CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks

Yousef Emami*
Bo Wei
Kai Li*
Wei Ni
Eduardo Tovar*

*CISTER Research Centre

CISTER-TR-210304

2021/06/28

Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks

Yousef Emami*, Bo Wei, Kai Li*, Wei Ni, Eduardo Tovar*

*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: emami@isep.ipp.pt, kai@isep.ipp.pt, Wei.Ni@data61.csiro.au, emt@isep.ipp.pt

https://www.cister-labs.pt

Abstract

Unmanned Aerial Vehicles (UAVs) can collaborate to collect and relay data for ground sensors in remote and hostile areas. In multi-UAV-assisted wireless sensor networks (MA-WSN), the UAVs' movements impact on channel condition and can fail data transmission, this situation along with newly arrived data give rise to buffer overflows at the ground sensors. Thus, scheduling data transmission is of utmost importance in MA-WSN to reduce data packet losses resulting from buffer overflows and channel fading. In this paper, we investigate the optimal ground sensor selection at the UAVs to minimize data packet losses . The optimization problem is formulated as a multi-agent Markov decision process, where network states consist of battery levels and data buffer lengths of the ground sensor, channel conditions, and waypoints of the UAV along the trajectory. In practice, an MA-WSN contains a large number of network states, while the up-to-date knowledge of the network states and other UAVs' sensor selection decisions is not available at each agent. We propose a Multi-UAV Deep Reinforcement Learning based Scheduling Algorithm (MUAIS) to minimize the data packet loss, where the UAVs learn the underlying patterns of the data and energy arrivals at all the ground sensors. Numerical results show that the proposed MUAIS achieves at least 46\% and 35\% lower packet loss than an optimal solution with single-UAV and an existing non-learning greedy algorithm, respectively.

Deep Q-Networks for Aerial Data Collection in Multi-UAV-Assisted Wireless Sensor Networks

Yousef Emami
CISTER Research Centre
Porto, Portugal
emami@isep.ipp.pt

Bo Wei
Northumbria University
Newcastle, U.K.
bo.wei@northumbria.ac.uk

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

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

Eduardo Tovar
CISTER Research Centre
Porto, Portugal
emt@isep.ipp.pt

***Abstract*—Unmanned Aerial Vehicles (UAVs) can collaborate to collect and relay data for ground sensors in remote and hostile areas. In multi-UAV-assisted wireless sensor networks (MA-WSN), the UAVs' movements impact on channel condition and can fail data transmission, this situation along with newly arrived data give rise to buffer overflows at the ground sensors. Thus, scheduling data transmission is of utmost importance in MA-WSN to reduce data packet losses resulting from buffer overflows and channel fading. In this paper, we investigate the optimal ground sensor selection at the UAVs to minimize data packet losses. The optimization problem is formulated as a multi-agent Markov decision process, where network states consist of battery levels and data buffer lengths of the ground sensor, channel conditions, and waypoints of the UAV along the trajectory. In practice, an MA-WSN contains a large number of network states, while the up-to-date knowledge of the network states and other UAVs' sensor selection decisions is not available at each agent. We propose a Multi-UAV Deep Reinforcement Learning based Scheduling Algorithm (MUAIS) to minimize the data packet loss, where the UAVs learn the underlying patterns of the data and energy arrivals at all the ground sensors. Numerical results show that the proposed MUAIS achieves at least 46% and 35% lower packet loss than an optimal solution with single-UAV and an existing non-learning greedy algorithm, respectively.**

***Index Terms*—Unmanned aerial vehicles, Communication scheduling, Multi-UAV Deep Reinforcement Learning, Deep Q-Network.**

I. Introduction

Unmanned Aerial Vehicles (UAVs) enjoy excellent mobility, low-cost technology and flexibility, these features allow them to be used in many civilian and commercial applications, e.g., weather monitoring, traffic control, package delivery [1] and crop monitoring [2]. UAVs are also employed to collect and process data from ground sensors deployed in harsh environments, such as natural disaster monitoring [3], border surveillance [1], and emergency assistance [4]. Fig. 1 depicts a typical multi-UAV-assisted wireless sensor network (MA-WSN). The sensors are deployed on oil pipelines in remote areas to measure humidity, oil flow rate, temperature changes, and detect potential leakages [5]. Sensory data are generated by the ground sensors and stored in a data queue for later transmission to the UAVs. The UAVs hover over the pipeline and move sufficiently close to each ground sensor to exploit short-distance line-of-sight (LoS) communications for data collection.

