ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 7-11 June 2020, Dublin, Ireland
DOI:10.1109/ICC40277.2020.9149151

Partially Explainable Big Data Driven Deep Reinforcement Learning for Green 5G UAV

Weisi Guo *IEEE Senior Member*

***Abstract*—UAV enabled terrestrial wireless networks enables targeted user-centric service provisioning to en-richen both deep urban coverage and target various rural challenge areas. However, UAVs have to balance the energy consumption of flight with the benefits of wireless capacity delivery via a high dimensional optimisation problem. Classic reinforcement learning (RL) cannot meet this challenge and here, we propose to use deep reinforcement learning (DRL) to optimise both aggregate and minimum service provisioning. In order to achieve a trusted autonomy, the DRL agents have to be able to explain its actions for transparent human-machine interrogation. We design a Double Dueling Deep Q-learning Neural Network (DDDQN) with Prioritised Experience Replay (PER) and fixed Q-targets to achieve stable performance and avoid over-fitting, offering performance gains over naive DQN algorithms. We then use a big data driven case study and found that UAVs battery size determines the nature of its autonomous mission, ranging from an efficient exploiter of one hotspot (100% reward gain) to a stochastic explorer of many hotspots (60-150% reward gain). Using a variety of telecom and social media data, our greener UAVs (30-40% energy saved) address both quantitative QoS and qualitative QoE issues with partial interpretability in the reinforcement learning actions during the course of learning and features extracted in some of the hidden layers, offering an initial step for explainable AI (XAI) connecting machine intelligence with human expertise.**

***Index Terms*—big data; machine learning; deep reinforcement learning; radio resource management; UAV; energy efficiency; XAI;**

I. Introduction

The wireless ICT industry is one of the fastest growing industries with millions of base stations providing service to billions of smart phones deployed worldwide, and collecting a large of amount of consumer data. To meet the rapidly increasing traffic volume in complex and demand environments, 5G and beyond mobile networks are expected to introduce a number of fundamental innovations apart from PHY and MAC layer technology enhancements. These advances complement existing PHY and MAC layer technologies via network densification, millimetre wave channels, and massive MIMO, *etc.*. To allow for centralised and large-scale network coordinated optimization, software defined network (SDN) and network function visualisation (NFV) have been proposed to optimise both the radio access network and mobile core network by integrating big data analytics and cloud control. Therefore, whilst real-time 5G radio resource management (RRM) remains critically important, but has become too complex for conventional optimisation. This brings the need to evolve towards an AI driven radio resource management (RRM) ecosystem [1], [2] to support more fine-grained user-centric service provision (see 3GPP Release 16 TR37.816). This fits into the wider ecosystem of AI driven tactile Internet for precision tasks [3] and the integration of big data into explainable AI reasoning.

Weisi Guo is with Cranfield University, Bedford, United Kingdom and Alan Turing Institute, London, United Kingdom. *Corresponding Author: wguo@turing.ac.uk.

A. UAV Enabled Wireless Networking

Unmanned Aerial Vehicles (UAVs) and drones have the autonomous capability to deliver tailored made wireless services to targeted areas [4]. They are an essential component to smarter cities, addressing a variety of urban infrastructure, surveillance, and personalised service needs.

Distinguishing it from earlier long endurance drone projects over a decade ago, 5G enabled UAVs now connect with massive MIMO mm-wave links and can operate on smaller and more efficient low-altitude platforms [5], [6]. Our own work [7] using triple validation of experimentation, ray-tracing simulation, and stochastic geometry have shown that micro-UAVs can provide flexible coverage solutions for poor coverage indoor entertainment venues, rural areas with rich diffraction loss, as well as deep fading and overcrowded urban areas [8]. Beyond wireless coverage, UAVs also provide vital physical delivery capability and potential for on-demand edge processing [9]. The aforementioned 5G RRM challenges become higher dimensional and difficult in highly dynamic environments involving UAV 3D heterogeneous channels. As a result, RRM is becoming increasingly complex and parameter optimization is a concern.

B. Deep Reinforcement Learning (DRL)

Classic reinforcement learning (RL) based solutions do not rely on accurate system models and is able to run in a model-free manner, which can be applied to online resource management. This in the past decade addressed some of the issues faced by traditional model dependent optimisation, such as dynamic programming and convex optimisation. However, traditional tabular RL approaches face the challenge in the scalability of the Q-table. When the state or action space becomes large ($10^3 - 10^{10}$), the sample complexity of RL approaches will result in non-convergence, incur excessive cost, even undermining the original learning gain.

