Fig. 2: The flight model of the UAV.

$$P_i(t) \approx \frac{\kappa_2^{-1} \ln \frac{\kappa_1}{\varepsilon}}{\|h_i(t)\|^2}(2^{r_i(t)} - 1), \tag{8}$$

where κ_1 and κ_2 are two channel constants, and $\|\cdot\|$ denotes norm.

C. WPT model

The distance between the UAV and sensor i along the flight trajectory at t is $q_i(t)$. The WPT transceiver alignment between the UAV and the ground sensor is $\gamma_i(t)$. The WPT efficiency factor $\phi(q_i(t), \gamma_i(t))$ depends on the distance between the UAV and the ground sensor, as well as the WPT transceiver alignment. Thus, the power transferred from the UAV to the ground sensor via WPT can be given by

$$\tilde{P}_i(t) = \phi(q_i(t), \gamma_i(t)) P_{\text{UAV}}^{tx} \|h_i(t)\|^2, \tag{9}$$

where P_{UAV}^{tx} is the transmit power at the UAV on WPT.

IV. Data-driven Deep Reinforcement Learning

In this section, the data-driven deep reinforcement learning, DDRL-RA is developed, which minimizes data losses due to buffer overflows and time-varying channels.

A. Problem formulation

Suppose that the number of ground sensors in the WPSN is N, where sensor $i \in [1, N]$. E denotes the battery capacity of the sensors, and the battery energy of sensor i has $e_i(t) \leq E$. Sensor i experiences random data arrivals, and the data to be transmitted is buffered in the queue. The number of data packets in the queue of sensor i is $d_i(t) \in [1, D]$. The buffers are finite with capacity of D, and the new data arrivals have to be dropped if $d_i(t) > D$ and start overflowing. The network state contains $e_i(t)$, $d_i(t)$, $e_{\text{UAV}}(t)$, $(x(t), y(t), z)$, and $h_i(t)$, where $i \in [1, N]$. Therefore, the battery energy of the UAV at t is

$$e_{\text{UAV}}(t) = e_{\text{UAV}}(t-1) + \Delta e_{\text{UAV}}(t) - \Delta E_{\text{UAV}}(t), \tag{10}$$

$$\Delta E_{\text{UAV}}(t) = \tilde{P}_i(t) * t, \tag{11}$$

where $\Delta E_{\mathrm{UAV}}(t)$ denotes the energy consumption on WPT, and $\Delta e_{\mathrm{UAV}}(t)$ is the amount of energy that the UAV harvests from its onboard solar panels.

At network state S_t, the UAV can conduct an action to determine the next location, i.e., $(x'(S_t), y'(S_t), z)$, and select a ground sensor to transmit data. Thus, the action can be given by

$$u_t = ((x'(S_t), y'(S_t), z), i_t), \tag{12}$$

where i_t denotes the selected sensor ID. When an action u_t is carried out at S_t, the packet loss can be measured as $C\{S_{t+1}|S_t, u_t\}$, i.e., network costs, and the next state is S_{t+1}.

B. Data-driven deep reinforcement learning

Fig. 3 depicts the proposed DDRL-RA, where LSTM is used to predict the time-varying channel fading in the environment. With the future network state prediction, DDPG optimizes the instantaneous heading of the UAV and sensor selection. DDRL-RA concurrently learns an action-value function and a policy. DDRL-RA utilizes an Actor-Critic architecture to combine the value iteration and the policy iteration to implement the proposition of the continuous state space and the continuous action space by using deep reinforcement learning. This is different from deep Q-networks (DQN) which focus on a discrete action space. Moreover, DDRL-RA can enlarge the continuous state and action space while minimizing $C\{S_{t+1}|S_t, u_t\}$ compared with reinforcement learning which suffers from the well-known curse of dimensionality [24]. The experience tuple $\left(S_t, S_{t+1}, u_t, C\{S_{t+1}|S_t, u_t\}\right)$ at each training step can be stored at the onboard replay memory at the UAV. Let i_t denote the selected ground sensor at state S_t. The network state that the UAV can observe is $\{e_{\mathrm{UAV}}(S_t), b_{i_t}(S_t), h_{i_t}(S_t), d_{i_t}(S_t), (x(S_t), y(S_t), z)\}$.

The action u_t can be optimized by a gradient-assisted training $\mu\{S_t\}$. Specifically, an actor neural network decides $(x'(t), y'(t), z)$ and the scheduled sensor i_t $(1 \leq i_t \leq N)$ to train $\mu\{S_t\}$. A critic neural network is trained to approximate the optimal action-value function $Q\{S_t, u_t\}$ to obtain the expected overall data loss. Moreover, $\mu\{S_t|w^{\mu}\}$ provides the flight resource allocation policy of the actor neural network. $\mu'\{S_t|w^{\mu'}\}$ is the target policy in the actor neural network. w^{μ} and $w^{\mu'}$ define the weights with regards to the policies training. The approximation loss Δ_{loss} can be minimized by adjusting the weights w^{Q} in the critic neural network.