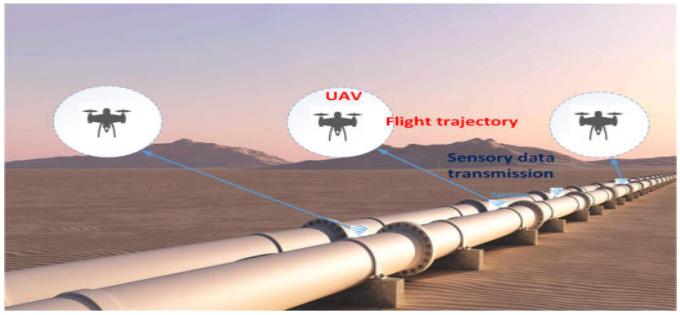

Fig. 1. The ground sensors in MA-WSN are deployed to monitor the oil pipeline. The UAVs hover over the pipeline and collect sensory data from the ground sensors.

In MA-WSN, the ground sensors undergo random data arrivals, since the data generation experiences a random environmental change of the temperature and humidity. The UAVs cooperatively move along the predetermined trajectories over the area of interest, while collecting the sensory data of the ground sensors. Scheduling a ground sensor for data collection can result in buffer overflows at other unselected sensors, since newly generated data at the unselected sensors have to be dropped if their buffers are already full and overflow. Thus, scheduling data collection of the UAVs for preventing data queue overflow and communication failure of the ground sensors is critical.

In this paper, we formulate the communication schedules of multiple UAVs as a multi-agent Markov Decision Process (MMDP). Each MMDP state contains the battery levels and data buffer lengths of the ground sensors, the channel conditions, and waypoints of the UAVs. The UAVs take the actions

* Corresponding author.

of selecting the ground sensors for data transmission. The UAVs' movements along trajectories with a large number of waypoints give rise to a plethora of channel conditions. This situation impact on data transmission and can diversify data queue lengths of the ground sensors given data arrival pattern. The main contributions of this paper can be summarized as follows:

- We formulate the problem of data collection scheduling as an MMDP to minimize the overall packet loss resulting from the buffer overflow and channel fading. To relieve the MMDP with large state and action spaces, a new Multi-UAV Deep Reinforcement Learning based Scheduling Algorithm (MUAIS) is proposed to optimize the selection of the ground sensor.
- We implement the proposed MUAIS in PyTorch which is an open source machine learning framework based on the Torch library [6]. Numerical results demonstrate that our MUAIS reduces the packet loss by up to 46% and 35%, as compared to a single-UAV case and an existing non-learning greedy algorithm, respectively.

The rest of this paper is organized as follows. Section II presents the related work on multi-UAV systems. Section III devotes to the system model. In Section IV, we formulate MMDP and multi-UAV Deep Q-networks (DQN) and present the proposed approach. Performance evaluation is presented in Section V. This paper is concluded in Section VI.

II. Related Work

Aerial data collection in the wireless sensor network using deep reinforcement learning is studied in [7]. The UAV follows a circular trajectory and equips with WPT transmitter schedules the ground sensor to prevent buffer overflows and charge the battery of the ground sensor to extend network lifetime. In [8], multiple UAVs provide energy supply and communication services to IoT devices in a wireless powered communication network. To improve throughput, a multi-UAV DQN based approach is used to adjust path planning of the UAVs and channel resource assignment. In [9], UAVs use DQN to conduct decisions for energy-efficient data collection. An energy-efficient deep reinforcement learning (DRL)-based UAV control strategy is developed in [10] to improve network coverage, fairness and connectivity.