The recent success of deep reinforcement learning (DRL) has opened new pathways to scalable optimization for high dimensional problems. DRL retains the model-free optimization capability of traditional reinforcement learning (RL), suitable for dynamic and online RRM. Meanwhile, in DRL, deep neural network (DNN) is used to approximate policy or

value functions for large-scale RL problem, overcoming the intrinsic scalability issue of traditional tabular RL approaches. The application of DRL in 5G and beyond [10] shows great promise and is receiving increased attention in the community at both the PHY and MAC layers. Specifically, the powerful function approximation and representation learning properties [11] of DNN empower RL with robust and high efficient learning. The application of DRL in 5G and beyond shows great promise and is receiving more and more attention in the community at both the PHY and MAC layers[1].

There has been a number of papers that have examined the use of AI and DRL in UAV assisted cellular communications [4], [12]. However, most existing DRL solutions applied in RRM use off-the-shelf algorithms with little consideration on the RRM feature set [13], [14]. This means that the resulting benefit and penalties incurred (e.g. latency and energy consumption) cannot be understood by the radio engineers monitoring and configuring the network. In order to achieve a trusted autonomy, the DRL agents have to be able to explain its actions for transparent human-machine interrogation [15]. The lack of full or partial interpretability / explainability in DRL is a major source of concern for transparent and explainable AI (XAI).

C. Novel Contribution, Big Data Case Study, & Organisation

Our novelty is to develop DRL that is proprietary to the task of UAV communications, where unstable traffic demand can cause erratic behaviour in UAV decisions. We design a Double Dueling Deep Q-learning Neural Network (DDDQN) with Prioritised Experience Replay (PER) and fixed Q-targets to achieve stable performance and avoid over-fitting. Then, a further key aspect of our novelty is to use big data to drive the DRL algorithm and UAV simulation, as well as offer explainability in the context of our input data. We examine: 1) 100+ base station data from London (over 3.2 million uplink and downlink demand requests), and 2) 430,000 geo-tagged Tweets; both posted in a 40km radius disc centred in Trafalgar Square, London over a recent 2 week period. Some of the data is available from Dryad and helped to shape earlier papers [16], [17].

In view of these challenges, this paper studies how to achieve partial explainable optimisation for joint UAV flight and RRM. We present a case study based on real experimental data. First, we set out the problem formulation and associated RL model in Section II. In Section III, we provide details on the neural network training and DRL process. In Section IV, we present our big data-driven case study results and we use a case study to partially explain our DRL results in a way that radio engineers can understand, offering a pathway to connect machine intelligence with human expertise and operational actions. Finally, we conclude and outline future challenges in Section V.

[1]see IEEE ComSoc Best Reading: https://www.comsoc.org/publications/best-readings/machine-learning-communications

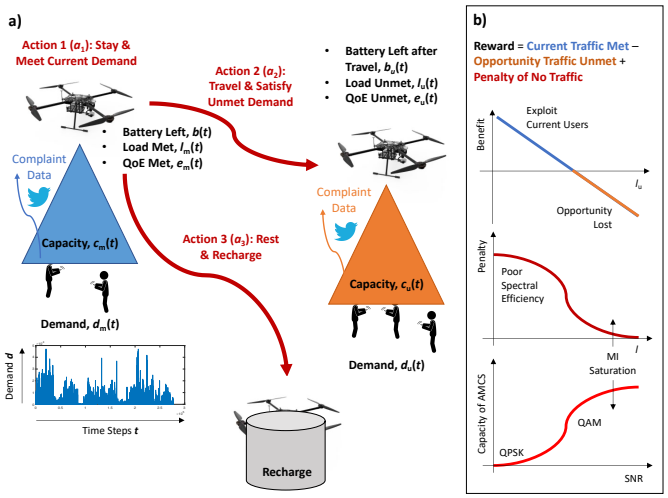

Fig. 1. System Model of UAV 5G Small-Cell: a) Action Space Choice of: 1 - Providing Service to Existing Users, 2 - Moving to Service New Users, 3 - Recharging; b) Reward based on Benefits and Penalty. Consumers demand traffic data d, and also provide real-time QoE complaint data via Twitter e, both of which we use as observable states to drive the DRL algorithm.

TABLE I
METHOD & DATA

System Parameter	Value
PHY Layer	5G AMCS [18]
Traffic Data, d	Real Traffic from BSs [16]
Complaint Data, e	Real Complaints from OSN [17]
Duration of Data	14 days
Frequency of Data	15 min
No. of BSs	105
No. of Hotspots, H	12
Area of Model	disk 40km radius
Area of Flight	disk 5km radius
Channel Model	UAV low-altitude urban [19]
Bandwidth	100 MHz
Drone Payload	1.36kg
Max Speed	48mph
Flight Time	78 min
Battery, B	Lithium $0.2 - 1.5$kWh

II. System Setup and Reinforcement Learning Method

A. UAV Enabled System

We consider a single 5G UAV small-cell that has the choice to service its existing users or meet the demand of new unmet users elsewhere (see Fig 1a). At each location, depending on the channel properties, the UAV can project a time-varying wireless capacity of c across its channels. This is to meet the traffic demand from users d, yielding a load of $l = d/c$ at any given location. Whilst the UAV can meet a traffic area, it is able to remove load from the unseen base stations (BSs) on the terrestrial cellular network. If it is not able to meet the load, it is extra demand on the BSs. Some of the system parameters are given in Table 1.

