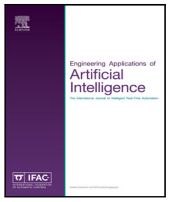
Engineering Applications of Artificial Intelligence 102 (2021) 104234

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survey paper

Deep Reinforcement Learning for QoS provisioning at the MAC layer: A Survey

Mahmoud Abbasi [a], Amin Shahraki [b,d,*], Md. Jalil Piran [c,**], Amir Taherkordi [b]

[a] Department of Computer Sciences, Islamic Azad University, Mashhad, Iran
[b] Department of Informatics (IFI), University of Oslo, Oslo, Norway
[c] Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[d] Faculty of Computer Sciences, Østfold University College, Halden, Norway

ARTICLE INFO

Keywords:
Quality of Service
Medium Access Control
Rate control
Resource sharing and scheduling
Deep Reinforcement Learning
Survey

ABSTRACT

Quality of Service (QoS) provisioning is based on various network management techniques including resource management and medium access control (MAC). Various techniques have been introduced to automate networking decisions, particularly at the MAC layer. Deep reinforcement learning (DRL), as a solution to sequential decision making problems, is a combination of the power of deep learning (DL), to represent and comprehend the world, with reinforcement learning (RL), to understand the environment and act rationally. In this paper, we present a survey on the applications of DRL in QoS provisioning at the MAC layer. First, we present the basic concepts of QoS and DRL. Second, we classify the main challenges in the context of QoS provisioning at the MAC layer, including medium access and data rate control, and resource sharing and scheduling. Third, we review various DRL algorithms employed to support QoS at the MAC layer, by analyzing, comparing, and identifying their pros and cons. Furthermore, we outline a number of important open research problems and suggest some avenues for future research.

1. Introduction

The development of networking technologies and paradigms such as cellular networks (Shahraki et al., 2021), Wireless Sensor Networks (WSNs) (Shahraki et al., 2020d), and the IoT (Stoyanova et al., 2020), along with emerging low-cost and small wireless devices supporting various communication protocols have enabled connecting billions of smart devices (Kaur et al., 2020). It is expected to have more than 21.5 billion connected devices around the world in 2025 (Özyılmaz and Yurdakul, 2020). The networking technologies allow different types of smart devices, *e.g.,* home appliances, vehicles, sensors, etc. to connect and host divers ubiquitous applications. The underlying networks need to support various QoS (White et al., 2017) requirements based on the priorities of applications and network traffic (Kaur and Kumar, 2019; Ayyasamy and Venkatachalapathy, 2015). Although QoS is a general term, it mainly refers to the use of appropriate techniques to achieve the Service Level Agreement (SLA) that is agreed between users and the network management agents. This can be evaluated using different parameters extracted from network traffic, *e.g.,* latency, jitter, and throughput. Although theoretically SLA must be fully supported to satisfy QoS requirements, in reality three commitment levels of QoS can be envisaged:

- *Best-effort Service (No QoS)*: In this case, network management agents try to provide the highest possible QoS for users, but there is no guarantee.
- *Guaranteed Service (Hard QoS)*: In this case, the network infrastructure must support the SLA. This requires reserving all resources for data transmission permanently regardless of whether the network traffic flow is running or not.
- *Differentiated Service (Soft QoS)*: In this case, the required QoS is not guaranteed, but resources are reserved, renewed, and released based on network traffic requirements.

Soft QoS is the most efficient commitment model in terms of resource efficiency as it allows introducing some techniques to manage the resources dynamically (Adhikari et al., 2019; Logambigai and Kannan, 2014). The definition of *resource* is very general in the field of wireless networks as it can vary from the Central Processing Unit (CPU) resources of machines to spectrum. As a general definition, QoS can be supported by each layer of wireless networks that manages some resources of the network. As an example, reducing delay can be performed by the MAC, network, and application layers through channel assignment, routing techniques (Kalidoss et al., 2020), and priority-based queue management in middlemen nodes, respectively. There are

* Corresponding author at: Department of Informatics (IFI), University of Oslo, Oslo, Norway.
** Corresponding author at: Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea.
E-mail addresses: am.shahraki@ieee.org (A. Shahraki), piran@sejong.ac.kr (Md. Jalil Piran).

https://doi.org/10.1016/j.engappai.2021.104234
Received 14 January 2021; Received in revised form 25 February 2021; Accepted 23 March 2021
Available online 16 April 2021

Nomenclature

UE	User Equipment
CDN	Content Delivery network
ABR	Adaptive Bitrate
QoE	Quality of Experience
DDPG	Deep Deterministic Policy Gradient
DASH	Dynamic Adaptive Streaming over HTTP
C-RAN	Cloud Radio Access Network
RRh	Remote Radio Head
NOMA	Non-Orthogonal Multiple Access
CR	Cognitive Radio
SVM	Support Vector Machine
AP	Access Point
LBT	Listen-Before-Talk
SU	Secondary User
QoS	Quality of Service
MAC	Medium Access Control
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
HetNet	Heterogeneous Networks
DL	Deep Learning
ML	Machine Learning
IoT	Internet of Things
IoV	Internet of Vehicle
WSNs	Wireless Sensor Networks
SDN	Software-Defined Networking
MDP	Markov Decision Process
POMDP	Partially Observable MDP
DQN	Deep Q-Network
CPU	Central Processing Unit
CSI	Channel State Information
i2A	Imagination-Augmented Agents
MBMF	Model-Based RL with Model-Free Fine-Tuning
MBVE	Model-Based Value Expansion
PG	Policy-Gradient
A2C	Advantage Actor–Critic
A3C	Asynchronous Advantage Actor–Critic
MSE	Mean Square Error
NN	Neural Network
HetNet	Heterogeneous Network
UAV	Unmanned Aerial Vehicle
PER	Prioritized Experience Replay
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ISP	Internet Service Provider
CIoT	Cognitive Internet of Things (IoT)
V2V	Vehicle to Vehicle
D2D	Device to Device
V2I	Vehicle to Infrastructure
WSN	Wireless Sensor Network
LR-WPAN	Low-Rate Wireless Personal Area Network
DRA	Dynamic Resource Allocation
MBS	Multi-Beam Satellite
CNN	Convolutional Neural Network
SON	Self-Organizing Network
SSN	Self-Sustaining Network
SLA	Service Level Agreement
eNB	evolved Node B
MARL	Multi-Agent RL
AIoT	Autonomous IoT
AI	Artificial Intelligence
DSA	Dynamic Spectrum Access

different technologies and techniques for QoS support from the application layer to the physical layer, *e.g.,* Software-Defined Networking (SDN) (Benzekki et al., 2016), cognitive radio management (Amjad et al., 2018), Machine Learning (ML) techniques (Klaine et al., 2017), etc. New networking paradigms face various challenges, *e.g.,* resource management and improving the network performance. Many networking challenges can be addressed by optimizing the medium access, thereby techniques that improve MAC layer performance have attracted much attention. Various performance aspects of the MAC layer can be improved, which can directly affect QoS, mainly network resource management and network access. To enhance the performance of such techniques, different technologies have been introduced, *e.g.,* automated decision-making techniques, which have attracted much attention, especially those based on ML techniques.

ML-based sequential decision-making (a.k.a RL) is one of the most important research directions in ML, as the model takes the dynamicity of the world of decision–actions into consideration (Sutton and Barto, 2011). Unlike other ML techniques, RL interacts with the environment to improve the efficiency of applying decisions made by the ML model. Simply, sequential decision-making means the model receives feedback from the environment after performing an action and updates its intelligence regarding the policies for decision making. RL has gained increasing attention thanks to its success in fulfilling difficult sequential decision-making needs, especially in dynamic and complicated environments.

Generally speaking, in RL models, an agent interacts with its environment, decides to take an action, and receives a reward based on the effect of the action on the environment. Then, it uses the newly earned experience to improve its decision-making policies to maximize its cumulative rewards based on a Myopia factor (Kaelbling et al., 1996). The environment refers to the entities that the decisions can affect them, *e.g.,* a heater can be considered as an "agent" and your home as an "environment". Although at the first glance, the agent incrementally learns new behaviors and strategies by interacting with its environment, the agent does not deeply depend upon full knowledge of its environment and can make a decision based on the limited cognition of the environment. As the main drawback of RL, the learning process takes considerable time. Another key challenge of RL is that it learns by the trial-and-error technique, which means that the agent possibly takes actions randomly to monitor the negative or positive reflects of the environment which causes inefficiency in result-sensitive decision-making environments, *e.g.,* smart vehicles.

In RL techniques, the agent needs to explore the environment based on a trial-and-error vision to provide information to exploit. Nevertheless, in RL, the trade-off between *exploration* (*i.e.,* gathering more information) and *exploitation* (*i.e.,* making the best decision given the current knowledge) is a fundamental issue. It can arise when one encounters an interactive decision-making problem as the decisions are not efficient when lacking the required information. RL techniques can be used in different environments for solving decision-making problems, *e.g.,* traffic light control, robotics (Gu et al., 2017), bidding and advertising (Wu et al., 2018), and networking (Luong et al., 2019). Unlike others, networking is a fast-paced changing environment in which the behavior of network traffic changes frequently. Due to the time-consuming learning process, RL is often unsuitable for and inapplicable to large-scale and dynamic networks which change their behavior frequently. Moreover, in large-scale ubiquitous network systems, the

volume of possible states and actions accessible to the agent is usually large, and RL may not find the optimal policy at an appropriate time.

To overcome the limitations of RL, a combination of RL with other modern and more efficient ML techniques, *e.g.,* deep neural networks, can be efficient. One attractive combination is RL with Deep Learning (DL) techniques, called DRL. In simple words, DL is a subset of ML in which the designed algorithms are inspired by the human brain functionality and can learn from vast amounts of data (*e.g.,* structured data, visual data, voice and text). DRL is most effective in tasks with high dimensional state-space that have a high volume of unlabeled or semi-labeled data to learn. Moreover, DRL solves the choice of features problem in RL due to its ability to extract complex abstractions as data representations. Consequently, DRL improves the speed of the learning process and the performance of RL. Unlike the limited real-world applications of RL, DRL has been successfully used in numerous networking applications (Luong et al., 2019).

As the main concern of this study in the context of networking, DRL has been used as a promising solution to many key challenges and problems from the application layer to the physical layer. In modern networking paradigms such as Heterogeneous Network (HetNet) (Zhang et al., 2019), the decentralized decision-making feature of DRL can be used to address the major challenges of such networks, *e.g.,* the coexistence of multiple wireless links of different users (Zhang et al., 2019). IoT (Shahraki et al., 2020c) and Internet of Vehicle (IoV) (Abbasi et al., 2021a) as two modern communication networks can also benefit from DRL when making local and autonomous decisions for configuring different network layers and supporting QoS.

There are several parameters for QoS provisioning in networks, *e.g.,* latency, network throughput, and jitter, to name a few. QoS is a broad topic that can be interpreted differently for different network layers. It can be achieved using various techniques such as managing network topology, configuring data rate, transmitting power control, joint user association, etc. It is worth noting that the optimum network configuration can dramatically affect QoS metrics such as throughput, delay, and data loss. In this paper, we focus on QoS support at the MAC layer, in particular, the role and advantages of DRL in this context, as listed below:

- In large-scale networks, the design and deployment of a centralized decision-making mechanism under uncertain and stochastic deployment environments is a resource-demanding task since a massive volume of data should be transferred to a central point causing high network overhead (Darivianakis et al., 2018). In such large-scale networks, the design of a decentralized and local decision-making mechanism is desirable—relying only on local information by the network nodes (Bannour et al., 2018). This is mainly due to the decentralized nature of these networks, *e.g.,* IoT networks (Luong et al., 2019). In other words, in such systems, local decision-making is desirable for network entities to promote the network performance under the uncertainty of the network environment (Lei et al., 2020). DRL techniques can address this challenge as they can interact with the networking environment, therefore each node that applies DRL techniques can make decisions locally (Dai et al., 2019; Luong et al., 2019). In other words, the interaction between the local networking environment and the DRL models can reduce the demand for global information for decision-making, resulting in reducing the communication overhead (Lei et al., 2020; Chen et al., 2020a). Interacting with the networking environment can potentially address the challenges of lacking global knowledge for decision-making in networks. Autonomous decision-making can be achieved by DRL. Therefore, network nodes can interact with their environments and adopt the best policy locally (Hoel et al., 2019). Besides, DRL can also decline the computational complexity of the previously proposed methods such as dynamic programming, value-iteration algorithms, and RL—an advantage to perform ML techniques in local resource-constrained machines (Arulkumaran et al., 2017).
- Considering the exponential growth in the number of networked devices, networking ecosystems face severe challenges in handling such massive numbers of devices and the generated big data. DRL techniques can introduce satisfactory and smart solutions for sequential decision-making, control problems, and optimization by integrating the abilities of DL to solve complex issues and the capabilities of RL to interact with the environment. Unlike other applications of ML, addressing the challenges of networking needs a tight interaction between the ML model and the networking environment (Shahraki et al., 2020a). Modern networking paradigms suffer from the lack of a unified theory of networking, thereby the ML models should be trained for each network separately (Ayoubi et al., 2018). DRL has the potential to address networking challenges as it can adapt the model with each unique networking environment interactively.
- In wireless networks, the efficient management of available resources, such as transmit power, subcarriers, time slots, and antennas is crucially important due to their enormous impact on the overall performance of the network. By adopting DRL for cellular networks, the following key advantages can be achieved: (1) a moderate-sized DRL model is able to quickly make forecasting because DRL models need a few number of simple operations to produce an output, (2) DRL is ideal for addressing incomplete information such as Channel State Information (CSI) since DRL agents can learn a long-term strategy by considering long-term rewards (Mao et al., 2018), and (3) model-free DRL, as a general and powerful technique, can be easily applied to a variety of contexts for learning complex behaviors (Sun et al., 2018).
- There is a good opportunity to integrate DRL and Recurrent Neural Network (RNN) models. Existing literature, *e.g.,* (Lin et al., 2020; Li et al., 2015; Lu, 2017) shows that by integrating DRL and RNN models, *e.g.,* Long Short-Term Memory (LSTM), DRL techniques can update their models based on the network behavior. Although there is a lack of literature on the use of DRL and RNN to solve the networking challenges, especially QoS-provisioning, based on the networking characteristics, they can solve serious problems as most of the modern networking paradigms change their behavior frequently based on different events, *e.g.,* intrusions, changes in network traffics and nodes joining and leaving (D'Alconzo et al., 2019).