In [11], the dueling DQN is used to adjust the deployment of UAVs so that the capacity of the communication links can be enhanced, while the ground nodes can be fully covered. The problem is modeled as a constrained MDP problem. The multi-agent reinforcement learning is used in [12] to address the resource allocation problem in UAV networks. A Q-learning based algorithm is used to enhance the long-term rewards, where the UAV selects its communication modes, power levels and sub-channels. In [13], spectrum sharing between the ground user and the UAVs is studied to improve the network utility. The UAVs can be used as relays for the primary network or can perform data transmission to the fusion center. The problem is formulated as a deterministic MMDP, which is addressed by Q-learning.

In [14], path planning of the UAVs is studied to reduce transmission latency, while improving energy efficiency. DRL based on echo state networks is used to perform path planning. In [15], trajectories and transmit power of UAVs are adjusted to improve the ground users' throughput, while satisfy the users' data rate requirement.

In [16], online velocity control and data capture are studied in UAV-enabled IoT networks. DQN is developed in the presence of outdated knowledge to determine the patrolling velocity and data transmission schedule of the IoT node. In [9], the joint flight cruise control and data collection scheduling in the UAV-aided IoT network is formulated as a POMDP to minimize the data lost due to buffer overflows at the IoT nodes and fading airborne channels.

An energy efficient and buffer-aware scheduling scheme for UAV-aided relaying is developed in [17]. The problem is formulated to minimize energy consumption of users by optimizing user selection, subject to overflow, power and fairness constraint. An energy-efficient transmission scheduling scheme of UAVs in a cooperative relaying network is developed in [18] such that the maximum energy consumption of all the UAVs is minimized, in which an applicable sub-optimal solution is developed and the energy could be saved up to 50% via simulations.

In the proposed MUAIS, every UAV learns the network state dynamics to minimize the overall packet loss due to the buffer overflow and channel fading. To alleviate the curse of dimensionality of the state space, MUAIS is equipped with DQN to optimize the selection of the ground sensor. [12] studied multi-agent reinforcement learning using Q-learning and stochastic game theory model for resource allocation in UAV networks. MUAIS is based on multi-UAV paradigm and has the merits of scalability and survivability over single UAV paradigm in [7].

III. System Model

The considered network contains J ground sensors and I UAVs. The UAVs fly along pre-determined routes which consist of a large number of waypoints to cover all the ground sensors in the field. The location of UAV i on its trajectory at t is denoted by $\zeta_i(t)$. The UAVs are responsible for collecting sensory data from the ground sensors.

The channel coefficient between UAV i ($\in [1, I]$) and ground sensor j ($\in [1, J]$) at time t is $h_j^i(t)$, which can be acquired by channel reciprocity. The modulation scheme of ground sensor j at t is denoted by $\phi_j(t)$. In particular, $\phi_j(t)$= 1, 2, and 3 indicates binary phase-shift keying (BPSK), quadrature-phase shift keying (QPSK), and 8 phase-shift keying (8PSK), respectively, and $\phi_j(t) \geq 4$ provides $2^{\phi_j(t)}$ quadrature amplitude modulation (QAM). Let $h_j^i(t)$ denote channel gain between ground sensor j and UAV i. The transmit power of the ground sensor, denoted by $P_j^i(t)$, is [19]

$$P_j^i(t) = \frac{ln\frac{k_1}{\epsilon}}{k_2 h_j^i(t)^2}(2^{\phi_j(t)} - 1) \quad (1)$$

TABLE I
NOTATION AND DEFINITION

Notation	Definition
J	number of ground sensors
I	number of UAVs
a_i	action of UAV i
a_u^{t-1}	past actions of other UAVs on a ground sensor
$S_{\alpha,i}$	state of UAV i
$S_{\beta,i}$	next state of UAV i
$P_j^i(t)$	transmit power between ground sensor j and UAV i
$h_j^i(t)$	channel gain between ground sensor j and UAV i
$\zeta_i(t)$	location of the UAV on its trajectory
$e_j(t)$	battery level of ground sensor j
$q_j(t)$	queue length of ground sensor j
D	maximum queue length of ground sensor
δ	discount factor for future states
θ	learning weight in deep Q-network

where k_1 and k_2 are channel constants, and ϵ denotes the required bit error rate (BER) of the channel.