B. Q-Learning

In DRL (see Fig 2a), the key aspects are: (i) state, (ii) action, (iii) reward, (iv) a model of environment dynamics, (v) policy and (vii) DNN implementation [11]. The former four elements define the underling Markov Decision Process (MDP) of a classical RL system. The goal is to find a policy that maximise the expected sum of the rewards. The DNN is used to approximate the optimal policy or value function for large-scale RL problems.

States - states that includes important observations such as the remaining battery of the UAV $b_m \leq B$, the current met traffic load on its channels l_m, the current met customer QoE complains e_m, the highest unmet load l_u and complaints e_u at another location, and the expected battery left after usage to travel to the unmet load area b_u. An example of the observation states s_t (time-varying) for n UAVs for 2 hotspots (1 met, 1 unmet) is as follows at a particular time step t:

$$\begin{pmatrix} b_m(1) & ... & b_m(n) \\ b_u(1) & ... & b_u(n) \\ l_m(1) & ... & l_m(n) \\ l_u(1) & ... & l_u(n) \\ e_m(1) & ... & e_m(n) \\ e_u(1) & ... & e_u(n) \end{pmatrix}$$

We consider $n \in N$ UAVs serving H hotspots that can mutually observe each others' states. Our big data from base station loads l and social media complaints c inform the load model given in the observation states.

Actions - a set of possible actions that correspond to the state observed. There are three actions: (1) continue to serve the current traffic area, (2) travel to the new unmet area at a project battery cost of b_u, offloading existing traffic load l_m to a nearby BS and serving the new traffic load l_u, or (3) move to recharge battery and offloading existing traffic load l_m. Our associated assumptions are as follows: (A1) any traffic that is not met or offloaded serves as discounted outage across the network; (A2) the projected travel cost and traffic load demand at new location is time invariant and accurate and the UAV performs no wireless services en route, and (A3) the travel cost to recharge battery is negligible (e.g., many charge stations around city).

Rewards & Punishment - we develop a value-function RL approach, whereby the action-value function is defined by the long-term return when starting in some state with some action and following a given policy. Subsequently, are estimated and the optimal action(s) for each state correspond to the one(s) with the largest action-value.

We maximise the cumulative reward R defined by the load l and complaints e currently met minus the unmet load elsewhere (see Fig 1b):

$$\begin{aligned} R &= \sum_t \gamma^t r_t \\ r_t &= (1-\lambda)[l_m - l_u + e_m - e_u] + \lambda\rho(l_m), \end{aligned} \tag{1}$$

and the summation is for as long as the UAV can remain operational for and $\gamma = \{0, 1\}$ is the standard discount factor that increases with time, and the parameter λ balances the merits of meeting demand with the penalty of low spectrum efficiency:

$$\rho(x) = \tanh(Bx) - 1, \tag{2}$$

where the total battery capacity of B augments the strength of the penalty. A larger battery is expected to meet lower traffic loads at the same penalty. This highlights the expensive nature of UAV operations, and the need to reduce its operation when there is no traffic. The functional form is inspired by the approximations of discrete modulation and coding schemes (MCS), recognising the mutual information saturation at high SNRs [20]. As such the penalty incurs a high value for very low loads, and decreases rapidly as the load is near unity. We utilise Q-learning to learn a policy, as is standard for many wireless RRM work (including our own work on cell expansion in 2013 [21]). $Q(s, a)$ for state s and action a describes the Q-value of performing an action in a given state:

$$\begin{aligned} Q'(s,a) = \alpha[r + \gamma \max_a Q(s', a)] \\ + (1-\alpha)Q(s,a), \end{aligned} \tag{3}$$

where α is the learning rate, $'$ denotes the next state, $Q(\cdot)$ is updated until convergence, and $\max_a Q_n(s', a)$ is the maximum possible reward for all possible actions.

III. Deep Neural Network (DNN) Implementation

DNNs are used to exploit the potential correlation of states, actions and policies for efficient approximation of high-dimensional RL problem. We use fully connected networks for feature extraction in our Deep Q-Networks (DQN). The input to the NN at the feature extraction stage (see Fig. 2a) consists is the states $3H \times N$ array for N UAVs serving H hotspots with 3 states each. The first hidden layer is fully-connected Rectified Linear unit (ReLU) with $3H$ neurons and the second hidden layer is fully-connected ReLU with $3H$ neurons. In the double dueling (DD) stage (see below for description of DD and Fig. 2b), the value fully connected and advantage fully connected layers are a fully- connected with a single state $V(s)$ and a single output for each valid action $A(s, a)$.