Since the time-varying channel fading leads to unknown network state transitions, the Actor-Critic architecture suffers from learning uncertainties, which reduces the training accuracy of the actions. Particularly, the DDRL-RA is carried out onboard at the UAV, where the network states are not fully observable. It can only make the observation of the UAV itself and the selected ground sensor, i.e., $\{e_{\mathrm{UAV}}(S_t), b_{i_t}(S_t), h_{i_t}(S_t), d_{i_t}(S_t), (x(S_t), y(S_t), z)\}$. As a result, the deep reinforcement learning accuracy can be compromised. To address this issue, LSTM is developed with DDRL-RA to predict the unobservable network states. The network state prediction achieved by LSTM is feed into the training environment of the DDPG. The output of the LSTM gives hidden states S_t^{hid}. The hidden state depends on the network activation in the previous time steps. Thus, LSTM is suitable for the proposed flight resource allocation problem, in which we wish to extract useful features from the actions of the UAV and predicted state dynamics, and reduce our state space.

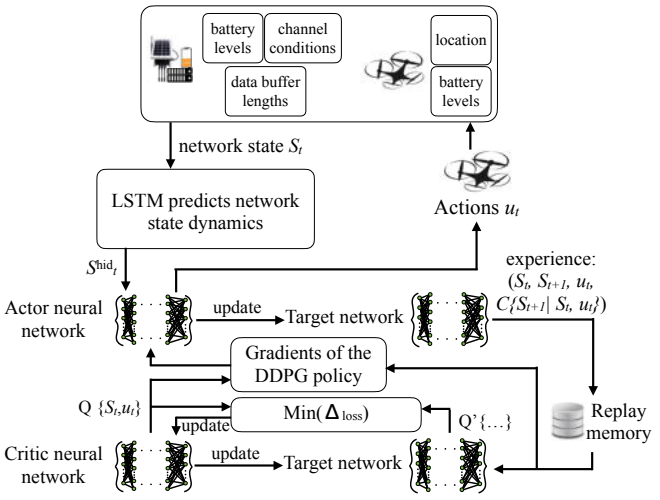


Fig. 3: The architecture of DDRL-RA, where LSTM is used to predict the time-varying channel fading in the environment, while DDPG optimizes the instantaneous heading of the UAV and sensor selection.

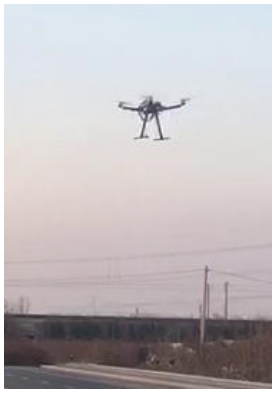
V. Datasets and Performance Evaluation

Datasets of channel gains are presented in this section, and the proposed DDRL-RA is implemented on Google TensorFlow. For performance evaluation, the packet loss is shown in accordance to the training episodes.

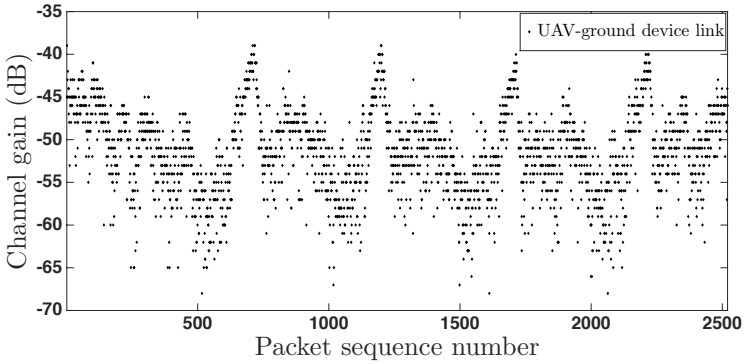
A. Datasets of channel gains

As shown in Fig. 4(a), a UAV is employed to communicate with a ground sensor to measure the channel gain, where the UAV patrols along a predetermined trajectory and broadcasts beacon packets to the ground sensor. Fig. 4(b) shows the dataset that records the link qualities between the UAV and the ground node. In particular, the link quality drops when the UAV moves away from the ground sensor.