We consider that UAV i moves in a low attitude for data collection, where the probability of LoS communication between UAV i and ground sensor j is given by [20]

$$Pr_{LoS}(\varphi_j^i) = \frac{1}{1 + a \cdot exp(-b[\varphi_j^i - a])} \tag{2}$$

where a and b are constants, and φ_j^i is the elevation angle between UAV i and ground sensor j. Furthermore, the path loss of the channel between UAV i and ground sensor j is modeled by

$$\gamma_j^i = Pr_{LoS}(\varphi_j^i)(\eta_{LoS} - \eta_{NLoS}) + 20log(r\sec(\varphi_j^i)) + 20log(\lambda) + 20log(\frac{4\pi}{v_c}) + \eta_{NLoS} \tag{3}$$

where r is the radius of the radio coverage of UAV i, λ is the carrier frequency, and v_c is the speed of light. η_{LoS} and η_{NLoS} are the excessive path losses of LoS or non-LoS, respectively.

IV. DEEP Q-NETWORKS WITH MULTIPLE UAVS

In this section, MMDP is formulated, while MUAIS is developed based on multi-UAV DQN.

A. MMDP Formulation

MMDP can be defined by the tuple $\{I,\ \{S_{\alpha,i}\},\ \{a_i\},$ C$\{S_\beta|S_\alpha,a\}\}$, where

1) I denotes the number of agents, i.e., UAVs.
2) $S_{\alpha,i}$ is the network state observed by agent i ($i \in I$). $S_{\alpha,i}$ consists of: channel quality $h_j^i(t)$, battery level $e_j(t)$, queue length $q_j(t)$ and the location of the UAV $\zeta_i(t)$,
i.e., $S_{\alpha,i}$=$\{(\mathrm{h}_j^i(t),\ e_j(t),\ q_j(t),\ \zeta_i(t)),\ \mathrm{i}=1,2,...\mathrm{I}\}$. Let S_α denotes the joint network state, where S_α=$S_{\alpha,1}\times.....\times S_{\alpha,I}$.
3) a_i represents the action of agent i, which schedules one of the sensors to transmit data to the UAV, i.e., a_i=$\{(j), i = 1, 2, ...I\}$. The joint action of the UAVs, denoted by a, can be given by a=$a_1\times.....\times a_I$.
4) C$\{S_\beta|S_\alpha,\ a\}$ is the network cost yielded when the joint action a is taken to transit from state S_α to S_β. The network cost is the packet loss of the ground sensors.

The size of state space for each agent is $((K+1)(D+1))J + H + W$, where K and D are the highest battery level and maximum queue length of the ground sensor respectively. H and W are the number of channel states and the number of waypoints on the UAV's trajectory. The action to be taken is to schedule one ground sensor to transmit data to the UAV. The size of the action space is J.

B. Multi-UAV DQN

The formulated MMDP has large state space since mobility feature of the UAVs give rise to a myriad of network states. Statistical methodologies cannot be applied to solve the MMDP due to lack of real-time knowledge and the size of the state space. For example, dynamic programming and linear programming are two typical statistical methodologies for solving the MMDP. Dynamic programming approaches such as policy iteration and value iteration derive the optimal policy based on explicit enumeration. This can create a computational bottleneck as the size of state space growing fast. A linear program can be used to formulate the problem of computing an optimal policy for the MMDP. However, linear programs solvable in polynomial time are impractical for the large MMDP [21]. DQN can alleviate the curse of dimensionallity of the large state space and deal with outdated knowledge of the network states. Particularly, the action of each UAV not only determines the future state, but also influences the actions of the other UAVs. Each UAV interacts with an unknown environment to learn a policy. In the learning process, UAV i observes the current environment state then takes an action in accordance with its policy to obtain the cost and the new environment state. After that, UAV i utilizes the gathered data to optimize its policy. UAV i interacts with the environment, performs the action, and optimizes its policy for many iterations until convergence to the optimal policy. In the multi-UAV setting, each UAV is designed to find an optimal policy π_i for minimizing the long-term expected accumulated discounted cost. From the perspective of UAV i, the accumulated cost by executing action a_i dependent on a_u^{t-1} at the current environment state $S_{\alpha,i}$ on the basis of policy π_i can be given by