A. Loss Function and Gradient Descent (GD)

We use the standard mean squared error (MSE) of the computed Q-value of the DNN to the target Q-value as the loss function. Mini-batch GD is used which splits the data into small batches that are used to calculate error and update coefficients: $\omega = \alpha\Delta\omega$, where $\Delta\omega = \partial L/\partial\omega$ and the loss function and target Q-value are given as:

$$\begin{aligned} L &= \frac{1}{2n}\sum[Q^*(s,a) - Q(s,a)]^2, \\ Q^*(s,a) &= R' + \gamma \max_{a+1} -Q(s', a'). \end{aligned} \tag{4}$$

Mini-batch GD seeks to find a balance between the robustness of stochastic GD and the efficiency of batch GD. The model update frequency is higher than batch gradient descent which allows for a more robust convergence - avoiding local minima, but all at the cost of an extra parameter.

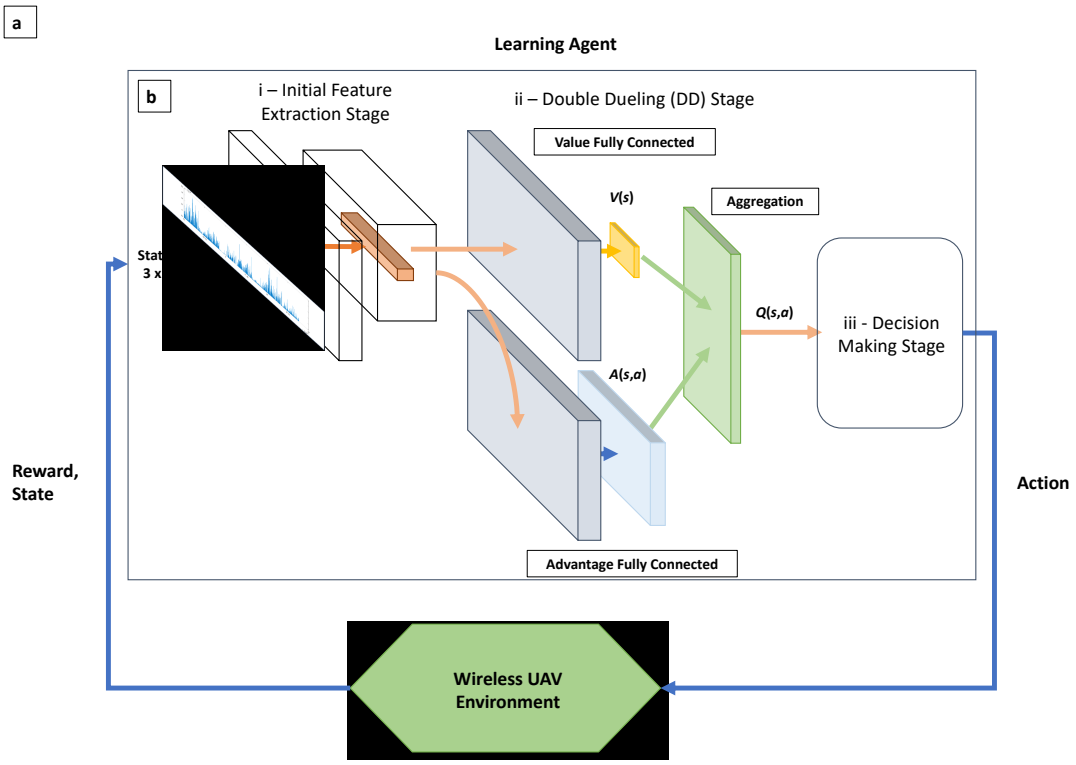

Fig. 2. DRL with DDDQN a) DRL architecture; b) DDDQN setup.

TABLE II
DRL HYPERPARAMETERS

Hyperparameter	Symbol	Value
Learning Rate	α	0.001
Discount Factor	γ	0.75
Mini Batch Size		24
Replay Memory Size	M	10240
Target Network Update		2 days
Initial Exploration		0.4
Final Exploration		0.2

B. DDDQN with PER and Fixed Q-Targets

Deep Q-learning (DQN) as previously described is known to learn unrepresentative high action values, because it includes a maximisation step over estimated action values, which tends to prefer overestimated to underestimated values, as it can be seen in Eq. 3. The idea of Double Deep Q-learning (DDQN) is to reduce overestimations by decomposing the max operation in the target into action selection and action evaluation. DDQN tries to decouple the action selection and evaluation. We employ a Double Dueling Deep Q-learning Neural Network (DDDQN) [22] with the following features to achieve several effects (e.g. stability, no overfitting):