The significant advantages of DRL models in addressing the networking challenges provide a great opportunity to present an overview of DRL applications in supporting QoS at the MAC layer (White et al., 2017) given its important role in improving different QoS metrics, *e.g.,* delay, throughput, reliability, etc. The related applications are categorized into two main categories, including network access and data rate control, and resource sharing and scheduling (see Fig. 1). The main contributions of this paper are summarized as follow:

- Presenting the concepts of QoS and DRL.
- Classifying the main topics in the field of providing QoS at the Medium Access Control (MAC) layer.
- Studying different DRL algorithms that are used in the field of QoS provisioning at the MAC layer.
- Providing a concrete list of open research problems and future directions in this field.

The rest of the paper is organized as follows. Section 2 discusses existing survey articles in this area. The basic concepts of RL, DRL and QoS are provided in Section 3. The applications of DRL for network access and data rate control, and resource sharing and scheduling are reviewed in Sections 4 and 5, respectively. In Section 6, we outline some important challenges and future research directions. Finally, Section 7 draws the conclusion.

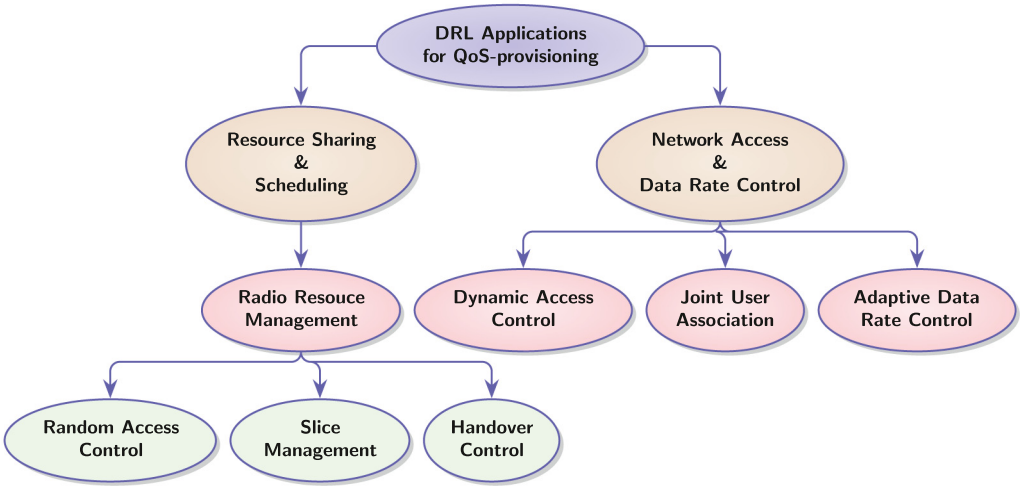


Fig. 1. A taxonomy of DRL applications for QoS-provisioning.

2. Review of related survey articles

To the best of our knowledge, this is the first paper that investigates the relation between DRL and QoS-provisioning at the MAC layer. A limited body of work exists that focuses on the applications of DRL, DL, and other classical ML models in the general area of communication systems and networks. However, our work is the first investigating the applications of DRL in networking from the QoS perspective. The closest paper to our work is the one conducted by Luong et al. (2019). In the paper, the authors provide a comprehensive review of DRL's applications in networking, such as network access/rate control, caching/offloading, and security/connectivity preservation. Nevertheless, the focus of that paper is not QoS-provisioning.

In Zhang et al. (2019), Zhang et al. surveyed the convergence of wireless and mobile networks and DL. The authors first provide a background on DL methods and their potential applications in networking. They also review various techniques and platforms that can be used to deploy DL over mobile networks. Moreover, a comprehensive list of papers related to mobile/wireless networks and DL is provided and classified based on different domains. A similar work has been conducted by Mao et al. in Mao et al. (2018). In this study, the applications of DL in different layers of wireless networks (*i.e.,* PHY, data link, network and upper layers) and network security are investigated. However, in these two surveys, the authors did not investigate the applications of DRL in wireless networks.

The work in Al-Garadi et al. (2020) by Al-Garadi et al. conducted an in-depth survey on classical ML and DL techniques for the security of IoT. The authors investigated different ML/DL-based techniques proposed for improving the security of IoT systems, and also they highlighted the pros and cons of each technique. That survey is different from our work because it focuses on IoT systems, and does not cover DRL techniques, while it reviews classical ML techniques for IoT security. In Mohammadi et al. (2018), DL models and their architectures for IoT big data and streaming data are studied by Mohammadi et al. Their survey covers DL-based methods that have been used to improve the analytics and learning from big and streaming data in IoT networks. The article also discusses the unique characteristics of IoT data (*i.e.,* 6V's features), and DL is proposed as a powerful ML model to deal with analyzing this type of data. We differentiate between that work and our paper since in Mohammadi et al. (2018) the authors focus only on DL models and the applications related to big/stream data analytics in IoT systems.

Lei et al. (2020) reviewed models, applications, and challenges of DRL in Autonomous IoT (AIoT). They refer to this fact that RL and DRL are two promising techniques to realize autonomy in AIoT systems. Regarding this fact, the authors first provide a general model for RL and DRL applications in AIoT systems. Then, they survey an extensive body of literature on DRL for AIoT systems and discuss research challenges and open issues for future research. Fadlullah et al. conducted a survey on the applications of DL in network traffic control systems (Fadlullah et al., 2017). In particular, they highlight DL's strengths and weaknesses for many aspects of network traffic control. Furthermore, the authors identify some unsolved research problems worth further research. However, the paper focuses only on old DL models as it was published three years ago.

Ahad et al. introduced Neural Networks (NN) as a way to tackle challenges related to decision making or parameter optimization in wireless networks, *e.g.,* the highly dynamic and unknown network environments (Ahad et al., 2016). One of the paper's severe disadvantages is that it focuses on traditional Neural Network (NN) systems. It also ignored recent advancements in DL models, and their successful applications in current communication systems and networks. Abu Alsheikh et al. reviewed the areas of WSNs and ML in Alsheikh et al. (2014). In the paper, ML-based techniques are discussed as useful tools for WSNs to adapt themselves to the network's dynamic environments and behaviors, resulting in resource utilization and increasing the network's lifetime. Moreover, the paper recognizes some research gaps between ML and WSNs that remain uninvestigated. Nevertheless, it pays full attention to the old classical ML algorithms such as Support Vector Machine (SVM), k-nearest neighbors algorithm (k-NN), and Bayesian, ignoring DL algorithms.

Bkassiny et al. discuss the potential of using ML (supervised and unsupervised models) related to Cognitive Radio (CR) issues, including decision making and feature classification (Bkassiny et al., 2012). The authors also identify several challenges related to the learning process in CR networks and highlight similarities and differences amongst the reviewed ML models. Klaine et al. deliver a survey on ML algorithms applied on Self-Organizing Networks (SONs), examine the advantages and disadvantages of these algorithms, and provide key research gaps for future research in this field (Klaine et al., 2017). The authors refer to several classical ML applications in SONs, including fault detection, fault classification, and cell outage management.

Zhou et al. reviewed research works on the convergence of ML and cognitive wireless networks (Zhou et al., 2018). They discuss current efforts that consolidate ML techniques into cognitive communication systems to improve the spectrum and energy resources in such systems. Moreover, the authors pinpoint future research challenges in this area. Mishra et al. provided a comprehensive survey on ML-driven intrusion detection techniques (Mishra et al., 2018). They performed a careful analysis of different ML algorithms (classical ML and DL algorithms)

that are used for intrusion detection purposes in order to identify the pros and cons of each algorithms. Last but not least, the work conducted in Abbasi et al. (2021b) investigates the capabilities of DL models for four key NTMA applications, including traffic classification, intrusion detection, traffic prediction, and fault management. In Table 1, we compare our survey with the survey works discussed in this section.

3. Basic concepts

In this section, we first provide the fundamental knowledge relevant to our review. As mentioned earlier, our goal is to review the research contributions on DRL-based improvement of QoS at the MAC layer. First, three basic concepts relevant to this survey are presented, including RL, DRL, and QoS.

3.1. RL

During the last decades, automated decision-making has attracted plenty of attention mainly thanks to using Artificial Intelligence (AI) techniques (Allen and Masters, 2020). AI is widely used in a wide range of applications to mimic the human brain in terms of decision-making, but also to improve performance of decisions that cannot be made by a human due to complexity. AI techniques are used in addressing complicated and large-scale decision–action problems, which are out of the power of the human brain. ML is a sub-field of AI to learn decision-making policies based on samples of decisions. Different techniques have been introduced as sub-fields of ML, including: (i) Supervised learning (SL): learning based on labeled data, (ii) Semi-supervised learning (SSL): learning based on a mix of unlabeled and labeled data, (iii) Unsupervised learning (UL): learning based on unlabeled data, and (iv) RL: learning based on interaction with the environment.

Labeled data refers to instances that have been labeled with one or more classes of labels, while unlabeled data does not have any classes or labels. It should be noted that all the aforementioned learning methods are somehow supervised by a cost function. As a distinguished form of ML, RL allows an agent to interact with its environment to achieve the goal without any prior information from the environment in which it operates. The agent takes actions and receives corresponding rewards from the environment while it is not aware of the best actions. Over time, by interacting with the environment and following a trial-and-error search, the agent tries to adopt the possible optimal policy (or a state–action pair) for decision-making. The optimal policy yields maximum long-term or short-term rewards based on the goals of the application.

RL algorithms are categorized based on the factor that whether an agent can lean a model of the environment in which it operates, *e.g.,* a function that forecasts state transitions and rewards. As illustrated in Fig. 2, the RL methods are categorized as follows:

- Model-based RL: in which a model of the world is learned and then using the learned model, the agent predicts the future and makes a plan accordingly. The agent updates and re-plans the model often. Model-based algorithms are sample efficient, meaning that how many samples of data are required by which the agent can successfully operate in the environment and subsequently maximize it rewards. It is due to the fact that the agent does not need to experience all the possibilities. The model is either learned, *e.g.,* world models, Imagination-Augmented Agents (i2A), Model-Based RL with Model-Free Fine-Tuning (MBMF), Model-Based Value Expansion (MBVE) or given (AlphaZero).
- Model-free RL: in which the agent does not try to build or receive a model, rather it constantly updates its knowledge to comprehend that how good an action is to be taken in a give state, *e.g.,* optimal actions. Hence, model-free algorithms do not directly learn a policy, but learn the state or state–action values. In a given state, the main goal for the agent is to take the best action. Therefore, the agent gets experience by exploration rather than exploitation. The algorithms in this category are either policy optimization or Q-Learning.

The model-free methods are mainly categorized into policy-based and value-based. Policy-based or Policy-Gradient (PG) are known as on-policy methods as well. In the policy optimization class, the algorithms optimize the parameters θ either using gradient ascent on the performance object, *e.g.,* cost function, directly, or indirectly, *e.g.,* by maximizing local approximation of the cost function. The main goal is to select a policy gradient that increases probability of good actions (which increases the rewards) and decreases the probability of bad actions (increases the penalties). The main advantage of policy-gradient is the fact that if Q-function is too complex to be learned, it is still able to learn an acceptable policy. It also converges faster and it is capable of learning stochastic policies. Moreover, it is much more easier to model continuous spaces. The main objective of policy-based methods is to maximize a performance measure, *e.g.,* the true value-function of the parametrized policy from all initial states.

Advantage Actor–Critic (A2C) and Asynchronous Advantage Actor–Critic (A3C) are two well-known policy-based algorithms (Wei et al., 2018; Chen et al., 2020b). These algorithms use multiple agents in which each of the agents has its own parameters. A2C combines DQN (value-based) and PG. A2C normally employs two neural networks including actor and critic. The actor is policy-based and its main duty is to sample the action from a policy. While the critic is value-based and is employed to measure how good the chosen action is. A2C synchronizes up for global parameter update and then starts each iteration with the same policy. However, in A3C, agents interact with their respective environments asynchronously. Both A2C and A3C utilize parallelism in the training process. Some other algorithms in this category are proximal policy optimization (PPO) (Wang et al., 2020), trust region policy optimization (TRPO) (Jha et al., 2020), Deep Deterministic Policy Gradient (DDPG) (Xu et al., 2020; Qiu et al., 2019), twin-delayed DDPG (TD3) (Dankwa and Zheng, 2019), and soft actor–critic (SAC) (Haarnoja et al., 2018).

The other subcategory of model-free methods is value-based (Watkins and Dayan, 1992). The ultimate goal of the value-based methods is to find an optimal policy to maximize the long-term reward as well as minimizing the loss function, *e.g.,* the Mean Square Error (MSE) the true Q-function and the parametrized Q-function. In this algorithm, a lookup table is kept, in which there is an entry for each state–action pair, $Q(state, action)$. The algorithm has a reward function to calculate the expected cumulative reward for each state–action pair, and consequently select optimal actions under different situations.

Markov Decision Process (MDP) is used to represent the RL mathematically. MDP is composed of: set of states S, start state s_0, the distribution of the initial states $p(s_0)$, set of actions A, the transition function $P(s_{t+1}|s_t, a_t)$, the instantaneous reward function $R(s_t, a_t, s_{t+1})$, the discount factor $\gamma \in [0, 1]$, the horizon H. Generally, in MDP, We represent a mapping from the states to a probability distribution over actions, $\pi : S \rightarrow (A = a|S)$. In case of episodic MDP, where the state is reset upon accomplishment of each episode, the sequence of S, A, and R in each episode forms a trajectory of the policy and every trajectory of a policy sum up rewards from the environment.

$$R = \sum_{t=0}^{T-1} \gamma^t r_{t+1} \tag{1}$$

where T is the length of an episode. By nature, RL intends to find out an optimal policy, *e.g.,* the policy that gives the maximum expected return, which is represented as follows:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}[R|\pi]. \tag{2}$$

Partially Observable MDP (POMDP) is a generalization of MDP, which is used in the cases that the states are not completely observable. The advantage of the POMDP is that it is general enough to be used to model a wide range of real-world sequential decision-making processes. POMDP is represented using a tuple $\langle S, A, \Omega, Tr, O, R\rangle$, where Ω is a finite set of observations, $O : S \times A \times \Omega \rightarrow [0, 1]$ is an observation

Table 1
Comparison between our survey and the related existing surveys.