$$Q_i^{\pi_i}(S_{\alpha,i}, a_i, a_u^{t-1}) = E[\Sigma_{t=0}^{\infty}\delta^t C(S_{\beta,i}|S_{\alpha,i}, a_i, a_u^{t-1})] \tag{4}$$

where $\delta \in$ [0, 1] is the discount factor. Each UAV aims to learn the optimal Q-value or the optimal policy. The Q-value of UAV i is updated as follows:

$$Q_i(S_{\beta,i}|S_{\alpha,i}, a_i, a_u^{t-1}) = (1-\nu)Q_i(S_{\beta,i}|S_{\alpha,i}, a_i, a_u^{t-1}) + \nu(C(S_{\beta,i}|S_{\alpha,i}, a_i, a_u^{t-1}) + \delta \min_{a_i'} Q_i(S_{\beta',i}|S_{\beta,i}, a_i', a_u') \tag{5}$$

where $\nu \in (0, 1]$ is the learning rate, $S_{\beta',i}$ is the next state and a'_i is the next action. Multi-UAV Q-learning cannot deal with the exponential growth of states and actions for the resource allocation problem in the MA-WSN. This is known as the curse of dimensionality. Multi-UAV DQN can circumvent the curse-of-dimensionality. It represents the action-value function of each UAV with a deep neural network parameterized by θ^{Q_i}. For each UAV, θ^{Q_i} is learned by sampling transitions from the replay memory and minimizing the squared temporal difference error:

$$\Gamma(\theta^{Q_i}) = y_i - Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}. \quad (6)$$

where y_i is the target Q-value which is set as a label and can be denoted by

$$y_i = C\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}\} + \delta \min_{a'_i} \mathrm{Q}'_i\{S_{\beta,i'} \mid S_{\beta,i}, a'_i, a'_u; \theta^{Q'_i}\} \quad (7)$$

Multi-UAV DQN uses target network and experience replay for each UAV to guarantee stability. In the multi-UAV DQN, experience replay removes correlations in the observation sequence and smooths over changes in the data distribution by randomizing over the states and the actions of MMDP at each time-step *t*. The provided multi-UAV DQN formulation is effective and promising for computing multi-UAV policies. In contrast to the traditional approaches for solving MMDP. It is able to deal with enormous size and complexity.

C. Proposed MUAIS

We present a multi-UAV version of DQN called MUAIS, as depicted in Fig.2. Overall, two separate Q-networks are maintained at each UAV, Q-network, denoted by $Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}$, and target network, denoted by $Q'_i\{S_{\beta,i'} \mid S_{\beta,i}, a'_i, a'_u; \theta^{Q'_i}\}$ with weights θ^{Q_i} and $\theta^{Q'_i}$, respectively. At first step, the Q-network and associated target of each UAV are initialized and then learning is ignited. Each UAV samples its state and computes its local state $S_{\alpha,i}$. Each UAV receives the local state $S_{\alpha,i}$ and selects a random action with probability ε, or exploits its knowledge and produces its action. Then, the corresponding cost and next state are sampled. The associated transition $(S_{\alpha,i}, S_{\beta,i}, a_i, C)$ is stored. θ^{Q_i} is learned by sampling batches of transitions from the replay memory and minimizing the squared temporal difference error:

$$\Gamma(\theta^{Q_i}) = y_i - Q_i\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}; \theta^{Q_i}\}. \quad (8)$$

where

$$y_i = C\{S_{\beta,i} \mid S_{\alpha,i}, a_i, a_u^{t-1}\} + \delta \min_{a'_i} Q'_i\{S_{\beta,i'} \mid S_{\beta,i}, a'_i, a'_u; \theta^{Q'_i}\} \quad (9)$$

finally for each UAV, the parameters of the Q-network θ^{Q_i} are copied into those of the target network $\theta^{Q'_i}$ after a predetermined number of iterations.

V. Numerical Results

In this section, we present network configurations and performance metrics. We investigate convergence of MUAIS, and evaluate the network cost of the proposed MUAIS scheme with regard to the network size. Here, the network cost defines the packet loss due to data buffer overflow and failed transmissions from the ground sensors to the UAVs.