1) Prioritised Experience Replay (PER): priority storage of experiences (state, action, reward, ...etc.) based on difference of the expected and target value. The probability of being selected is given by the stochastic prioritisation in Boltzmann form: $P(i) = \frac{p_i^{1/T}}{\sum_k P_k^{1/T}}$ [21], where $p(i)$ is the priority value of i being selected and hyper-parameter T is the temperature, which is a balancing factor between exploitation (T small) and exploration (T large). A replay memory size M is used to reduce

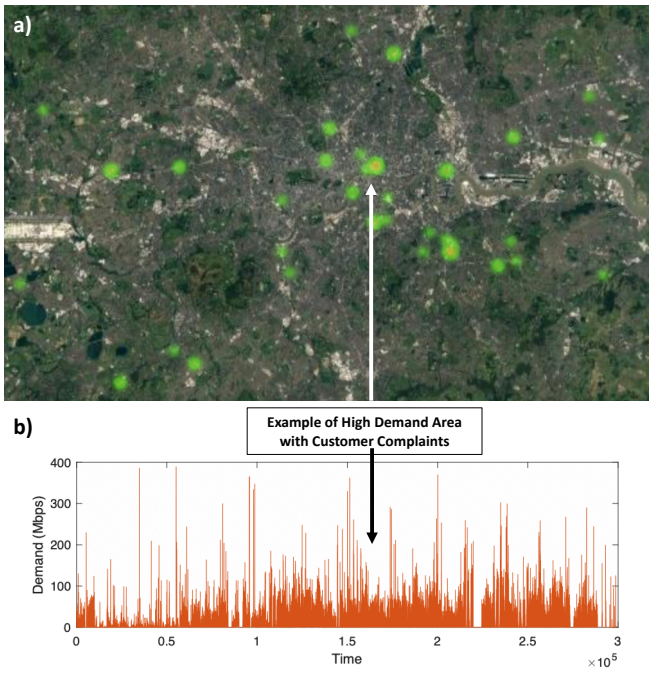

Fig. 3. London with areas of: a) high demand and poor customer service mined from Natural Language Processing of consumer complaints (see our earlier paper [17]); b) temporal demand profile of that area.

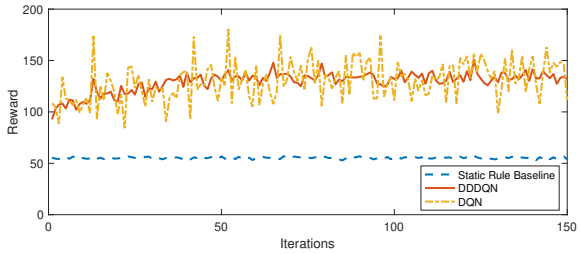

Fig. 4. Absolute Reward (R) as a function of learning iteration for baseline static rule UAV, DQN UAV, and DDDQN UAV.

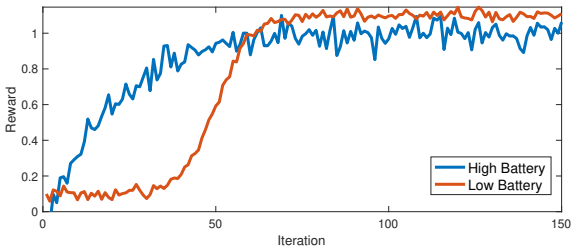

Fig. 5. Reward Gained (G) as a function of learning iteration for a high battery (1200 kWh) and a low battery (500 kWh) UAV.

the weights of frequently seen samples.

2) DQN with Fixed Q-Targets for Stability: A separate network with fixed parameter for estimating target value is used for stability, where the DQN update the target network. The change in weights is given by the predicted Q-value $\hat{Q}(\cdot)$:

$$\Delta w = \alpha[(r + \gamma \max_{a} \hat{Q}(s', a, \omega^{-})) - \hat{Q}(s, a, \omega))]\nabla_{\omega}\hat{Q}(s, a, \omega)), \quad (5)$$

where $\nabla_{\omega}\hat{Q}(s, a, \omega))$ is the gradient of the predicted Q-value of the target network.

3) Double DQN to Avoid Overfitting: We improve over taking the maximum Q-value (which will be noisy) towards avoiding the over-fitting problem by taking the value computed by the DQN network and the target Q-value of taking that action at the next state from the target network.

The hyper-parameters of the DRL replay buffer size are given in Table II.