Research	Networking technology	ML model				Contribution
		RL	DL	DRL	Other	
Klaine et al. (2017)	Cellular SONs, mobile big data				✓	Insights into applying ML algorithms for SONs.
Luong et al. (2019)	Cellular, CR networks, Ad-hoc, satellite communications, IoT	✓		✓		A comprehensive survey on the applications of DRL for communication systems and networks.
Mao et al. (2018)	Wireless networks, mobile big data		✓		✓	A survey of DL models for wireless networks.
Zhang et al. (2019)	Cellular, CR networks, Ad-hoc, satellite communications, IoT		✓		✓	A comprehensive survey of DL models for mobile and wireless network.
Al-Garadi et al. (2020)	IoT		✓		✓	An overview of ML- and DL-driven IoT security methods.
Mohammadi et al. (2018)	IoT, mobile big data		✓		✓	A survey of DL models for IoT data analytics applications.
Lei et al. (2020)	IoT		✓	✓		A extensive literature review on DRL for autonomous IoT.
Fadlullah et al. (2017)	Cellular, CR networks, IoV, IoT		✓			Insights into employing DL models for network traffic control.
Ahad et al. (2016)	Cellular, CR networks, WSNs, Ad-hoc		✓		✓	An comprehensive overview of NNs for wireless networks.
Alsheikh et al. (2014)	WSNs				✓	An overview of ML applications in wireless sensor networks.
Bkassiny et al. (2012)	CR networks				✓	A comprehensive overview of the application of ML models in cognitive radio.
Zhou et al. (2018)	CR networks	✓	✓		✓	A comprehensive survey of ML for cognitive wireless communications.
Mishra et al. (2018)	Cellular networks, mobile big data		✓		✓	A survey of ML-based intrusion detection techniques.
Abbasi et al. (2021b)	Cellular, IoT, mobile big data		✓			A comprehensive survey of DL techniques for Network Traffic Monitoring and Analysis (NTMA) applications.
Our study	**Cellular, CR network, Ad-hoc, satellite communications, IoT, mobile big data**	✓		✓		**A comprehensive survey of RL/DRL for mobile and wireless networks, with a QoS perspective.**

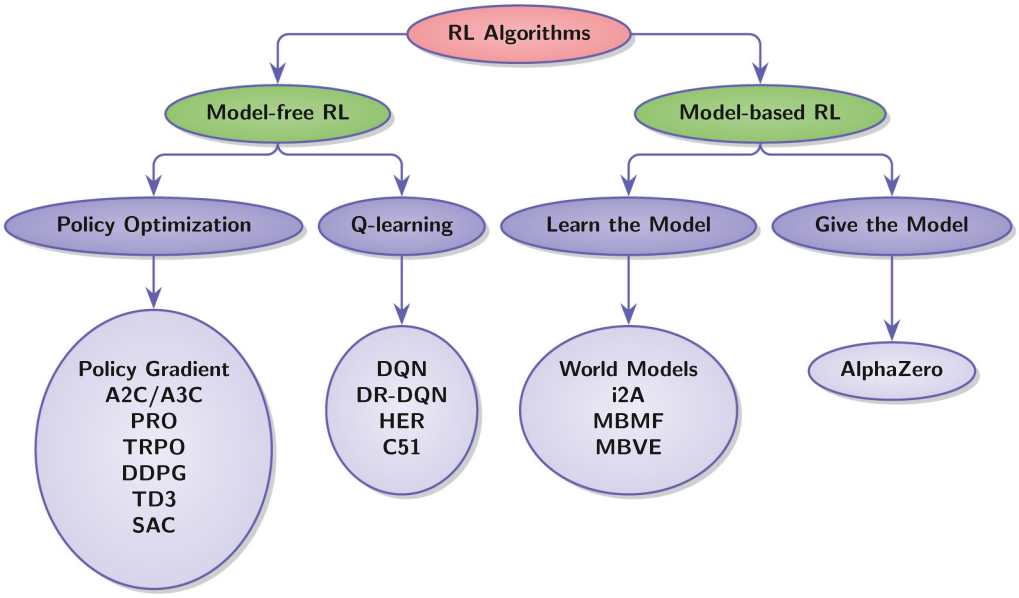


Fig. 2. A taxonomy of the RL methods.

function, $Tr : S \times A \times S \rightarrow [0, 1]$ is a transition function, and $R : S \times A \times S \rightarrow \mathbb{R}$.

Q-learning utilizes any policy to estimate Q in such a way that the future reward is maximized. It requires neither a prior model nor a prior policy. It directly approximates Q^* using the Belman optimality equation and then updates it in every iteration.

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha(R_{t+1} + \gamma max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)) \quad (3)$$

where, $Q_{t+1}(s_t, a_t)$ is the new Q value of the next state, $Q_t(s_t, a_t)$ is the Q value of the previous state, α is the learning rate, R_{t+1} is the reward, and γ is the discount factor. The key challenges in this algorithm are exploration and exploitation.

We can utilize neural networks to approximate the Q-function, which results in a Deep Q-Network (DQN). In DQN, the Q-function, for the value-based methods, uses a neural network. The main goal is as the agent navigates the environment and experiences different actions and results using backpropagation, loss function, gradient, the agent tries to form a good representation of the optimal Q-function. DQN, as a off-policy method, approximates Q and infers optimal policy while policy-gradient directly optimizes policy space. Compared with PG, DQN is sample-efficient, *e.g.,* it needs less data and is more stable during the training process, and strong credit assignment to (state, action) pairs for a delayed reward. Some other examples of algorithms in this category are quantile regression DQN (DR-DQN), categorical 51-Atom DQN (C51), and hindsight experience replay (HER).

In the area of communication systems and networks, RL is an effective solution for dealing with a wide range of resource management issues, scheduling, and other optimization problems. To this end, the problem of interest can be modeled as a RL system, and then the agent, also called learner, must calculate and update the lookup table through the reward function to discover which actions yield the most reward. This calls for updating the reward function in an iterative manner because the agent uses its current knowledge of the rewards of the actions and selects one of these actions (exploiting). However, in the long run, the agent may discover better actions, and therefore the rewards function needs to be updated. In ML terminology, the rewards function updating process is referred as *training*.

3.2. Evolution From RL to DRL

Many modern networks are designed to be decentralized, ad-hoc, and autonomous. Such networks can be modeled as a MDP and solved by RL algorithms. Nevertheless, modern networks, such as IoT networks, are expected to be large in scale and enormously complicated (Shahraki and Haugen, 2018). As a result, RL-based methods become costly in terms of learning time, computational complexity, and maintaining for such large-scale networks.

DRL is a combination of RL and DL, which includes DQN and policy-gradients. DQN techniques intend to forecast the rewards that will follow certain actions taken in a particular state. On the other hand, the policy-gradient techniques either map states directly to actions (deterministic policies) or produce probability distributions for actions (stochastic).

Since DRL has the capability to handle large-scale and dynamic systems, it has been introduced to overcome these issues. In DRL, when the number of states and actions are very large, DL can be used for value function approximations and consequently alleviate the challenges related to memory and computation demands. This new approach is known as DQN. In DQN, the state is considered as the input of the network and the reward value of all possible actions is produced as the output of the network. Fig. 3 shows the difference between Q-learning and DQN. From this figure, one can see that in the DQN the current observed system's states are considered as the input of the network. The inputs go through different network layers with particular weight to extract high-level features from the inputs. In the final phase, the DQN considers all of the possible actions and produces a set of Q-values as its outputs.

DQN can be adopted to speed up the learning process and also reduce memory usage for storing the model parameters (Luong et al., 2019). This is especially important for electronic devices with limited resources, such as IoT devices. Therefore, one can conclude that DRL is an ideal solution to deal with many challenges in communication systems and networks, such as resource management, non-convex and complex problems, scheduling, etc. Moreover, DRL can enormously decline the complexity of the model, in particular when it is used as a solution for a wide range of challenges in upcoming communication systems and networks. More specifically, the Q-learning algorithm's performance will drop-off significantly when one applies this algorithm in complex and sophisticated environments, such as communication systems (*e.g.,* due to the updates in the Q-table). As a result, DRL can be used as a function approximation for the Q-function to deal with the high dimensional state-space environments. DRL can be applied to a wide range of communication models and fields including IoT, HetNets, Unmanned Aerial Vehicle (UAV), ad-hoc, etc. (Luong et al., 2019). Generally, the key advantages of DRL over RL are as follows:

- In RL, the agent employs trail-and-error techniques to conclude the optimal policy and it can receive the reward as the only feedback.
- In RL, The agent must deal with long-range time dependencies, where it faces the issue of credit assignment (Arulkumaran et al., 2017).
- DRL is able to manage the cases in which the inputs are constantly changing (Zhang et al., 2020a).
- DRL can reduce the computational complexity and maintain costs in large-scale problems.
- DRL algorithms can learn directly from visual inputs, such as images and videos (Levine et al., 2016, 2018). This feature enables DRL to be a promising tool for many real world problems, such as robotics, playing video games, and indoor navigation (Mnih et al., 2015; Zhu et al., 2017). It can also be used to analyze the time-series gathered from networks (Shahraki et al., 2017b; Shahraki and Haugen, 2019).

Due to its unique advantages, DQN has received much attention over the recent years and hence several new improvements on DQN have been introduced including Rainbow DQN (Hessel et al., 2018; Canaan et al., 2020), SAC (Haarnoja et al., 2018), DDPG (Wu et al., 2020), PPO (Schulman et al., 2017), etc.

The main intention of the Rainbow DQN is to generate a comprehensive DL training infrastructure by integrating a bunch of DQN improved versions including:

- Double DQN (DDQN) to consider an overestimation bias of DQN by decoupling the selection and evaluation of the bootstrap action (Tao and Hafid, 2020).
- Prioritized Experience Replay (PER) to enhance data efficiency by replaying more often transitions and hence improve the training procedure (Hou et al., 2017).
- Multi-step DQN to improve the learning process by utilizing multi-step boot-strap targets (Hernandez-Garcia and Sutton, 2019).
- Distributional DQN to approximate the distributions of returns rather than the expected returns (Bellemare et al., 2017).
- Dueling DQN algorithm includes two streams of computation, including value and advantage streams, sharing a convolutional encoder, and emerged by a particular aggregator (Wang et al., 2016).
- Noisy DQN to improve the exploration process by utilizing the stochastic network layers (Fortunato et al., 2018).

3.3. QoS

QoS refers to a set of mechanisms to control bandwidth, delay, jitter, and packet loss, to name a number of parameters (Shahraki et al., 2017a). QoS mechanisms aim at improving at least one of these parameters. Intrinsically, QoS mechanisms have no direct influence on the amount of actual resources of any given network, but they are used to improve the performance of using existing resources by efficiently assigning them to tasks. That is to say that, these mechanisms enable the network management agents to more efficiently allocate the existing network resources, such as bandwidth.

Considering the ever-increasing number of connected smart devices and the volume of generated traffic, QoS becomes an essential means to deal with when the services are increased in number and variety. Internet Service Providers (ISPs) are selling QoS in explicit (*i.e.,* as an option) or implicit (*i.e.,* embedded into services) ways (Xiao, 2008). ISPs guarantee a certain level of performance for many services, such as voice, data, and multimedia services. To this end, they have to take into account multiple QoS-related factors, *e.g.,* end-to-end latency (Rafsanjani et al., 2014), loss rate, and reliability.

Cellular-based communications play a key role in our daily life, *e.g.,* business, industry, and personal affairs. Hence, QoS has to be given a high priority. In the context of cellular networks, one can define QoS as the ability of the telecom providers companies to offer an appropriate service in terms of quality of voice, signal strength, low call dropping probability (CDP), and high data rates for multimedia or data services. QoS in cellular networks depends on five main factors, including bandwidth (or throughput), delay, packet loss rate, packet error rate, and reliability (Cheng et al., 2015). It is crucially essential to differentiate between traffic with different priority levels. Some types of traffic have a higher priority than the other ones, *e.g.,* voice service should have a higher priority than data service due to the fact that users consider voice as the most essential service. In the context of QoS, the users who pay more for better services need to receive more preference, without negative influence on the remaining users with normal pay rates. To achieve all these things, precise and effective QoS mechanisms are necessary.

Broadly speaking, one may classify the QoS mechanisms based on how the service traffic is treated into two major groups: (1) resource management, and (2) traffic handling. In the MAC layer, the

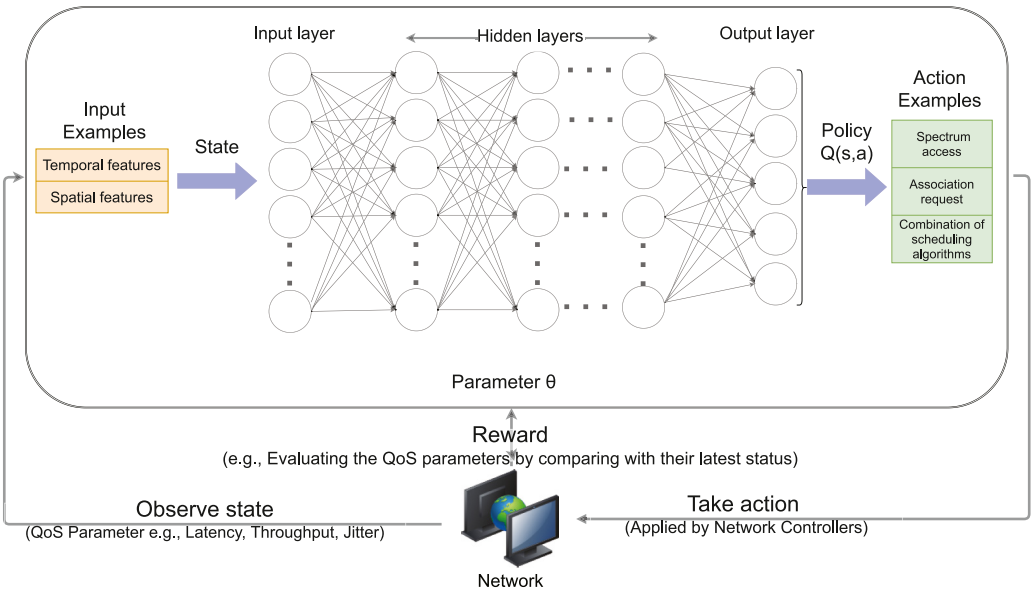


Fig. 3. An example model of DRL for QoS-provisioning.

former refers to the mechanisms and techniques that aim at managing the network resources (Shahraki et al., 2020b), *e.g.,* radio resource, through configuration and coordination of network elements like BS, User Equipment (UE), and Access Point (AP). One can refer to admission control and resource reservation as two main mechanisms for bandwidth management. The latter, *i.e.,* traffic handling refers to mechanisms that optimize traffic transmission in a network. The major traffic handling mechanisms are categorized into classification, traffic policing, channel access, and packet scheduling.