A. Implementation of MUAIS

The simulation platform is a Lenovo Workstation running 64-bit Ubuntu 16.04 LTS, with Intel Core i5-7200U CPU @ 2.50GHz × 4 and 8 G memory. MUAIS is implemented in Python 3.5 by using Pytorch (the Python deep learning library). The region of interest is set to be a square area with a size of 1000 m x 1000 m, and 20 to 120 ground sensors are distributed in the region. Each ground sensor has the maximum discretized battery capacity 50 Joules, the highest modulation = 5, and the maximum transmit power 100 milliwatts. For calculating the power $P_j^i(t)$ of the ground sensor, the two channel constants, k_1 and k_2, are set to 0.2 and 3, respectively. The required BER is 0.05. The carrier frequency is 2000 MHz. ε is set to 0.05. The value of ε can be configured based on the traffic type and quality-of-service (QoS) requirement of the sensor's data, as well as the transmission capability of the UAV. Other simulation parameters are listed in Table II.

TABLE II
PyTorch Configuration

Parameters	Values
Number of ground sensors	20-120
Queue length	40
Energy levels	50
Discount factor	0.99
Learning rate	0.001
Replay memory size	10^6
Batch size	100
Number of episodes	1000

For performance evaluation, we use Random scheduling policy (RSA) and DRL-SA [7] algorithms as two baselines. In RSA, the UAVs schedule the ground sensors randomly and scheduling process is independent of the battery level and data queue length of the ground sensors, the channel quality, and the UAV's trajectory. In DRL-SA, a single UAV leverages DQN to schedule the ground sensors based on the data queue length, energy level, channel quality and UAV's trajectory.

B. Performance Evaluation

Fig. 3 shows the network cost at each episode of MUAIS with I=7 and DRL-SA. The MUAIS with I=7 has better performance than DRL-SA. The reason is that the involvement of multiple UAVs results in reduction of overflow cost. At the beginning of the learning process, MUAIS with I=7 witnesses a high network cost. With an increasing number of episodes, the network cost drops gradually until it stabilizes.

Fig. 4 studies the network cost with an increasing number of ground sensors, where the data buffer size of MUAIS is set