In DDDQN, the value $Q(s, a)$ is computed as a combination of the value of being in a state $V(s)$ and the advantage of taking action at that state $A(s, a)$:

$$Q(s, a, \theta; a, b) = V(s; \theta, b) + \left[A(s, a; \theta, a) - \frac{1}{A}\sum_{a'} A(s, a'; \theta, a)\right] \quad (6)$$

where θ is the common network parameters, a is the advantage stream parameters, b is the value stream parameters. DDQN can learn the value of states without learning the impact of each action in that state – this avoids choosing the local maxima/minima. Fig 2b shows the DDDQN architecture, where the aggregation layer performs the Q-value estimation from the advantage and value functions, based on the aforementioned equation Eq.6.

IV. Big Data Driven Case Study

A. City Scale Telecom Demand Data & UAV Operational Scenario

We model the capacity of the N UAVs on an orthogonal 5G UAV channel using: Vienna 5G simulator [18] and a low altitude urban UAV statistical channel model [19]. We use two data sets to identify key areas in London pertaining to (see Fig. 3):

1) **QoS Hotspot** - high traffic demand (Quality-of-Service issue) from 100+ 4G base stations (BSs) to simulate load demand across Greater London [16].
2) **QoE Blackspot** - poor customer service mined from Natural Language Processing (NLP) of consumer complaints (Quality of Experience issue - see our earlier paper [17]);

We observe the rewards of one UAV to meet the demand between 1 or more peak demand situations with a fixed disk with radius of 5km. The goal of the RL agent is to maximise the reward and the performance KPIs are as follows: we define the relative reward gained and the energy efficiency as the following:

$$G = \frac{R_{\text{DRL}} - R_{\text{static}}}{R_{\text{static}}}, \quad E = \frac{D}{B - b_m}\text{bits/kWh}, \quad (7)$$

where D is a fixed demand traffic met. The parameters of the DDDQN is given in Table 2. A static **rule-based baseline** is established whereby the UAV periodically goes between recharge and a nearby hotspot without data-driven adaptation.

B. Results with Explainability

Our baseline results in Fig. 4 show the absolute reward R from: a) the static baseline UAV operation, b) the DQN driven UAV which suffers from over-fitting and unstable performance, and c) the DDDQN driven UAV which is more stable partly due to the PER and fixed Q-value.

1) Efficient Exploiter and the Fast Learning Explorer: Our results first look at how results vary with learning iteration in Fig. 5 for both a low battery and high battery capacity UAV. We see that the low battery UAV is unable to achieve any significant reward gain (below 20%) until iteration 50, because it is too conservative in its approach to save battery power and avoid the penalty in Eq.2. This is also likely to be because it cannot find a satisfactory demand hotspot to serve, but once it does beyond iteration 50, it is able to achieve a very exploitative behaviour, oscillating between saving battery and exploiting the nearest hotspot of demand – achieving a consistently high reward over a static case - 110%. For the high battery case, the UAV has the energy capacity to rapidly exploit hotspots over a range of areas and leads to a steady rise in reward over time, but cannot exceed 100% gain reliably due to its more explorative nature. Our partial explainability is that the DRL agent actually enables a smaller battery UAV to more reliably serve the demand hotspot, but it

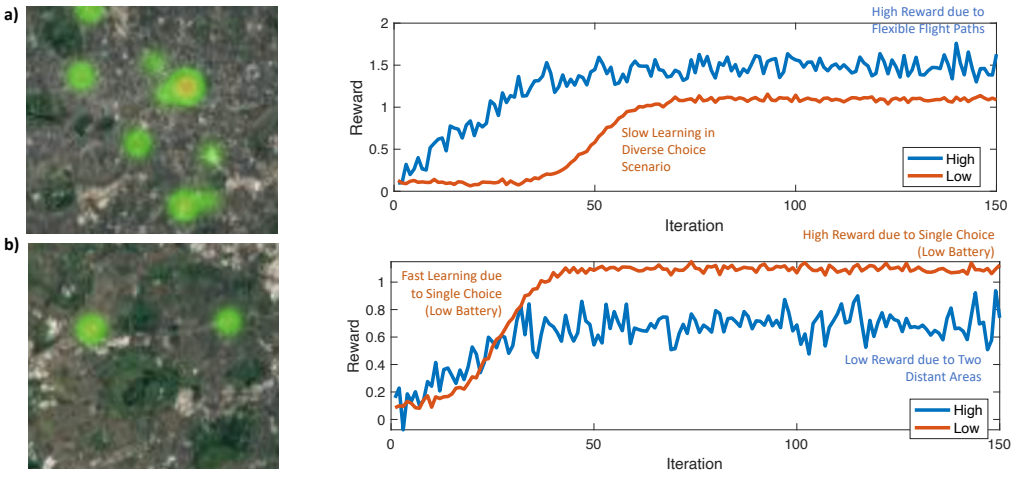

Fig. 6. Two diverse urban locations: a) numerous demand hotspots, and b) two distant demand hotspots.