In the next section, we review several recent works that used DRL to support QoS. As mentioned, we review the main DRL-based contributions supporting QoS in the MAC layer, which are classified into two main categories, including network access and data rate control, and resource sharing and scheduling (see Fig. 1).

It is worth noting that in this study, we do not focus on specific network infrastructures, but our aim is to conduct the literature review based on a problem-solution approach in DRL-based QoS support for the MAC layer.

4. Network access and data rate control

In this section, we review the contributions of deep RL towards dealing with three issues, *i.e.,* dynamic spectrum access, joint user association and adaptive data rate control. Note that throughout this paper we follow a problem-solution strategy for reviewing the related work. Additionally, our main aim is to cover different types of networks, such as cellular, V2V, sensor networks, etc. This allows us to structure our paper according to the theme of the reviewed papers.

4.1. Dynamic spectrum access

Dynamic Spectrum Access (DSA) includes a group of mechanisms to enhance the efficiency of spectrum through the real-time adoption of radio resources. DSA is important in supporting QoS (Ali et al., 2019). For example, an inefficient DSA mechanism can waste network resources by unfair spectrum resources allocation to users (Tian et al., 2013). Generally, spectrum access mechanisms can be categorized into two main classes, including Listen-Before-Talk (LBT) and the spectrum sharing scheme. The former refers to the strategy, in which Secondary Users (SUs), also known as unlicensed users, only send over a frequency band when it is sensed to be idle. By adopting the LBT scheme, it is possible to effectively avoid strong interference to primary users (or licensed users). However, given the random nature of wireless communication systems, the lack of/no cooperation among SUs, and other considerations, this scheme is not appropriate for achieving spectrum co-existence between network users. More specifically, miss detection of primary users may lead to poor decisions by SUs in terms of channel access (Fu et al., 2017). Another spectrum access mechanism is spectrum sharing (Stotas and Nallanathan, 2011). Spectrum sharing refers to the simultaneous utilization of a specific radio frequency band in a specific geophysical zone by primary and secondary users, leveraged by techniques such as adjusting the transmit power level. This scheme is based on the fundamental assumption that a priori knowledge on CSI of the link between a SU transmitter and a receiver primary user is available for power adjustment. Considering this fact that modern networks become more decentralized in nature, it is an unrealistic assumption that a central controller is necessary for obtaining these CSI. Moreover, with a central controller, exchanging these CIS is costly in terms of control overhead. To address these challenges, DRL techniques have recently been proposed in the context of DSA because of their ability to adapt to decentralized, dynamic and unknown environments.

Dynamic spectrum access mechanisms based on DRL can not only use the current spectrum sensing result but also leverage the outcome from past spectrum conditions. By adopting such mechanisms, one can reduce the negative impact of poor spectrum sensing caused by users. Moreover, using DRL-based spectrum access mechanisms, it is possible to acquire useful information about the channel status, *e.g.,* the pattern behavior of primary users and the load of other SUs. This information is especially important for collision reduction among primary and secondary users (Chang et al., 2018).

Motivated by the above-mentioned advantages, the work in Chu et al. (2018) proposes an uplink scheduling algorithm, based on RL and DQN for IoT networks. The proposed algorithm is applied to a small-cell IoT network with multiple energy harvesting UEs, one BS and limited uplink access channels. MDP is used to model the network control, with the assumption that complete prior knowledge is not available at the BS. Moreover, a RL based algorithm is introduced for the battery prediction problem. The proposed algorithms show better results compared to existing benchmarks. DRL can be a promising technique in energy harvesting IoT networks as in such networks it is very difficult to acquire a priori knowledge about dynamics of energy sources. In Zhu et al. (2018), DRL was used as a method for transmission scheduling in Cognitive IoT (CIoT) networks. In their work, Zhu et al. use Q-learning to develop an effective strategy for transmitting packets of

different buffers by multiple channels in order to maximize the network throughput. They deployed MDP to model the state transformation of the network. In their method, Q-learning helps a relay achieve the optimal strategy, and consequently maximizes the network throughput. In He et al. (2018), He et al. focus on green resource allocation in content-centric-based IoT. More specifically, they introduce DQN-based QoE-driven download resource allocation algorithms to manage the network through experience, and improve QoE. Similar to the above-mentioned researches, in this work MDP is also used to model the IoT network. The model includes states, actions and rewards; and the agent grasps the optimal decision and control mechanism in an online manner.

Vehicle to Vehicle (V2V) is another communication system that derived substantial benefit from DRL. V2V communications based on Device to Device (D2D) communications are considered as an important technology in 5G mobile networks to improve intelligent transportation and road safety. This is mainly due to the fact that D2D communications are able to provide satisfactory performance in terms of V2V QoS requirements. However, effective resource allocation techniques is essential to manage the interference between D2D and the cellular users. This is becoming more challenging as in V2V networks the nodes (vehicles) are highly mobile, and hence changes in wireless channels are rapid. Consequently, the traditional spectrum access mechanisms with a complete CSI assumption can no longer be used in modern V2V networks because it can be almost impossible to track channel changes in such networks with high mobility. Centralized mechanisms introduced for spectrum allocation in D2D-based V2V networks will impose a large transmission overhead to the networks since each node needs to send its local CSI and interference information to the central controller. In this case, the DRL scheme can meet the aforementioned challenges by considering a V2V link as an agent and training the agent to learn from the environment in order to make optimal decisions on spectrum and power for transmission (Ye et al., 2019). The agent leverages its experience to maximize the reward, which in this case is a function of the capacity of the V2V and Vehicle to Infrastructure (V2I) links and the corresponding latency. In Liang et al. (2019), Liang et al. look at the spectrum sharing issue in communication systems through a multi-agent RL algorithm. The authors model the issue as a multi-agent RL problem. Then, they use a fingerprint-based DQN to solve the problem. Analysis of the proposed method reveals that by adopting a right reward design and training mechanism, it is possible to train the V2V transmitters to learn from the environmental interactions and discover an optimal strategy for cooperative work in order to maximize system level performance through leveraging local information.

DRL has attracted much attention in recent years in enabling dynamic spectrum access, especially in wireless networks. According to Knopp and Humblet (1995), dynamic spectrum access is important for the efficient spectrum utilization in wireless networks to deal with the increasing demand for more capacity. This is especially important when a wireless network operates with other networks in the same spectrum. There is a large body of research that targets spectrum access in wireless networks through design and implementation of algorithms. However, the majority of them are based on a simple independent-channel model, which may be not true in real world deployments. For example, consider a Wireless Sensor Network (WSN) operates on IEEE 802.15.4 standard. This standard uses three possible frequency bands, *i.e.,* 868/915/2450 MHz, that are shared by different technologies and applications such as Bluetooth, Wi-Fi and industrial/scientific equipment. As a result, external interference in Low-rate Wireless Personal Area Networks (LR-WPANs) with a focus on enabling WSNs is highly possible, thereby adopting novel mechanisms for dynamic multichannel access, and consequently providing QoS, is necessary. Motivated by this challenge, the work in Wang et al. (2018) deploys DQN to deal with dynamic multi-channel access challenge. This DRL algorithm enables an agent to learn in unknown environments and decline the computational costs. The authors in Wang et al. (2018) implement a mechanism based on DQN which is able to shape a channel access policy through online learning. The mechanism can handle large-scale communication systems, and shape an appropriate policy or even optimal policy by leveraging the past experience, and without the need to prior knowledge about the dynamics of system. In Naparstek and Cohen (2017), Naparstek et al. investigate the dynamic spectrum access problem in multichannel wireless networks. The authors introduce a distributed dynamic spectrum access algorithm that gains advantages from multi-user DRL. That is to say that, at each time slot, a constructed DQN is employed by each user to map its present state to actions (*i.e.,* spectrum access) in order to maximize the expected Q-value. The proposed algorithm is more effective than random access protocol since the user can figure out proper policies to improve the performance of the network, without having to perform online coordination, communications between users, or carrier sensing. In Yu et al. (2019), a DRL-based MAC protocol is proposed for heterogeneous wireless networks. The main goal of the proposed protocol is to learn an optimal strategy for channel access to meet a pre-defined global objective, *e.g.,* sum throughput maximization. By investigation of this MAC protocol, one can see that with precise definitions of the state space, action space, and rewards in DRL, the protocol is able to maximize the sum throughput through judiciously selecting certain time slots for transmission. One must note that the benefits of DRL-based MAC protocol in wireless networks is twofold; first, the ability to converge fast for near-optimal solutions and second, robustness against sub-optimal parameter values.

Cellular networks have to meet many challenges, such as co-existence with other technologies, multiple access, and resource allocation, to provide the promised QoS. Fortunately, DRL is a practical approach to tackle such complex challenges. For example, in Zhao et al. (2019), Zhao et al. target user association and resource allocation problems in HetNets. To this end, they determine the state, action and reward functions for UEs and propose dueling a double deep Q-network (D3QN) to figure out the near-optimal strategy. The proposed method enables UEs to achieve the global state space with a few message transmissions, *i.e.,* with a small communication overhead. Moreover, D3QN can guarantees the QoS requirements of UEs in the downlink of cellular HetNets. Zhong et al. (2019) proposed a framework for dynamic multichannel access based on deep actor–critic RL. The authors argued that the multichannel problem can be modeled as a partially observable Markov decision process (POMDP) because of this fact that each user can observe only the states of channels selected by him/herself. Then, they propose two actor–critic DRL frameworks to find the optimal policies in single-user and multi-user scenarios. The authors report that the proposed frameworks have great abilities to handle a huge number of channels, high tolerance against uncertainty, and large action spaces. Feng and Mao in Feng and Mao (2019) explored two key problems in millimeter-wave systems, *i.e.,* limited backhaul capacity and extremely dynamic data rates of users. The limited backhaul capacity is mainly due to the mmWave channels random blockage. To deal with these problems, the authors introduce a DRL method. More specifically, the blockage pattern in mmWave channels can be learned using the introduced method, and then the learned patterns can be utilized to predict the system dynamics, which results in efficient utilization of backhaul resources. In this case, DRL-based methods have considerable advantages over traditional methods. For example, they do not need explicit/instantaneous knowledge on the network or explicit information about the inter-dependent patterns of different nodes. Moreover, deep the RL-based method can alleviate the online overhead, resulting in better performance.

According to Aravanis et al. (2015), Dynamic Resource Allocation (DRA) can be used in resource-limited Multi-Beam Satellite (MBS) systems to improve the network performance. Motivated by this fact that the existing DRA approaches, *e.g.,* iterative metaheuristics optimization algorithms, are highly complex in terms of computation, the authors in Hu et al. (2018) proposed a DRL-based framework for DRA in MBS systems. They mapped the DRA problem in MBS systems into

a sequence of DRA decisions, and then modeled the problem as the interactions between satellite and user terminals services. The authors also report the maximum cumulative rewards that the agent (*i.e.,* satellite) is able to gain from the environment (user terminals services) which is the optimal solution to the problem. The main contribution of the paper is that the state is considered as a series of images to fully represent the spatial and temporal features. The work of Liu et al. (2018) also proposed a DRA algorithm in MBS systems based on DRL. In this paper, the DRA problem is modeled as a MDP. Similar to what the authors have done in Hu et al. (2018), the system state is reformulated into an image-like tensor and informative features are extracted through Convolutional Neural Network (CNN). The proposed algorithm shows better results in terms of the blocking probability and spectrum efficiency compared to existing algorithms.

Kassab et al. investigate the challenge of dynamic spectrum access through multi-agent deep deterministic policy gradient (MADDPG) (Kassab et al., 2020). They assume an IoT network with N devices that compete for T time slots, which constitute a frame. All the devices monitor K events where various IoT devices could monitor each event. When at least one of its monitored events becomes active, each IoT device selects an event and a time slot to send the related active event data. In such a network, to avoid collisions and redundant transmission events, the authors proposed an event-based spectrum access scheme by employing DRL. Through the numerical simulations, the authors demonstrated the superiority of their scheme over benchmark schemes, such as ALOHA and independent DQN. A comparison of the reviewed papers in this section is provided in Table 2.

4.2. Joint user association

Regarding the ever-increasing in the number of mobile devices and deployment of IoT solutions, dealing with the rapid growth of wireless applications/services will be of paramount importance for the 5th generation mobile network (Huang et al., 2017). Fortunately, network heterogeneity, especially by deploying small BSs, *e.g.,* pico BSs (PBSs) and femto BSs (FBSs), is a promising solution to tackle the challenge. These BSs are different in terms of the transmitter power, the size of the coverage area, physical size, cost, etc. HetNets are more flexible in terms of deployment than traditional macro BSs, which are pretty costly to deploy and maintain. Additionally, in order to enhance the spectrum efficiency of the HetNets, small BSs (*i.e.,* PBSs and FBSs) are able to reuse and share the same channels with macro BSs. Hence, HetNets have been considered as an effective solution to increase the capacity of modern communication systems and networks.