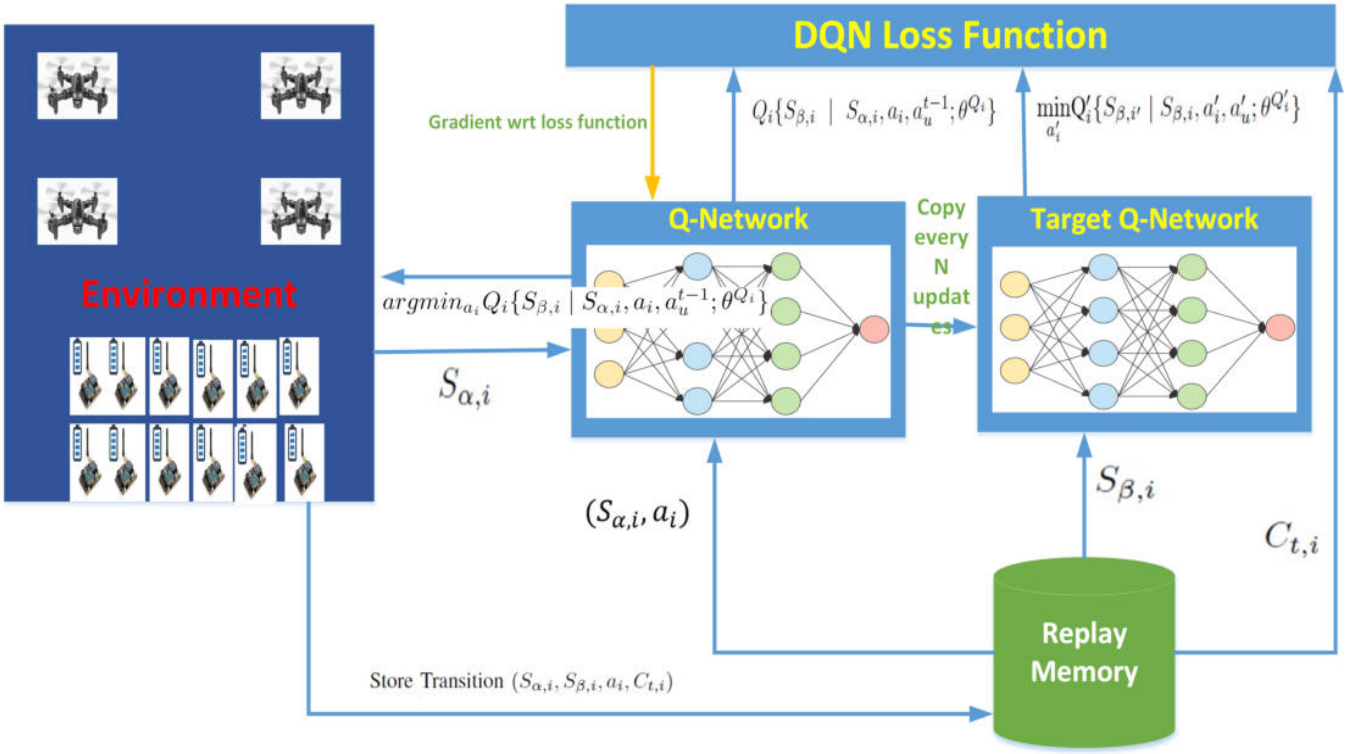

Fig. 2. The architecture of MUAIS for optimizing the selection of the ground sensor for data transmission.

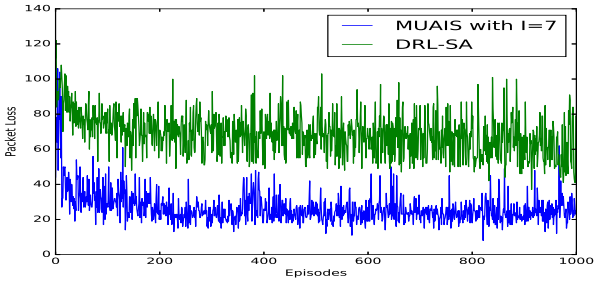

Fig. 3. The network cost at each episode of MUAIS with I=7 and DRL-SA.

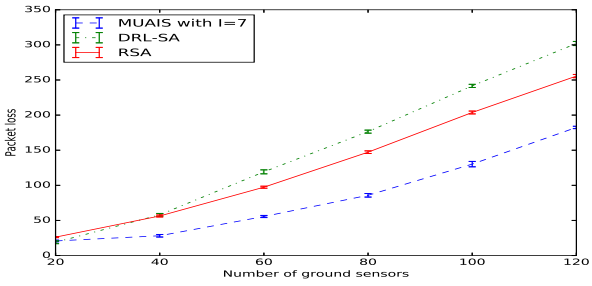

Fig. 4. Packet loss in terms of the number of ground sensors, where the standard deviation is achieved based on 30 experiments.

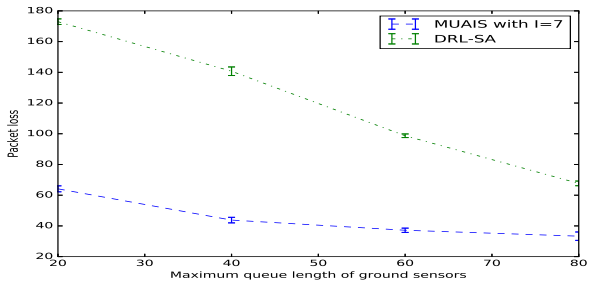

Fig. 5. Packet loss with regards to data queue sizes

to 40 packets. Overall, the proposed MUAIS with I=7 achieve lower network cost than RSA and DRL-SA. When the number of ground sensors is 20, the network cost of MUAIS is almost equal to that of DQN. The network cost of RSA is higher in this point. With increasing ground sensors, MUAIS receives the lowest cost. In contrast, RSA receives the highest cost and DQN is the middle. Particularly, when J=100, the packet loss of MUAIS with I=7 is lower than RSA and DRL-SA by around 35% and 46%, respectively.