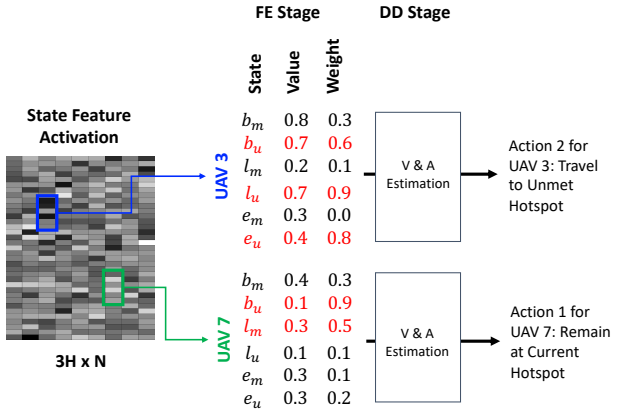

Fig. 7. Partial Explainability with Weights of States from feature extraction (FE) stage of DDDQN DRL interpreted for 2 UAVs.

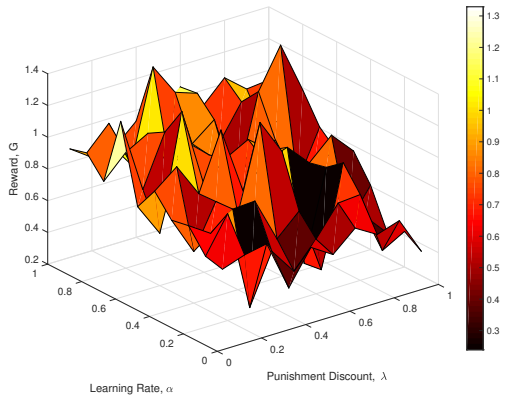

Fig. 8. Reward Gained (G) as a function of punishment discount factor λ and learning rate α.

has a delay in learning the right course of actions. However, the UAV with a low battery never ventures beyond its small area (efficient exploiter), and is unable to track dynamic demands very well. On the other hand, the larger battery UAV (fast learning explorer) learns quickly and converges on a 10% smaller reward than the small battery UAV, but can serve a wide area.

2) Impact of Hotspot Spatial Distribution: Our results next look at reward variation with learning iteration in Fig. 6 for two diverse urban locations: a) numerous demand hotspots in Westminster area, and b) two distant demand hotspots near Heathrow Airport area. We show that for a), the high battery UAV is able to achieve a higher performance due to the close proximity of many hotspots and it is able to leverage on its larger battery to serve a number of demands over time with 150% reward. However, the smaller battery UAV in comparison can only achieve 110% as before with 1-2 hotspots. For case b), two distant hotspots prove inefficient for the high battery UAV, achieving only 60% reward gain with high variational behaviour. The smaller battery UAV in comparison can only achieve a consistent 110% as before with only 1 hotspots, and more importantly a rapid learning rate due to the lack of choice.

C. Explainable AI (XAI) Impact on Design

Using a $H = 12$ hotspot and $N = 10$ UAV example, we attempt to interpret the feature extraction stage of DDDQN DRL interpreted for 2 UAVs in Fig. 7, where we highlight the high weights (≥ 0.5). We can see that the high remaining battery and the relatively higher load and complaint of the unmet hotspot drove UAV-3 to move to the new position and abandon its current hotspot. Conversely for UAV-7, the fact that it would have very little battery left if they move. This interpretation supports our general learning results shown above that battery power remaining is a key determinant of the UAV's actions.

Our partial XAI have shown that when there are diverse demand hotspots in close proximity (e.g. Central London), a larger UAV with high battery is suitable and can achieve a very high reward gain. However, when there is only 1-2 hotspots or hotspots separated by larger distances, then a number of dedicated smaller UAV with smaller batteries

tailored for single hotspots is more suitable (e.g., Heathrow Airport). As a result of the more optimal learning actions, we have been able to reduce the energy used by the UAVs for the same targeted data rate provisioning by 30-40%, leading to a much more energy efficient and green ecosystem. However, we have not considered the impact of energy costs in our DRL algorithm, and the debate between energy efficiency and AI is left open [23].

Our results for reward variation with punishment for poor spectral efficiency (SE) and learning rate in Fig. 8. Here, we can see that in general a high learning rate is preferable given the stable behaviour of the DDDQN and the lack of overfitting in the Q-learning process. The punishment factor for poor SE utilisation can actually vary across the a non-zero range and is not an important parameter as first thought in the design phase, further demonstrating how explainability of the DRL is important.