In order to benefit from the full functionality of the small BSs, network users should be assigned to these types of BSs. This is mainly due to the fact that the small BSs have often light load, and consequently can provide a higher data rate. Assigning the users to the small BSs, also known as user association, is an optimization problem in HetNets. While the user association problem is investigated in the literature (Fooladivanda and Rosenberg, 2012; Lin et al., 2015), regarding the combinatorial and non-convex nature of the problem, it is difficult to find an optimal and nearoptimal solution. The previously proposed methods for the user association problem, such as game-theoretic, linear programming, and the Markov chain approximation technique seem to be inefficient since they require almost complete and accurate information on network to obtain a globally optimal strategy, which may not be typically accessible. Hence, scholars investigate deep RL-based methods to deal with such a challenging problem. For example, in Yi et al. (2018), Yi et al. leverage multi-agent deep RL to develop an adaptive user association approach. In their work, the authors consider different UE types and FBS access mechanisms to create a trade-off between QoE and load balancing. In comparison with the conventional approaches, such as distributed multiple attribute decision making (MADM), the proposed approach shows better performance in terms of QoE, load balancing and blocking probability. In Ding et al. (2020), the user association and power control problems in HetNets are explored. They propose a multi-agent deep DQN method to tackle the problem. As the authors describe, the multi-agent DQN method needs less network information, unlike the widely used methods, such as convex optimization and fractional programming, that require more and accurate network information. Furthermore, the proposed method is able to ensure the requirements related to the user's QoS and network utility. The work in Zhang et al. (2019) by Zhang et al. focuses on the user association problem in symbiotic radio networks (SRN), in which a primary network supports not only the primary communications but also the IoT communications. The authors' goal is to link each IoT device to a fitted UE to provide the maximum sum rate for all IoT devices. Similar to what we mentioned in Ding et al. (2020), complete real-time CSI needs to obtain an optimal policy in this case. To deal with this challenge, the authors propose two deep RL algorithms that utilize the past information in order to deduce the current information to take appropriate decision. In Sana et al. (2019), the authors investigate the user association problem in dense mmWave networks. They propose a multi-agent deep RL-based distributed algorithm to deal with this problem. To be specific, they consider each user as an agent, which maps its local observations to an action, *i.e.,* an association request to a BS in its coverage area in order to maximize the network sum-rate. The proposed algorithm offers considerable advantages. For example, it operates in a fully distributed fashion and without the need for information exchange between UEs. In Khan et al. (2019), the authors explore the issue of vehicle-cell association in mmWave networks. To this end, the authors carefully model the vehicle association as a discrete non-convex optimization problem. Then, they utilize distributed deep RL (DDRL) and the asynchronous actor–critic algorithm (A3C) to introduce an algorithm for solving the problem. The simulation results reveal the superiority of the proposed method over several baseline methods. Note that in such networks due to the mobility and the use of mmWave, traditional optimization techniques pose huge computational overhead, whereas deep RL-based techniques can lead to low computational and signaling overhead. Another interesting study that investigates the joint issue of user association and resource allocation is the one carried out by Zhao et al. (2018). In this study, the authors propose a novel technique based on double DQN (DDQN) in order to maximize the long-term overall network utility, as well as meet the QoS requirements of UEs. The choice of deep RL is mainly due to the fact that the joint of user association and resource allocation is a non-convex and combinatorial problem, and hence DDQN is a promising solution for obtaining the optimal policy with a little communication overhead.

Chou et al. address the joint user association and resource management problem in mobile edge computing (MEC) by means of DRL (Chou et al., 2020). To be specific, online video streaming is one of the services that can benefit from MEC. Nonetheless, user association and video quality adoption are still challenging in this scenario. To deal with these challenges, the authors model it as an MDP and use the DDPG algorithm to solve it. Analysis of the simulation results reveals that the proposed method delivers notable QoE improvement.

A comparison of the reviewed papers in this section is provided in Table 3.

4.3. Adaptive data rate control

The continuing growth in the number of smart mobile devices, and user friendly and innovative high data-rate services, such as video game streaming and condition monitoring, calls for investigation solutions to accommodate the enormous wireless traffic demands (Zheng et al., 2015). According to the Cisco's global mobile data traffic forecast (Index, 2017), 75% of the world's mobile data traffic will be video by 2020. Demand for mobile video content and video usage on the mobile network by mobile users is endless, hence the user's Quality of Experience (QoE) is a key indicator for telecommunication provider companies.

Table 2
DRL for DSA.

Research	Year	Network	Model	Learning algorithm	Improved QoS factor(s)	Technical contribution
Naparstek and Cohen (2017)	2017	Wireless networks	Game	DQN+LSTM	Data rate Packet loss	Develops a distributed dynamic spectrum access technique based on deep multi-user RL for wireless networks.
Chu et al. (2018)	2018	IoT	MDP	DQN+LSTM	Data rate Reliability	Investigates both the access control and battery prediction issues through a two layer LSTM network with DQN enhancement.
Zhu et al. (2018)	2018	IoT	MDP	Q-learning+SEA	Throughput Packet Loss	Proposes a novel deep RL-based transmission scheduling technique to discover the proper strategy in the presence multiple channels for the cognitive node.
He et al. (2018)	2018	IoT	MDP	DQN	QoE	Proposes a model for evaluation the quality of the whole IoT and resource allocation algorithms in CIoT.
Wang et al. (2018)	2018	WSNs	POMDP	DQN	Throughput Packet Error Rate	Applies DQN to dynamic multichannel access problem to increase the number of succeed transmissions. Moreover, compares the proposed method with well-known methods such as Myopic policy and Whittle Index-based heuristic.
Hu et al. (2018)	2018	Satellite communications	MDP	DQN+CNN	Bandwidth, Round trip delay	Provides a deep RL-based framework for dynamic resource allocation in MBS systems.
Liu et al. (2018)	2018	Satellite communications	MDP	DQN+CNN	Throughput Delay	Introduces a dynamic channel allocation algorithm based on deep RL, which considers temporal correlation among the sequential channel allocation events.
Ye et al. (2019)	2019	Vehicular networks	MDP	DQN+LSTM	Delay Throughput	Introduces a decentralized resource allocation technique through deep RL to deal with the latency constraints on V2V links.
Liang et al. (2019)	2019	Vehicular networks	MDP	DQN + three layers fully connected network	Throughput Reliability	Introduces a distributed spectrum access mechanism in vehicular networks based on multi-agent RL.
Yu et al. (2019)	2019	HetNets	MDP	DQN+ResNet	Throughput Proportional Fairness	Establishes a framework to utilize deep RL in the construction of a MAC protocol for HetNets. Also, investigates the benefits of deep RL over conventional RL for heterogeneous wireless networks.
Zhao et al. (2019)	2019	Cellular networks	MDP	DDQN+ dueling DQN	Throughput Proportional Fairness System capacity Network Utility	Introduces a distributed multi-agent deep RL technique for the jointly user association and resource allocation in heterogeneous cellular networks.
Zhong et al. (2019)	2019	HetNets	POMDP	DQN+ neural network with two layers	Delay System capacity	Provides an actor–critic deep RL framework for dynamic multichannel access in a single-user scenario and demonstrates that the framework can handle a relatively larger number of channels.
Feng and Mao (2019)	2019	Cellular networks	MDP	DQN+LSTM	Data rate backhaul Capacity	Proposes a deep RL-based method to deal with the problem of limited backhaul capacity in mmWave networks.
Kassab et al. (2020)	2020	IoT	Dec-POMDP	MADDPG	Throughput	Investigates the application of multi-agent RL algorithm in event-based dynamic spectrum access.

Dynamic Adaptive Streaming over HTTP (DASH) has been considered as the predominant mechanism for video transmission (IEC 23009-1, 2014). This is mainly due to the fact that it is able to use the existing HTTP server and Content Delivery network (CDN) infrastructures, as well as it is suitable for many of client-side applications. By adopting the DASH technique, the video content breaks into short segments, where each segment contains a few seconds of the content. These segments are encoded at different bit rates in order to create an adaptation set of representations. When a client plays back the video content, it deploys a Adaptive bitrate (ABR) algorithm to automatically choose a representation, *i.e.,* the segment with the highest possible bit rate that can be downloaded. Indeed, the client attempts to maximize the QoE for the current video content, based on current network conditions. A large number of commercial systems use basic heuristic methods as the DASH adaptation approach. This can cause annoying quality variations (Hoßfeld et al., 2014) and a poor network resources utilization. Towards optimization of the client's QoE, the adaptation approach should consider both the video content and the playout buffer state, *i.e.,* the time the freezed video plays out. In such a complex case, DRL has been considered as an effective and feasible remedy to the video adaptation problem. The video adaptation mechanisms based on DRL can learn from the past experience through trial-and-error, and optimal policy can gradually be derived. Based on our previous knowledge, it is easy to understand that one can model this problem as an MDP, in which the agent is the client and the action is the selection of a representation of the video segments to download. Applying DRL techniques in this challenging scenario can bring some advantages, such as, (1) the ability of DRL to handle the systems with a very large state space in an efficient manner, and (2) achieving much better QoE, *i.e.,* visual quality of each video segment and the stability variations across video segments.

The overwhelming advantages of DRL-based approaches motivate researchers to investigate the applications of DRL in adaptive rate control problems. For example, Zhang et al. in Zhang et al. (2018) applied self-transfer DRL to the cache-enabled video rate allocation problem. They provided a mathematical model for the problem and introduced a DRL method for solving it. More specifically, the cache-enabled video rate allocation problem has been modeled as a MDP and the Q-learning scheme has been leveraged to figure our the optimal allocation policy. The simulation results verified the performance of the proposed method in terms of user's QoE, while the user moves among small cells. The work presented in Gadaleta et al. (2017) by Gadaleta

Table 3
DRL for joint user association.

Research	Year	Network	Model	Learning algorithm	Improved QoS factors	Technical contribution
Yi et al. (2018)	2018	HetNets	MDP	DQN	Transmission Rate QoE	Designs and develops a multi-agent DQN algorithm for joint user association and power control problems in HetNets.
Zhao et al. (2018)	2018	HetNets	MDP	DDQN	System Capacity Network utility QoE	Provides a distributed deep RL framework to discover the optimal strategy for the joint user association and resource allocation optimization problem in the HetNets.
Zhang et al. (2019)	2019	SRN	Gauss Markov process	DQN	Transmission Rate	Investigates the user association problem in SRN and introduces two algorithms based on deep RL, without the need of the full real-time CSI.
Sana et al. (2019)	2019	Cellular networks (MmWave)	Multi-agent MDP	Hysteretic deep recurrent Q-network (HDRQN)	Throughput	Introduces a flexible method for the user association problem through deep RL in dense networks with different radio access technologies
Khan et al. (2019)	2019	Cellular networks (Vehicular)	MDP	A3C+ deep NN	QoE, Network-Wide Sum Rate	Proposes a low complexity deep RL-based algorithm for vehicle association problem in mmWave communication networks.
Ding et al. (2020)	2020	HetNets	MDP	DQN+ fully-connected neural network with two hidden layers	Data Rate, QoE	Introduces a semi-distributed multi-agent deep RL-based framework to obtain the optimal strategy for user association problem.
Chou et al. (2020)	2020	Cellular networks	MDP	DDPG	QoE	Provides a DRL-based method for improving online video streaming services in MEC by addressing joint user association/resource management problem.

et al. combined DL and RL to establish the framework, called D-DASH, for optimization of the DASH's QoE. Accordingly, the D-DASH framework adopted different DL structures, *e.g.,* combination of feed-forward and RNN, to approximate the Q-values. The authors asserted that their framework based on DRL achieves better performance as there is a better trade-off between policy optimality and the speed of convergence compared to the advanced Q-learning DASH adaptation methods. Similarly, the authors in Mao et al. (2017) also focused on adaptive video streaming in the DASH system. They proposed a system based on RL, namely Pensieve, that generates ABR algorithms. Pensieve uses the past decisions results to make decisions, unlike the conventional schemes that depend on assumptions about the environment or pre-programmed models. Pensieve adopts A3C as the training algorithm, which simultaneously trains two NNs, *i.e.,* the actor network and the critic network. The actor network involves bit rates decisions for the client, where the critic network merely is responsible for training the actor network. The simulation results reveal that Pensieve outperforms the existing ABR algorithm in terms of the average QoE.

Making a trade-off between sending the bit rate and video quality is the main driver of the work by Huang et al. in Huang et al. (2018). For this reason, they designed and implemented a rate control algorithm, called QARC, to get higher video quality with possibly sending at a lower sending rate. QARC leverages DRL for training a NN in order to choose the future bit rates for video segments based on previous observations of the network status and the past video segments. The aim of this work is to minimize a fundamental weakness of a set of conventional rate control approaches (*e.g.,* the loss-based approach and the delay-based approach), in which effort is made to choose the bit rate as high as possible. Nevertheless, because of the occurrence of imbalance between high video quality and high bit rate, these approaches may lead to waste of bandwidth resources. In Kan et al. (2019), an adaptive bit rate control algorithm for 360-degree video streaming based on DRL was proposed. The main goal of the algorithm is to maximize the QoE of clients. To achieve this goal, the algorithm chooses an optimal bit rate for each video segment, before downloading the video chunks by client. The rate adaptation problem is modeled as a MDP and A3C algorithm is used for solving this problem. One can refer to video playback quality, bandwidth efficiency, and alleviation of the stalling video problem as the main advantages of the proposed algorithm. Last but not least, in a study that is conducted by Mao et al. (2019), the authors investigated the adaptive bit rate problem in a real-world application. They accurately assess Facebook's ABR methods for its web-based video streaming platform and provide the deployment experience of a RL-based ABR approach in this platform. The authors refer to the fact that the deployment of control policies based on RL in real-world applications needs custom-built designs beyond RL algorithms. As a result, they propose to design new components in the adaptive BRL's learning procedure to deal with some exciting challenges. The major challenges that they refer includes: (1) RL needs a single reward value as the training feedback, however, in many ABR algorithms there are multiple optimization objectives together, such as maximization bit rate and minimization stalls, (2) videos have different bit rate encodings (*e.g.,* HD and SD). Nevertheless, standard RL techniques employ NNs that produce fixed outputs, both in the number of bit rates and the corresponding bit rate levels. The interested reader may refer to Mao et al. (2019) for a detailed description of the challenges.

Fu et al. explored the issue of live streaming services in IoV networks (Fu et al., 2020). More specifically, the authors stated this fact that live streaming services face serious challenges to offer a service with high quality, low bit rate variation, and low latency for IoV applications. This is mainly because of the highly dynamic nature of IoV channels. To deal with this challenge, they used the SAC algorithm to design a new live video transcoding and streaming approach. The proposed approach can increase the bit rate and reduce delay/bit rate variations by optimizing resource allocation (spectrum and computational), bit rate selection, and vehicle scheduling in IoV networks. MDP has been used to model the joint optimization problem, and then the A3C algorithm has been employed to solve the modeled problem. Similar work was conducted in Jiang et al. (2020), in which the A3C algorithm was adopted for solving video offloading and resource allocation problems, and consequently improve the transaction throughput. The authors claimed that the proposed algorithm is designed in a decentralized manner with low complexity. Based on the simulation results, this algorithm provided better performance than its counterparts in terms of convergence and throughput. A summary of the reviewed papers in this section is provided in Table 4.

Table 4
DRL for adaptive data rate control.