The collected datasets of channel gains are used to train the proposed DDRL-RA, which enables LSTM for predicting the time-changing channel fading. The dataset is firstly normalized

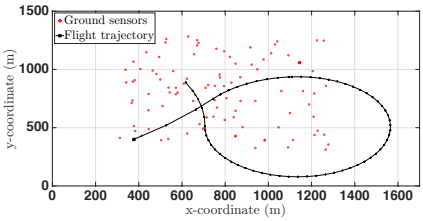


(a) The UAV broadcasts beacon packets.

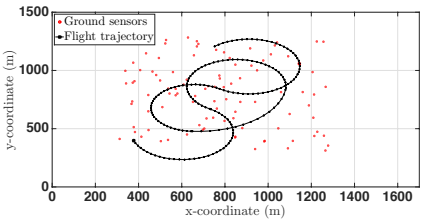


(b) The channel quality dataset.

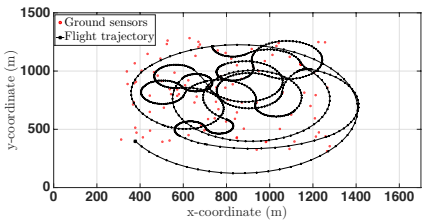
Fig. 4: A real-world UAV patrols along its trajectory while broadcasting beacon packets to the ground sensor (as shown in (a)). The channel gains are measured and 2500 data samples are plotted in (b).



(a) LSTM = 10, DDPG = 100.



(b) LSTM = 500, DDPG = 100.



(c) LSTM = 500, DDPG = 800.

Fig. 5: The flight trajectories of the UAV.

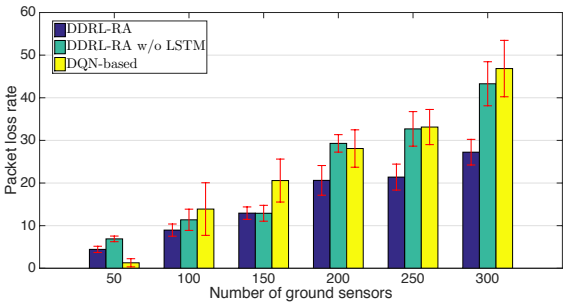


Fig. 6: Packet loss rate. The standard deviation is calculated over 20 experiments.

by *MinMaxScaler()* in TensorFlow. Next, the LSTM model is implemented by *Sequential()* in Keras to predict the future channel gain. In addition, the LSTM model is configured by *LSTM_model.compile (loss = 'mean_squared_error', optimizer = 'adam')*, which applies adam optimizer in TensorFlow for minimizing Δ_{loss}.

B. Performance of DDRL-RA

Fig. 5(a), (b), and (c) study the cruise routes of the UAV in terms of training duration of LSTM and DDPG. It is observed that DDRL-RA can adjust the cruise control onboard at the UAV in the continuous action space. The actions of $(x'(S_t), y'(S_t), z)$ and i_t for data collection and WPT are constantly optimized. In Fig. 5(a), as the experience in the replay memory is not ample, the trajectory of the UAV and sensor selection are hardly optimized given a small number of LSTM epochs and learning iterations. In Fig. 5(b), the training of LSTM and DDPG is extended to 500 and 100, respectively. The UAV starts to adjust the flight to schedule more ground sensors for data transmission and WPT. In Fig. 5(c), since the training of DDRL-RA is extended, LSTM and DDPG are sufficiently trained for minimizing Δ_{loss}.

Fig. 6 plots the packet loss rate of the DDRL-RA, DDRL-RA without LSTM, and DQN-based solution. The number of sensors increases from 50 to 300. More ground sensors will keep the data in the queue and wait for the UAV, until the sensor that is scheduled finishes the data transmission and WPT. As a result, the packet loss rate increases steadily with the growth of the WPSN size. In the case of actor-critic based policies such as DDRL-RA and DDRL-RA without LSTM, their packet loss rates are similar when $N \leq 150$ sensors. Moreover, the packet loss rate of DDRL-RA is about 15% and 19% lower than the DDRL-RA without LSTM and

DQN-based one when $N = 300$. This is because LSTM of DDRL-RA predicts the channel dynamics of all the ground sensors, which efficiently adapts the sensor selection for data transmission and WPT to reduce the data packet loss.

VI. Conclusion

This paper studied a data-driven deep reinforcement learning to train resource allocation in UAV-aided WPSN for minimizing the data loss. The proposed DDRL-RA leveraged LSTM to predict channel randomness while DDPG is developed to determine the UAV's trajectory as well as the scheduling of data transmission and WPT. A UAV-ground sensor testbed was built, which measures the link quality of the UAV-ground channel in real world. The collected experimental datasets were utilized to train the LSTM. DDRL-RA was implemented on Tensorflow, and numerical results showed that DDRL-RA achieves 19% lower packet loss than other deep reinforcement learning frameworks.