V. Conclusion & Future Challenges

This paper presents a DRL based on a big data driven approach to control an UAV based 5G system. UAVs are an essential component to smarter cities, addressing a variety of urban infrastructure, surveillance, and personalised service needs, but their ability to efficiently perform tasks in an explainable autonomous manner and their high energy demand is a critical problem. First we showed that our DDDQN data-driven UAV is proven to be significantly more efficient than classic rule-based UAVs, and also more stable than naive DQN based UAVs. We then use a big data driven case study and found that UAVs battery size determines the nature of its autonomous mission, ranging from an efficient exploiter of one hotspot (100% reward gain) to a stochastic explorer of many hotspots (60-150% reward gain). Using a variety of telecom and social media data, our greener UAVs (30-40% energy saved) address both quantitative QoS and qualitative QoE issues with partial interpretability in the reinforcement learning actions during the course of learning and in some of the hidden layers (in terms of features regarded as influential in decisions), offering an explainable AI (XAI) pathway to connect machine intelligence with human expertise with impact on operational actions. Future work will focus on full explainability, based on opening up the DDDQN value and advantage stream layers to understand which features are being propagated, the estimation of value based on state information, and their influence on the subsequent actions.

Acknowledgement: This paper is partly funded by EC H2020 grant 778305: Data Aware Wireless Networks for Internet-of-Everything and also funded by UK Geospatial Commission. I wish to thank team members Dr. W. Qi and Dr. S.Chotvijit on their map of hotspots in London.

References

[1] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5g: When cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, October 2017.

[2] D. Wang, B. Song, D. Chen, and X. Du, "Intelligent cognitive radio in 5g: Ai-based hierarchical cognitive cellular networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 54–61, June 2019.

[3] M. Dohler, T. Mahmoodi, M. A. Lema, M. Condoluci, F. Sardis, K. Antonakoglou, and H. Aghvami, "Internet of skills, where robotics meets ai, 5g and the tactile internet," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[4] M. Mozaffari, A. Taleb Zadeh Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5g with uavs: Foundations of a 3d wireless cellular network," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 357–372, Jan 2019.

[5] B. Li, Z. Fei, and Y. Zhang, "Uav communications for 5g and beyond: Recent advances and future trends," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2241–2263, April 2019.

[6] M. Gapeyenko, V. Petrov, D. Moltchanov, S. Andreev, N. Himayat, and Y. Koucheryavy, "Flexible and reliable uav-assisted backhaul operation in 5g mmwave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2486–2496, Nov 2018.

[7] W. Guo, C. Devine, and S. Wang, "Performance analysis of micro unmanned airborne communication relays for cellular networks," in *2014 9th International Symposium on Communication Systems, Networks Digital Sign (CSNDSP)*, July 2014, pp. 658–663.

[8] Q. Zhang, M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Machine learning for predictive on-demand deployment of uavs for wireless communications," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.

[9] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a uav-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2049–2063, March 2018.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2019.

[11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov 2017.

[12] U. Challita, A. Ferdowsi, M. Chen, and W. Saad, "Machine learning for wireless connectivity and security of cellular-connected uavs," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 28–35, February 2019.

[13] 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)*, May 2018, pp. 1–7.

[14] ——, "Interference management for cellular-connected uavs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, April 2019.

[15] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie, "A reinforcement learning framework for explainable recommendation," in *2018 IEEE International Conference on Data Mining (ICDM)*, Nov 2018, pp. 587–596.

[16] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, "Estimating mobile traffic demand using twitter," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 380–383, Aug 2016.

[17] W. Qi, R. Procter, J. Zhang, and W. Guo, "Mapping consumer sentiment toward wireless services using geospatial twitter data," *IEEE Access*, vol. 7, pp. 113 726–113 739, 2019.

[18] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S. Schwarz, and M. Rupp, "Versatile mobile communications simulation: the vienna 5g link level simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 226, Sep 2018. [Online]. Available: https://doi.org/10.1186/s13638-018-1239-6

[19] M. Simunek, F. P. Fontan, and P. Pechac, "The uav low elevation propagation channel in urban areas: Statistical analysis and time-series generator," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 7, pp. 3850–3858, July 2013.

[20] W. Guo, S. Wang, and X. Chu, "Capacity expression and power allocation for arbitrary modulation and coding rates," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 3294–3299.

[21] W. Guo and T. O'Farrell, "Dynamic cell expansion with self-organizing cooperation," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 851–860, May 2013.

[22] T. H. et al., "Deep q-learning from demonstrations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, April 2018.

[23] Z. Du, Y. Deng, W. Guo, A. Nallanathan, and Q. Wu, "Green deep reinforcement learning for radio resource management: Architecture, algorithm compression and challenge," *under submission - IEEE Communications Magazine*, 2019.

Partially explainable big data driven deep reinforcement learning for green 5G UAV

Guo, Weisi

2020-07-27

Guo W. (2021) Partially explainable big data driven deep reinforcement learning for green 5G UAV. In: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 7-11 June 2020, Dublin

https://doi.org/10.1109/ICC40277.2020.9149151