Research	Year	Network	Model	Learning algorithm	Improved QoS factors	Technical contribution
Gadaleta et al. (2017)	2017	DASH system	MDP	DQN+ LSTM	QoE bit rate	Surveys existing RL-based approaches for video adaptation and highlights the limitations of these approaches in terms of QoE, time/memory requirements. Also, establishes a deep RL-based framework for improve the QoE in DASH system.
Mao et al. (2017)	2017	DASH system	MDP	DQN+ A3C	QoE, bit Rate	Provides a system that uses the past observations to generate ABR algorithms, and consequently optimize its policy for choosing a bit rate for each video segment.
Zhang et al. (2018)	2018	Wireless networks	MDP	Deep continuous Q-learning	Network Capacity, bit Rate	Models the allocation problem as a MDP and utilizes deep RL techniques to solve it, and consequently achieves the extended network capacity and bit rate.
Huang et al. (2018)	2018	DASH system	MDP	DQN+ CNN+RNN	QoE, bit Rate, Delay	The first work that proposes a deep RL technique to choose sending bit rate for future video segments by considering previous observations of network status and the past video segments.
Kan et al. (2019)	2019	Field of view (FoV) applications	MDP	DQN+ A3C	QoE, Bandwidth Capacity,Delay	Introduces a deep RL-based rate adaptation algorithm based on the A3C algorithm for the adaptive 360-degree video streaming.
Fu et al. (2020)	2020	IoV	MDP	A3C	Bit rate variation, Delay	A DRL scheme based on the A3C algorithm for trans-coding and streaming in IoV networks.
Jiang et al. (2020)	2020	Autonomous IoV	MDP	A3C	Throughput	Used the MDP to model video offloading and resource allocation problems in autonomous IoV and then adopt the A3C algorithm to solve them.

5. Resource sharing and scheduling

Considering wireless networks, resource sharing/scheduling mechanisms have a key role in fulfilling QoS requirements, such as bandwidth, delay, jitter, and packet loss (Liu et al., 2001). Due to the exclusive features of wireless channels (*e.g.,* dynamic channel conditions, and restricted bandwidth and radio range), the traditional resource allocation and scheduling mechanisms, such as those deployed in the wired networks, can no longer be used. The conditions of the wireless channels are varying in terms of path-loss, fading, interference, and slow log-normal shadowing. Hence, network users may receive time-varying service quality. For example, in a cellular voice service, mobile users who are in good channel conditions may receive better voice quality.

In the MAC layer, the scheduling algorithms are crucially important to the careful management of radio resources because they can help with saving resources. Resource scheduling refers to the assignment of physical resource blocks, *e.g.,* spectrum access by users and prioritizing users using existing approaches to comply with some QoS requirements such as delay and packet loss. Over the past decade, communication systems and networks have adopted scheduling mechanisms that can guarantee a certain level of QoS for a user. However, in recent years, a shift has taken place from QoS-oriented mechanisms to QoE-oriented mechanisms, due to the fact that the end-users have concerns about the received service quality.

DRL has been successfully used in wireless networks to respond to serious challenges such as dynamic spectrum access and access control (Chang et al., 2018; Chu et al., 2018). DRL algorithms such as A3C, DQN and distributed proximal policy optimization (DPPO) have great potential to solve problems with a high dimensional state-space. Hence, researchers exploit these potentials for many resource management problems, such as resource sharing and scheduling.

Chen et al. (2020c) investigate the problem of radio resource management in vehicle-to-vehicle networks. To be specific, the number and radio coverage of mounted RSU devices along roads determines the capability of traveling vehicles to communicate with the infrastructure to receive the necessary traffic information such as time and location. Nevertheless, RSU devices are fairly expensive. Thus, authorities prefer to install a limited number of RSU devices, especially in sparsely populated regions. Regarding the scarcely of RSU resources, effective management of these resources is critically important in vehicular environments. The authors in Chen et al. (2020c) propose a proactive DRL algorithm, by which RSU devices can allocate the frequency band and do packet scheduling in an optimal manner. They model the problem as a discrete-time single-agent MDP and use DQN+LSTM for solving it. One can refer to high-spatially mobile and temporally-varying incoming traffic as one of the most challenging aspects of this problem.

In Zhang et al. (2020b), the authors target the spectrum sharing problem in a cognitive radio network, in which there are primary users, SUs, and wireless sensors. They introduce two algorithms based on DRL, namely A3C and DPPO, to adjust the power control of users. A3C-based and DPPO-based algorithms benefit from two different algorithms for the optimization of the power allocation function, namely RMSProp and Adam optimization, respectively. The authors claim that the proposed algorithms can guarantee the QoS requirements of primary users/SUs. One of the big advantages of the proposed algorithms is that they can operate in a distributed manner. Moreover, the authors use the Hogwild (Recht et al., 2011) training technique to decline the cost of network parameters.

Similarly, Nasir and Guo (2019) investigates the problem of transmission power allocation in wireless networks. The main motivation for conducting this research is that existing methods for power allocation are not scalable and pose high computational complexity. Hence, this work proposes a model-free DRL method which can execute in a distributed fashion. In this method, each transmitter sets its transmission power based on the QoS information and CSI gathered from multiple neighbors. The main goal of the proposed method is to maximize network sum-rate/fair scheduling. The proposed method is compared with two well known centralized methods and shows better performance than its counterparts in terms of the speed of achieving sub-optimal power allocation. What distinguishes this method from its counterparts is its ability to utilize limited local CSI, where the other methods need the full CSI which is unrealistic in real-world applications.

Xu et al. (2017) investigate the resource allocation problem in Cloud Radio Access Network (C-RAN). C-RANs is an important technology towards enabling 5G mobile networks. Nevertheless, further research is needed to enhance resource allocation in C-RAN, and consequently reduce power consumption and satisfy the essential requirements of UEs over a long working time. To this end, the authors in Xu et al. (2017) exploit the potential of DRL to provide a power-efficient resource allocation technique in C-RAN. In the proposed technique, to lower the size of the state–action space, a novel two-step decision approach has been adopted. In the first step, an agent decides on the set of Remote Radio Heads (RRhs) to be turned on/asleep in order to reduce the state–action space. Given the set of active RRhs, the agent could achieve an optimal resource allocation policy by finding

a solution for a convex optimization problem. Unlike many resource allocation techniques, which ignore state transition overheads such as energy-wasting resulting from the transition from sleep/active to active/sleep state, the work in Xu et al. (2017) takes into account this aspect of the problem.

The work in Yang et al. (2020) by Yang et al. exploits the capabilities of the social IoT paradigm in cognitive D2D-enabled IoT systems. They provide a network model in order to combine these two items (*i.e.,* social IoT and D2D-enabled IoT), and consequently to meet different QoS requirements of IoT devices and optimize the network performance in terms of energy efficiency. Next, the authors model resource management as a multi-agent optimization problem and use DRL to solve it. An essential and missing aspect of the previously proposed research works is that they did not take into account different QoS requirements, such as latency and reliability; whereas, the work in Yang et al. (2020) provides a QoS-driven network model, in which QoS requirements are considered to carry out resource optimization.

Wang et al. (2019) improve the performance of network slicing by providing a ML-based dynamic resource scheduling method. They refer to the data produced by mobile devices and applications as an available opportunity to manage the network slicing resources. To this end, they extract knowledge and insight from the data to implement a dynamic resource scheduling technique for network slicing based on ML. The main objective of the proposed technique is realizing end-to-end service reliability and optimization of resources in an automatic and efficient manner. In this technique, DRL has been utilized to retrieve knowledge from the data, especially the data related to users, which is mainly due to privacy issues. DRL can retrieve knowledge through interacting with the environment (*i.e.,* the network) and gaining experience. Subsequently, the retrieved knowledge enables DRL to dynamically allocate the resources to different slices. By doing this, one not only maximizes resources utilization but also guarantees QoS requirements.

The work presented in Xu et al. (2020), investigated the issue of resource allocation in vehicular networks by employing multi-agent DDPG. In the paper, V2V communication was considered an agent to share the pre-allocated V2I frequency spectrum by the Non-Orthogonal Multiple Access (NOMA) scheme. Several essential considerations in vehicular environments must be taken into account when one designs resource allocation techniques, including (1) the fast-changing channel condition, and consequently obtaining CSI, and (2) maximization of V2I communications' sum-rate and meeting the latency and reliability requirements in V2V communications simultaneously. In response to these considerations, in Xu et al. (2020), decentralized Discrete-time and Finite-state Markov Decision Process (DFMDP) was used to model the resource allocation problem. For solving the modeled problem, the DDPG algorithm has been utilized as the algorithm that can manage high dimensional action spaces. Based on the simulation results, the proposed scheme could maximize V2I communications' sum-rate while guaranteeing V2V communications with the desired level of latency and reliability constraints.

A similar work based on the advantage of the actor–critic algorithm was presented in Chen et al. (2020b). To be more specific, virtualization technology and a DRL-based resource allocation technique were combined for resource allocation in vehicular networks to enhance QoS. Markov chains have formulated the dynamic changes of the network, and A3C has been used as a learning algorithm for solving it. Their simulation results confirmed the performance of the proposed technique in terms of convergence speed and the total reward.

Gu et al. adopted a knowledge-assisted DRL technique for wireless schedulers in the 5G network with time-sensitive data (Gu et al., 2020). In their work, the authors highlighted this fact that DDPG, as a widespread used algorithm for scheduling tasks in cellular networks, leads to poor performance in terms of convergence speed of QoS, especially in 5G networks which are highly non-stationary. Hence, they proposed knowledge-assisted DDPG (K-DDPG) as an extension to deal with these challenges. The proposed algorithm utilized the scheduler's expert knowledge to improve the convergence time and each UE's QoS. The simulation result reveals that the proposed algorithm reaches better QoS (decreasing packet losses) and convergence time than its counterparts.

In Table 5, we presented a comparison of the reviewed papers in this section.

5.1. Radio resource management (RRM)

The main goal of Radio resource management (RRM) is to make the most efficient use of the existing network resources, such as radio resources. RRM is considered enormously important for controlling power consumption. For example, during a low traffic load period in a BS, it is possible to save energy through RRM techniques (Oh and Krishnamachari, 2010). During an RRM process, different operations and management functions may be performed to monitor and control the resources. Broadly speaking, one can refer to the following items as the main RRM functions:

- Radio admission control: Admission control (AC) is an important process towards the optimization of radio resource usage and provisioning QoS through accepting or rejecting requests for the establishment of a new connection.
- Packet scheduling (PS): The PS is a key step towards RRM through allocation and deallocation of the network resources (*e.g.,* physical resource block) to users for packet transmission. Effective PS algorithms can maximize spectrum efficiency in the network.
- Load balancing (LB): The LB involves the distribution of the traffic load among different cells to achieve the maximum radio resource efficiency, as well as keep the QoS level as high as possible.

It should be noted that we list only functionalities that are related to the MAC layer. There are other functionalities for RRM, such as radio bearer control (RBC), and inter-cell interference coordination (ICIC) to name a few that are related to the PHY layer.

In modern large-scale networks (*e.g.,* IoT networks), RRM is a challenging task. This is mainly due to the complexity and high dimensional state-space of these networks. In other words, there is a direct correlation between the complexity of RRM and dimensionality of the problem. In addition, to realize RRM in modern networks, there are many network operations with multiple time scales, ranging from seconds to milliseconds (Calabrese et al., 2018). Hence, the conventional optimization techniques, such as convex optimization and fractional programming can no longer be used for these networks because these techniques are not scalable and are based on this assumption that a known optimal solution is available. Regarding the fact that DRL has remarkable abilities to solve scalable optimization problems on highly dimensional systems, such as the modern networks, we investigate the potential of DRL for RRM applications. DRL benefits from model-free learning strategy, the so-called model-free optimization, which is inherited from RL. Model-free optimization is a powerful technique for launching control operations when a priori knowledge on the environment and the actions/reactions between entities is not accessible. By applying model-free optimization techniques, an entity (*e.g.,* IoT device) learns and controls its behavior based on the interaction with its environment. These distinguishing characteristics of model-free optimization will work properly for online and dynamic RRM. Moreover, to apply RL to large-scale and dynamic systems such as cellular networks, NNs have been used to approximate the reward function in RL, under the umbrella of DRL. In summary, DRL can address some major challenges related to RRM, including large state-space dimensionality, complexity and fast running time, scalability, heterogeneity, and partial observations. In the following paragraphs, we review the contributions of DRL towards tackling RRM task.

Du et al. (2019) investigate green DRL for RRM. In this work, DRL comes under fierce criticism for its high energy consumption. The authors refer to this fact that the deep neural networks used in DRL

Table 5
DRL for resource sharing and scheduling.

Research	Year	Network	Model	Learning algorithm	Improved QoS factors	Technical contribution
Xu et al. (2017)	2017	C-RAN	MDP	DQN+ feedforward NN	QoE	Establishes a framework based on deep RL to decline power consumption in C-RAN and fulfill user requirements. In addition, the proposed framework is able to support highly dynamic conditions.
Nasir and Guo (2019)	2019	Wireless networks	MDP	DQN+five-layer NN	Throughput, Delay in CSI	Provides a distributed power allocation method based on deep RL. The method can achieve sub-optimal power allocation much faster than some well-known centralized methods.
Wang et al. (2019)	2019	5G network	MDP	DQN+CNN	Support SLA	Proposes a data-driven dynamic resource scheduling technique for 5G network slicing.
Chen et al. (2020c)	2020	Cellular networks (Vehicular)	discrete-time single-agent MDP	DQN+LSTM	Packet drops, Average Utility	Proposes a deep RL-based proactive resource management technique for vehicular environments, with a focus on RSU resources.
Zhang et al. (2020b)	2020	CR networks	MDP	DQN+NN	SINR- it has been supposed that SINR can delineate the QoS of a user.	Explores the spectrum sharing issue in cognitive radio networks and proposes QoS-aware algorithms for power allocation in these types of networks based on deep RL.
Yang et al. (2020)	2020	CIoT	MDP	DQN+ Improved deep NN	Delay, Reliability	To combine the social IoT with D2D-enabled IoT systems, this paper proposes a QoS-driven network model and consequently optimizes the performance of the systems in terms of energy consumption.
Xu et al. (2020)	2020	Cellular networks (Vehicular)	DFMDP	DDPG	Throughput, Delay, Reliability	Uses the DDPG algorithm to simultaneously guarantee the throughput and the latency/reliability requirements for V2I and V2V communications, respectively.
Chen et al. (2020b)	2020	Cellular networks (Vehicular)	MDP	A3C	Operator's revenue (or the received signal-to-noise ratio (RSNR)	Leveraged the network virtualization and A3C algorithm for resource allocation in connected vehicle networks, consequently improving QoS.
Gu et al. (2020)	2020	Cellular networks	MDP	K-DDPG	Packet losses	To improve convergence time and QoS it proposes a K-DDPG-based wireless scheduler for the cellular networks for time-sensitive traffic.

have a high number of parameters (weights) that need to be tuned. Tuning such a huge number of parameters calls for high computational and memory resources, and consequently high energy consumption. This is more challenging for energy- and resource-constrained electronic devices, such as IoT devices. This work investigates the feasible solutions for green DRL for RRM in terms of architecture and algorithm.