Fig. 5 depicts the packet loss of MUAIS with I=7 when data buffer size is extended from 20 to 80. It is observed that MUAIS reduces the packet loss to a greater extent than DRL-SA. In particular, given D=20, MUAIS outperforms DRL-SA in term of the packet loss by 62%. From D=20 to D=80, the packet loss of MUAIS drops by 52%. This confirms that

MUAIS significantly reduces buffer overflow for all the ground sensors when enlarging their buffer length.

VI. CONCLUSIONS

In this paper, we studied data collection for preventing data queue overflow of the ground sensors. The problem was formulated as an MMDP, with the states of the battery level and queue length of the ground sensors, the channel conditions, and waypoints of the UAVs, to minimize the packet loss due to buffer overflows at the ground sensors and fading airborne channels. We proposed MUAIS to solve the MMDP, where all UAVs use DQN to conduct respective decisions. In MUAIS, the UAVs learn the underlying patterns of the data and energy arrivals at all the ground sensors, as well as the scheduling decisions of the other UAVs. PyTorch deep learning library was used for simulation and results confirm 46% and 35% reduction of the proposed MUAIS in packet loss as compared to az single-UAV case and an existing non-learning greedy algorithm, respectively.

VII. ACKNOWLEDGEMENT

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).

REFERENCES

[1] H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019.

[2] J. Kim, S. Kim, C. Ju, and H. I. Son, "Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications," *IEEE Access*, vol. 7, pp. 105 100–105 115, 2019.

[3] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, and K. Wong, "Uav-assisted emergency networks in disasters," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 45–51, 2019.

[4] Y. Gao, X. Chen, J. Yuan, Y. Li, and H. Cao, "A data collection system for environmental events based on unmanned aerial vehicle and wireless sensor networks," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 2020, pp. 2175–2178.

[5] C. Gómez and D. R. Green, "Small unmanned airborne systems to support oil and gas pipeline monitoring and mapping," *Arabian Journal of Geosciences*, vol. 10, no. 9, p. 202, 2017.

[6] "PyTorch," http://pytorch.org/, accessed: 2021-02-01.

[7] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep q-network for uav-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 215–12 226, Dec 2019.

[8] J. Tang, J. Song, J. Ou, J. Luo, X. Zhang, and K.-K. Wong, "Minimum throughput maximization for multi-uav enabled wpcn: A deep reinforcement learning method," *IEEE Access*, vol. 8, pp. 9124–9132, 2020.

[9] B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-based energy-efficient data collection by unmanned vehicles in smart cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1666–1676, 2017.

[10] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, 2018.

[11] Q. Wang, W. Zhang, Y. Liu, and Y. Liu, "Multi-uav dynamic wireless networking with deep reinforcement learning," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2243–2246, 2019.

[12] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for uav networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, 2020.

[13] A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, and J. Ashdown, "Distributed cooperative spectrum sharing in uav networks using multi-agent reinforcement learning," in *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019, pp. 1–6.

[14] U. Challita, W. Saad, and C. Bettstetter, "Deep reinforcement learning for interference-aware path planning of cellular-connected uavs," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7.

[15] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-uav assisted wireless networks: A machine learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, 2019.

[16] K. Li, W. Ni, E. Tovard, and A. Jamalipour, "Online velocity control and data capture of drones for the internet of things: An onboard deep reinforcement learning approach," *IEEE Vehicular Technology Magazine*, vol. 16, no. 1, pp. 49–56, 2020.

[17] Y. Emami, K. Li, and E. Tovar, "Buffer-aware scheduling for uav relay networks with energy fairness," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[18] 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*, vol. 15, no. 6, pp. 1377–1386, 2015.

[19] K. Li, Y. Emami, W. Ni, E. Tovar, and Z. Han, "Onboard deep deterministic policy gradients for online flight resource allocation of uavs," *IEEE Networking Letters*, vol. 2, no. 3, pp. 106–110, 2020.

[20] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[21] M. L. Littman, T. L. Dean, and L. P. Kaelbling, "On the complexity of solving markov decision problems," 2013.