Acknowledgements

This work was partially supported by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology), within the CISTER Research Unit (UIDP/UIDB/04234/2020); also by national funds through the FCT, under CMU Portugal partnership, within project CMU/TIC/0022/2019 (CRUAV).

This work was in part supported by the Federal Ministry of Education and Research (BMBF, Germany) as part of the 6G Research and Innovation Cluster 6G-RIC under Grant 16KISK020K.

References

[1] C. Wang, J. Li, Y. Yang, and F. Ye, "A hybrid framework combining solar energy harvesting and wireless charging for wireless sensor networks," in *IEEE INFOCOM*, 2016, pp. 1–9.

[2] P. Machura, V. De Santis, and Q. Li, "Driving range of electric vehicles charged by wireless power transfer," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 5968–5982, 2020.

[3] Z. Chen, K. Chi, K. Zheng, G. Dai, and Q. Shao, "Minimization of transmission completion time in UAV-enabled wireless powered communication networks," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1245–1259, 2019.

[4] A. Al-Hourani, "On the probability of line-of-sight in urban environments," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1178–1181, 2020.

[5] K. Li, W. Ni, and F. Dressler, "Continuous maneuver control and data capture scheduling of autonomous drone in wireless sensor networks," *IEEE Transactions on Mobile Computing*, 2021.

[6] P. Spachos and S. Gregori, "Integration of wireless sensor networks and smart UAVs for precision viticulture," *IEEE Internet Computing*, vol. 23, no. 3, pp. 8–16, 2019.

[7] S.-W. Dong, X. Li, X. Yu, Y. Dona, H. Cui, T. Cui, Y. Wang, and S. Liu, "Hybrid mode wireless power transfer for wireless sensor network," in *2019 IEEE Wireless Power Transfer Conference (WPTC)*. IEEE, 2019, pp. 561–564.

[8] P. Luong, F. Gagnon, L.-N. Tran, and F. Labeau, "Deep reinforcement learning-based resource allocation in cooperative UAV-assisted wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7610–7625, 2021.

[9] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 131–141, 2020.

[10] K. Li, W. Ni, and F. Dressler, "LSTM-characterized deep reinforcement learning for continuous flight control and resource allocation in UAV-assisted sensor network," *IEEE Internet of Things Journal*, 2021.

[11] S. Yin and F. R. Yu, "Resource allocation and trajectory design in UAV-aided cellular networks based on multi-agent reinforcement learning," *IEEE Internet of Things Journal*, 2021.

[12] S. Lhazmir, O. A. Oualhaj, A. Kobbane, J. Ben-Othman *et al.*, "UAV for wireless power transfer in IoT networks: A GMDP approach," in *IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[13] Y. Hu, X. Yuan, J. Xu, and A. Schmeink, "Optimal 1D trajectory design for UAV-enabled multiuser wireless power transfer," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5674–5688, 2019.

[14] H. Yan, Y. Chen, and S.-H. Yang, "UAV-enabled wireless power transfer with base station charging and UAV power consumption," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 12 883–12 896, 2020.

[15] Y. Wang, M. Hua, Z. Liu, D. Zhang, B. Ji, and H. Dai, "UAV-based mobile wireless power transfer systems with joint optimization of user scheduling and trajectory," *Mobile Networks and Applications*, pp. 1–15, 2019.

[16] X. Mo, Y. Huang, and J. Xu, "Radio-map-based robust positioning optimization for UAV-enabled wireless power transfer," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 179–183, 2019.

[17] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy region characterization," in *IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2017, pp. 1–7.

[18] H. Kurunathan, K. Li, W. Ni, E. Tovar, and F. Dressler, "Deep reinforcement learning for persistent cruise control in UAV-aided data collection," in *IEEE Conference on Local Computer Networks (LCN)*. IEEE, 2021, pp. 347–350.

[19] K. Li, W. Ni, E. Tovar, and M. Guizani, "Joint flight cruise control and data collection in UAV-aided internet of things: An onboard deep reinforcement learning approach," *IEEE Internet of Things Journal*, 2020.

[20] Y. Emami, B. Wei, K. Li, W. Ni, and E. Tovar, "Joint communication scheduling and velocity control in multi-UAV-assisted sensor networks: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 986–10 998, 2021.

[21] C. Liu, T. Q. Quek, and J. Lee, "Secure UAV communication in the presence of active eavesdropper," in *International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2017, pp. 1–6.

[22] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *2014 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2014, pp. 2898–2904.

[23] K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, no. 6, pp. 1377–1386, 2016.

[24] Y. Emami, B. Wei, K. Li, W. Ni, and E. Tovar, "Deep Q-networks for aerial data collection in multi-UAV-assisted wireless sensor networks," in *International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2021, pp. 669–674.