The work in Jiang et al. (2019) by Jiang et al. propose to use DRL for random access control in NB-IoT Networks in a real-time manner. More specifically, they address the problem of how much radio resources should be allocated to a set of IoT devices for random access and for data transmission to increase the maximum number of connected devices. To this end, the authors propose three DRL-based methods, including tabular Q-learning, Linear Approximation-based Q-learning (LA-Q), and DQN. Moreover, they compare the performance of these methods with the traditional methods (*i.e.,* load-based estimation methods). The simulation result demonstrates the superiority of the DRL-based methods over their rivals. Unlike the traditional methods, the proposed methods do not need historical information at the evolved Node B (eNB) for optimal radio resource allocation. This is crucially important because the factors such as the stochastic nature of the traffic at eNB, random collision, path-loss, and fading make it more and more difficult to have suitable a prior knowledge. Similar works have been performed in Tello-Oquendo et al. (2018) and Bello et al. (2014).

DRL has been proposed in Li et al. (2018) by Li et al. to manage resources in network slicing architectures. Resource management is a key step towards improving the performance and providing cost-efficient services in network slicing. This is mainly due to the fact that radio spectrum is scarce and expensive, the stochastic pattern of resource demand in each slice, and strict QOE requirements in network slicing techniques. Hence, network slicing looks forward to using AI-based solutions, such as DRL, to tackle the related challenge in this context. One of these challenges is having a trade-off between the level of users' activities per slice and the allocated resources (e.g radio resource). In this case, DRL has been employed to deal with the dynamics of requests from slice tenants to get adequate services, while the costs associated with the spectrum and computing are reasonable for providing operators. Similarly, Van Huynh et al. (2019) leverages DRL to establish a optimal and fast real-time resource slicing framework. The proposed framework enables real-time allocation of different resources, such as radio and computing resources to different slices. In this paper, a semi-Markov decision process has been employed to handle the dynamic and uncertain nature of demands in network slicing.

To optimize the handover (HO) process in large-scale communication systems, such as IoT networks, Wang et al. (2018) propose a technique to control the HO process based on DRL. First, they categorize UEs into different clusters according to UEs' patterns of mobility. Then, to control the HO processes in each cluster, they implement an asynchronous multi-user DRL technique. Using DRL can lower the HO process occurring rate while satisfying the QoS requirement in terms of system throughput. When it comes to Millimeter wave (mmWave) in 5G mobile networks, the HO process control becomes a more challenging task. This is mainly due to the small footprint mmWave antennas, and consequently high-density BSs deployments. The HO process in mmWave networks may affect QoS and QoE, as well as increase the signaling overhead. To minimize the HO process rate, Mollel et al. (2020) propose an offline double DRL scheme. The authors leverage offline learning in order to mitigate the side effects of online learning policies (*e.g.,* high computational costs).

Guo et al. (2020) tackled the problem of HO/power allocation in HetNets through using Multi-Agent RL (MARL). This joint optimization problem was modeled as a cooperative multi-agent task. Then, the authors introduced a MARL scheme based on the PPO method to solve the task, and consequently found decentralized policies for each agent (*i.e.,* UE). The comparison between the proposed scheme and the existing literature (*e.g.,* MADDPG) showed the superiority of the proposed

scheme, as it maximizes the overall throughput while lessening HO's frequency.

Motivated by resource management complexity in an emerging distributed learning paradigm, Tseng et al. focused on radio resource scheduling in 5G through DDPG (Tseng et al., 2019). In the proposed method, the combination of scheduling algorithms was considered as actions; hence, it operates more efficiently in training and performance. We presented a comparison of the reviewed papers in this sub-section in Table 6.

6. Challenges and future directions

In this section, we review the challenges of using DRL in QoS-provisioning at the MAC layer and discuss future directions and open issues. Most of existing challenges lie in the essence of modern networking and limitations of DRL as listed below:

- Theory of Network (ToN): It is a central challenge of using ML techniques in networks defined by David Meyer in Meyer (2016). ToN claims that networks suffer from the lack of a unified theory; therefore, ML models should be trained for each network separately. ToN also causes inefficiency in using benchmark datasets as the ML models trained based on them have lower performance in real-world networks. The shortage of representative datasets that can address the ToN challenge motivates us to combine heuristic and meta-heuristic techniques with ML techniques. In the beginning, heuristic techniques are used for networking. Then, the ML models are learned gradually by the results of the heuristic techniques resulting in reducing the negative impacts of heuristic methods, *e.g.,* network overheads (Xie et al., 2018). DRL techniques have a high potential to be integrated with heuristic and meta-heuristic methods as they can use explore–exploit techniques to learn the environment, but as the DRL model should be learned for each network, they need a considerable amount of resources and time to be learned.
- Non-stationary networks: modern networks can be very dynamic. The high dynamicity is considered as a big challenge which causes *concept drift* in ML techniques. Different ML techniques are introduced to solve the problem of concept drift, *e.g.,* online learning (Bifet et al., 2018). In some cases, *e.g.,* network behavior changes, high dynamicity triggers the ML techniques to be retrained (Shahraki et al., 2020a), but it is costly. Different light-weight DL and DRL techniques have been introduced, *e.g.,* Doriguzzi-Corin et al. (2020), Mnih et al. (2016), but the DRL techniques that can adapt the existing method with the changes by updating the model have a higher priority to be used in networking as retraining is very costly.
- Speed and Accuracy: Regarding the ToN and dynamicity concerns, it is evident that updating or retraining the DRL models is an inevitable task when DRL is used in networking. In most modern networking paradigms, the speed of the ML model is an important issue as the networks have not much time to retrain the ML model because of security concerns in unattended network management (Shahraki et al., 2020a). Reputation of DL models is that they are very time-consuming especially in the training phase. Some methods help reduce the delay of retraining, *e.g.,* lightweight deep learning and distributed learning, especially federated learning. There is no considerable literature showing the use of them in network management techniques.
- Online Learning and Stream Processing: Generally, network management data is growing, exposed as data streams. QoS parameters can be gathered as time-series datasets from the network to represent the network behavior. Most existing ML models are proposed for batch learning, but also there are some DL techniques to process streams, *e.g.,* RNN and LSTM. RL is also designed to explore–exploit a learning environment interactively. Although DRL techniques can potentially process streams, there are modern techniques, tools, and models for processing streams. Integrating DL, RL, and DRL techniques with online learning techniques helps explore–exploit the evolving data extracted from network management interactively and update their models gradually to avoid concept drift.
- Distributed DRL Processing: Distributed learning is not a new paradigm, but due to the characteristics of networks, *Federated learning*, as an emerging distributed learning paradigm (Yang et al., 2019), can satisfy the security and privacy concerns in terms of network management. Federated learning can enable DL or DRL to be (re)-trained in different machines, and then the trained model updates a global model interactively. The idea behinds Federated Learning is to avoid the sharing of users' data and update the ML model in a distributed manner. As the network management data is gathered from multiple machines in a network, distributed learning can have an excellent opportunity to use network management models. As DL and DRL models are compatible with distributed learning paradigms, they can be used in a distributed manner to analyze the performance of networks.
- SONs and Self-Sustaining Networks (SSNs): Modern networking paradigms, *e.g.,* 5G and 6G suffer from the shortage of intelligent network management models. SON is a paradigm to organize the networks based on automating different network management tasks, *e.g.,* joining and leaving the nodes. In real-world applications, SON is integrating with existing cellular generations, but also there is a strong need for SSN to maintain the network's key performance indicators, *e.g.,* SLA. SSN can be provided by ML techniques, especially by using distributed learning techniques. The 6G network, as a new emerging generation of cellular networks, needs to use ML techniques incredibly. At the intersection of ML techniques and SSN, DRL can be used to manage and sustain the network and nodes' resources automatically (Saad et al., 2019) as it can interact with the network continuously.
- Edge Intelligence: Edge computing is emerging as a solution to reduce resource consumption and delay of accessing services in large scale and complex networks. Edge intelligence is an aspect of edge computing in which ML techniques are applied at the level of edge nodes. Edge allows the distributed machine learning techniques to be realized over Edge platforms. On the other hand, ML techniques can help the Edge networks be established and maintained efficiently. DRL allows Edge networks to monitor the performance of the network and provide QoS. As DRL can interact with the networking environment, it has a high potential to improve the Edge intelligence (Deng et al., 2020).

7. Conclusion

DRL and modern networks, such as IoT, HetNets and UAVs have gained considerable attention in recent years. This is mainly due to the fact that DRL has shown its effectiveness in solving and realizing sequential decision-making tasks in complex real-world applications. Furthermore, new communication systems, *e.g.,* IoT, have proven to make a significant effect on various human live aspects, such as health monitoring, agriculture, autonomous cars, etc. Modern communication systems work in collaboration with DRL to deal with complexity, high dimensionality, and heterogeneity of such systems. In this paper, we surveyed the applications of DRL in network-level QoS provisioning. Specifically, we discussed three key aspects, in which DRL techniques have been used to deal with complex problems, such as network access and data rate control, and resource sharing and scheduling. We investigated dynamic spectrum access, joint user association and adaptive data rate control as three sub-aspects where DRL makes significant contributions. Highlighting the important advantages of DRL-based methods over traditional methods for various problems is also another focus of this survey. Finally, we highlighted a number of open research problems for future investigation.

Table 6
DRL for radio resource management.

Research	Year	Network	Model	Learning algorithm	Improved QoS factors	Technical contribution
Jiang et al. (2019)	2018	NB-IoT networks	POMDP	LA-Q, Tabular-Q, DQN, Multi-Agent DQN	Throughput	The paper leverages deep RL for the radio resource allocation problem in the NB-IoT network to maximize the throughput of the network and improve the training efficiency.
Li et al. (2018)	2018	Cellular networks	MDP	DQN	Delay, QoE	Introduces deep RL as a promising approach to deal with the technical challenges in network slicing concept, and consequently delivering better services in a cost-effective manner.
Wang et al. (2018)	2018	Wireless networks	MDP	DQN+ LSTM	Throughput	The work in this paper provides a two-layer framework in order to lower the HO rate and increase the system throughput.
Van Huynh et al. (2019)	2019	Cellular networks	semi-MDP	QN, Double DQN, deep dueling	System capacity	Proposes a fast and optimal deep RL-based framework for managing resources, such as radio and computation in the network slicing architecture.
Mollel et al. (2020)	2020	Wireless networks	MDP	DDQN	Throughput, Service delay	Establishes a framework based on deep RL for the HO process management in mmWave networks. The main goal of the paper is to reduce the HO process occurrence rate.
Tseng et al. (2019)	2019	Cellular networks	MDP	DDPG	Throughput, packet loss	Radio resource scheduling method based on DDPG, which picks a radio resource scheduling policy among the available combinations (*i.e.,* schedulers).
Guo et al. (2020)	2020	HetNets	MARL	PPO	Throughput	A multi-agent PPO algorithm for HO/power allocation problem in the HetNets with several UEs to increase the network throughput.

CRediT authorship contribution statement

Mahmoud Abbasi: Conceptualization, Investigation, Writing - original draft, Writing - review & editing, Resources. **Amin Shahraki:** Conceptualization, Validation, Methodology, Writing - original draft, Writing - review & editing, Data curation, Resources, Supervision, Project administration, Formal analysis. **Md. Jalil Piran:** Writing - review & editing, visualization, Resources. **Amir Taherkordi:** Writing - review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbasi, M., Shahraki, A., Barzegar, H.R., Pahl, C., 2021a. Synchronization techniques in "device to device-and vehicle to vehicle-enabled" cellular networks: A survey. Comput. Electr. Eng. 90, 106955.

Abbasi, M., Shahraki, A., Taherkordi, A., 2021b. Deep learning for network traffic monitoring and analysis (NTMA): A survey. Comput. Commun..

Adhikari, M., Amgoth, T., Srirama, S.N., 2019. A survey on scheduling strategies for workflows in cloud environment and emerging trends. ACM Comput. Surv. 52 (4), 1–36.

Ahad, N., Qadir, J., Ahsan, N., 2016. Neural networks in wireless networks: Techniques, applications and guidelines. J. Netw. Comput. Appl. 68, 1–27.

Al-Garadi, M.A., Mohamed, A., Al-Ali, A., Du, X., Ali, I., Guizani, M., 2020. A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun. Surv. Tutor..

Ali, R., Nauman, A., Zikria, Y.B., Kim, B.-S., Kim, S.W., 2019. Performance optimization of qos-supported dense WLANs using machine-learning-enabled enhanced distributed channel access (MEDCA) mechanism. Neural Comput. Appl. 1–9.

Allen, R., Masters, D., 2020. Artificial intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. In: ERA Forum, Vol. 20. Springer, pp. 585–598.

Alsheikh, M.A., Lin, S., Niyato, D., Tan, H.-P., 2014. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. IEEE Commun. Surv. Tutor. 16 (4), 1996–2018.

Amjad, M., Rehmani, M.H., Mao, S., 2018. Wireless multimedia cognitive radio networks: A comprehensive survey. IEEE Commun. Surv. Tutor. 20 (2), 1056–1103.

Aravanis, A.I., MR, B.S., Arapoglou, P.-D., Danoy, G., Cottis, P.G., Ottersten, B., 2015. Power allocation in multibeam satellite systems: A two-stage multi-objective optimization. IEEE Trans. Wireless Commun. 14 (6), 3171–3182.

Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A., 2017. A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866.

Ayoubi, S., Limam, N., Salahuddin, M.A., Shahriar, N., Boutaba, R., Estrada-Solano, F., Caicedo, O.M., 2018. Machine learning for cognitive network management. IEEE Commun. Mag. 56 (1), 158–165.

Ayyasamy, A., Venkatachalapathy, K., 2015. Context aware adaptive fuzzy based qos routing scheme for streaming services over MANETs. Wirel. Netw. 21 (2), 421–430.

Bannour, F., Souihi, S., Mellouk, A., 2018. Distributed SDN control: Survey, taxonomy, and challenges. IEEE Commun. Surv. Tutor. 20 (1), 333–354.

Bellemare, M.G., Dabney, W., Munos, R., 2017. A distributional perspective on reinforcement learning. In: International Conference on Machine Learning. pp. 449–458.

Bello, L.M., Mitchell, P., Grace, D., 2014. Application of Q-learning for RACH access to support M2m traffic over a cellular network. In: European Wireless 2014; 20th European Wireless Conference. VDE, pp. 1–6.

Benzekki, K., El Fergougui, A., Elbelrhiti Elalaoui, A., 2016. Software-defined networking (SDN): a survey. Secur. Commun. Netw. 9 (18), 5803–5833.

Bifet, A., Gavaldà, R., Holmes, G., Pfahringer, B., 2018. Machine Learning for Data Streams: with Practical Examples in MOA. MIT Press.

Bkassiny, M., Li, Y., Jayaweera, S.K., 2012. A survey on machine-learning techniques in cognitive radios. IEEE Commun. Surv. Tutor. 15 (3), 1136–1159.

Calabrese, F.D., Wang, L., Ghadimi, E., Peters, G., Hanzo, L., Soldati, P., 2018. Learning radio resource management in RANs: Framework, opportunities, and challenges. IEEE Commun. Mag. 56 (9), 138–145.

Canaan, R., Gao, X., Chung, Y., Togelius, J., Nealen, A., Menzel, S., 2020. Behavioral evaluation of hanabi rainbow dqn agents and rule-based agents. In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Vol. 16. pp. 31–37.

Chang, H.-H., Song, H., Yi, Y., Zhang, J., He, H., Liu, L., 2018. Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach. IEEE Internet Things J. 6 (2), 1938–1948.

Chen, S., Chen, J., Chen, J., 2020a. A deep reinforcement learning based network management system in smart identifier network. In: Proceedings of the 2020 4th International Conference on Digital Signal Processing. pp. 268–273.

Chen, M., Wang, T., Ota, K., Dong, M., Zhao, M., Liu, A., 2020b. Intelligent resource allocation management for vehicles network: An A3C learning approach. Comput. Commun. 151, 485–494.

Chen, X., Wu, C., Chen, T., Zhang, H., Liu, Z., Zhang, Y., Bennis, M., 2020c. Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective. IEEE Trans. Wireless Commun. 19 (4), 2268–2281.

Cheng, W., Zhang, X., Zhang, H., 2015. Optimal power allocation with statistical qos provisioning for D2d and cellular communications over underlaying wireless networks. IEEE J. Sel. Areas Commun. 34 (1), 151–162.

Chou, P.-Y., Chen, W.-Y., Wang, C.-Y., Hwang, R.-H., Chen, W.-T., 2020. Deep reinforcement learning for MEC streaming with joint user association and resource management. In: ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, pp. 1–7.

Chu, M., Li, H., Liao, X., Cui, S., 2018. Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems. IEEE Internet Things J. 6 (2), 2009–2020.

Dai, Y., Xu, D., Maharjan, S., Chen, Z., He, Q., Zhang, Y., 2019. Blockchain and deep reinforcement learning empowered intelligent 5g beyond. IEEE Netw. 33 (3), 10–17.

D'Alconzo, A., Drago, I., Morichetta, A., Mellia, M., Casas, P., 2019. A survey on big data for network traffic monitoring and analysis. IEEE Trans. Netw. Serv. Manag. 16 (3), 800–813.

Dankwa, S., Zheng, W., 2019. Twin-Delayed DDPG: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent. In: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing. pp. 1–5.

Darivianakis, G., Georghiou, A., Lygeros, J., 2018. Decentralized decision making for networks of uncertain systems. arXiv preprint arXiv:1803.07660.

Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., Zomaya, A.Y., 2020. Edge intelligence: the confluence of edge computing and artificial intelligence. IEEE Internet Things J..

Ding, H., Zhao, F., Tian, J., Li, D., Zhang, H., 2020. A deep reinforcement learning for user association and power control in heterogeneous networks. Ad Hoc Netw. 102, 102069.

Doriguzzi-Corin, R., Millar, S., Scott-Hayward, S., Martinez-del Rincon, J., Siracusa, D., 2020. Lucid: A practical, lightweight deep learning solution for ddos attack detection. IEEE Trans. Netw. Serv. Manag..

Du, Z., Deng, Y., Guo, W., Nallanathan, A., Wu, Q., 2019. Green deep reinforcement learning for radio resource management: Architecture, algorithm compression and challenge. arXiv preprint arXiv:1910.05054.

Fadlullah, Z.M., Tang, F., Mao, B., Kato, N., Akashi, O., Inoue, T., Mizutani, K., 2017. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun. Surv. Tutor. 19 (4), 2432–2455. http://dx.doi.org/10.1109/COMST.2017.2707140.

Feng, M., Mao, S., 2019. Dealing with limited backhaul capacity in millimeter-wave systems: A deep reinforcement learning approach. IEEE Commun. Mag. 57 (3), 50–55.

Fooladivanda, D., Rosenberg, C., 2012. Joint resource allocation and user association for heterogeneous wireless cellular networks. IEEE Trans. Wireless Commun. 12 (1), 248–257.

Fortunato, M., Azar, M.G., Piot, B., Menick, J., Hessel, M., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., et al., 2018. Noisy networks for exploration. In: International Conference on Learning Representations.

Fu, F., Kang, Y., Zhang, Z., Yu, F.R., Wu, T., 2020. Soft actor-critic DRL for live transcoding and streaming in vehicular fog computing-enabled iov. IEEE Internet Things J..

Fu, Z., Xu, W., Feng, Z., Lin, X., Lin, J., 2017. Throughput analysis of LTE-licensed-assisted access networks with imperfect spectrum sensing. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, pp. 1–6.

Gadaleta, M., Chiariotti, F., Rossi, M., Zanella, A., 2017. D-dash: A deep Q-learning framework for dash video streaming. IEEE Trans. Cogn. Commun. Netw. 3 (4), 703–718.

Gu, S., Holly, E., Lillicrap, T., Levine, S., 2017. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3389–3396.

Gu, Z., She, C., Hardjawana, W., Lumb, S., McKechnie, D., Essery, T., Vucetic, B., 2020. Knowledge-assisted deep reinforcement learning in 5g scheduler design: From theoretical framework to implementation. arXiv preprint arXiv:2009.08346.

Guo, D., Tang, L., Zhang, X., Liang, Y.-C., 2020. Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning. IEEE Trans. Veh. Technol..

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. pp. 1861–1870.

He, X., Wang, K., Huang, H., Miyazaki, T., Wang, Y., Guo, S., 2018. Green resource allocation based on deep reinforcement learning in content-centric IoT. IEEE Trans. Emerg. Top. Comput..

Hernandez-Garcia, J.F., Sutton, R.S., 2019. Understanding multi-step deep reinforcement learning: A systematic study of the DQN target. arXiv preprint arXiv:1901.07510.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., Silver, D., 2018. Rainbow: Combining improvements in deep reinforcement learning. In: Thirty-Second AAAI Conference on Artificial Intelligence.

Hoel, C.-J., Driggs-Campbell, K., Wolff, K., Laine, L., Kochenderfer, M.J., 2019. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. IEEE Trans. Intell. Veh. 5 (2), 294–305.

Hoßfeld, T., Seufert, M., Sieber, C., Zinner, T., 2014. Assessing effect sizes of influence factors towards a qoe model for HTTP adaptive streaming. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (Qomex). IEEE, pp. 111–116.

Hou, Y., Liu, L., Wei, Q., Xu, X., Chen, C., 2017. A novel DDPG method with prioritized experience replay. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 316–321.

Hu, X., Liu, S., Chen, R., Wang, W., Wang, C., 2018. A deep reinforcement learning-based framework for dynamic resource allocation in multibeam satellite systems. IEEE Commun. Lett. 22 (8), 1612–1615.

Huang, Y., Tan, J., Liang, Y.-C., 2017. Wireless big data: transforming heterogeneous networks to smart networks. J. Commun. Inf. Netw. 2 (1), 19–32.

Huang, T., Zhang, R.-X., Zhou, C., Sun, L., 2018. Qarc: Video quality aware rate control for real-time video streaming based on deep reinforcement learning. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 1208–1216.

IEC 23009-1, 2014. Dynamic adaptive streaming over http (dash)–part 1: media presentation description and segment formats. International Organization for Standardization (ISO).

Cisco Visual Networking Index, 2017. Global Mobile Data Traffic Forecast Update, 2016–2021. White Paper.

Jha, D.K., Raghunathan, A.U., Romeres, D., 2020. Quasi-newton trust region policy optimization. In: Conference on Robot Learning. PMLR, pp. 945–954.

Jiang, N., Deng, Y., Nallanathan, A., Chambers, J.A., 2019. Reinforcement learning for real-time optimization in NB-IoT networks. IEEE J. Sel. Areas Commun. 37 (6), 1424–1440.

Jiang, X., Yu, F.R., Song, T., Leung, V.C., 2020. Intelligent resource allocation for video analytics in blockchain-enabled internet of autonomous vehicles with edge computing. IEEE Internet Things J..

Kaelbling, L.P., Littman, M.L., Moore, A.W., 1996. Reinforcement learning: A survey. J. Artif. Intell. Res. 4, 237–285.

Kalidoss, T., Rajasekaran, L., Kanagasabai, K., Sannasi, G., Kannan, A., 2020. Qos aware trust based routing algorithm for wireless sensor networks. Wirel. Pers. Commun. 110 (4), 1637–1658.

Kan, N., Zou, J., Tang, K., Li, C., Liu, N., Xiong, H., 2019. Deep reinforcement learning-based rate adaptation for adaptive 360-degree video streaming. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4030–4034.

Kassab, R., Destounis, A., Tsilimantos, D., Debbah, M., 2020. Multi-agent deep stochastic policy gradient for event based dynamic spectrum access. arXiv preprint arXiv:2004.02656.

Kaur, T., Kumar, D., 2019. Qos mechanisms for MAC protocols in wireless sensor networks: a survey. IET Commun. 13 (14), 2045–2062.

Kaur, K., Kumar, S., Baliyan, A., 2020. 5g: a new era of wireless communication. Int. J. Inf. Technol. 12 (2), 619–624.

Khan, H., Elgabli, A., Samarakoon, S., Bennis, M., Hong, C.S., 2019. Reinforcement learning-based vehicle-cell association algorithm for highly mobile millimeter wave communication. IEEE Trans. Cogn. Commun. Netw. 5 (4), 1073–1085.

Klaine, P.V., Imran, M.A., Onireti, O., Souza, R.D., 2017. A survey of machine learning techniques applied to self-organizing cellular networks. IEEE Commun. Surv. Tutor. 19 (4), 2392–2431.

Knopp, R., Humblet, P.A., 1995. Information capacity and power control in single-cell multiuser communications. In: Proceedings IEEE International Conference on Communications ICC '95, Vol. 1. pp. 331–335. http://dx.doi.org/10.1109/ICC.1995.525188.

Lei, L., Tan, Y., Zheng, K., Liu, S., Zhang, K., Shen, X., 2020. Deep reinforcement learning for autonomous internet of things: Model, applications and challenges. IEEE Commun. Surv. Tutor..

Levine, S., Finn, C., Darrell, T., Abbeel, P., 2016. End-to-end training of deep visuomotor policies. J. Mach. Learn. Res. 17 (1), 1334–1373.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D., 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int. J. Robot. Res. 37 (4–5), 421–436.

Li, X., Li, L., Gao, J., He, X., Chen, J., Deng, L., He, J., 2015. Recurrent reinforcement learning: a hybrid approach. arXiv preprint arXiv:1509.03044.

Li, R., Zhao, Z., Sun, Q., Chih-Lin, I., Yang, C., Chen, X., Zhao, M., Zhang, H., 2018. Deep reinforcement learning for resource management in network slicing. IEEE Access 6, 74429–74441.

Liang, L., Ye, H., Li, G.Y., 2019. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. IEEE J. Sel. Areas Commun. 37 (10), 2282–2292.

Lin, Y., Bao, W., Yu, W., Liang, B., 2015. Optimizing user association and spectrum allocation in hetnets: A utility perspective. IEEE J. Sel. Areas Commun. 33 (6), 1025–1039.

Lin, L., Guan, X., Hu, B., Li, J., Wang, N., Sun, D., 2020. Deep reinforcement learning and LSTM for optimal renewable energy accommodation in 5g internet of energy with bad data tolerant. Comput. Commun..

Liu, X., Chong, E.K.P., Shroff, N.B., 2001. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. IEEE J. Sel. Areas Commun. 19 (10), 2053–2064.

Liu, S., Hu, X., Wang, W., 2018. Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems. IEEE Access 6, 15733–15742.

Logambigai, R., Kannan, A., 2014. QEER: Qos aware energy efficient routing protocol for wireless sensor networks. In: 2014 Sixth International Conference on Advanced Computing (ICoAC). IEEE, pp. 57–60.

Lu, D.W., 2017. Agent inspired trading using recurrent reinforcement learning and lstm neural networks. arXiv preprint arXiv:1707.07338.

Luong, N.C., Hoang, D.T., Gong, S., Niyato, D., Wang, P., Liang, Y.-C., Kim, D.I., 2019. Applications of deep reinforcement learning in communications and networking: A survey. IEEE Commun. Surv. Tutor. 21 (4), 3133–3174.

Mao, H., Chen, S., Dimmery, D., Singh, S., Blaisdell, D., Tian, Y., Alizadeh, M., Bakshy, E., 2019. Real-world video adaptation with reinforcement learning.

Mao, Q., Hu, F., Hao, Q., 2018. Deep learning for intelligent wireless networks: A comprehensive survey. IEEE Commun. Surv. Tutor. 20 (4), 2595–2621.

Mao, H., Netravali, R., Alizadeh, M., 2017. Neural adaptive video streaming with pensieve. In: Proceedings of the Conference of the ACM Special Interest Group on Data Communication. pp. 197–210.

Meyer, D., 2016. Machine intelligence and networks. In: IETF97.

Mishra, P., Varadharajan, V., Tupakula, U., Pilli, E.S., 2018. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Commun. Surv. Tutor. 21 (1), 686–728.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. pp. 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature.

Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M., 2018. Deep learning for IoT big data and streaming analytics: A survey. IEEE Commun. Surv. Tutor. 20 (4), 2923–2960.

Mollel, M.S., Abubakar, A.I., Ozturk, M., Kaijage, S., Kisangiri, M., Zoha, A., Imran, M.A., Abbasi, Q.H., 2020. Intelligent handover decision scheme using double deep reinforcement learning. Phys. Commun. 42, 101133.

Naparstek, O., Cohen, K., 2017. Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks. In: GLOBECOM 2017 - 2017 IEEE Global Communications Conference. pp. 1–7. http://dx.doi.org/10.1109/GLOCOM.2017.8254101.

Nasir, Y.S., Guo, D., 2019. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. IEEE J. Sel. Areas Commun. 37 (10), 2239–2250.

Oh, E., Krishnamachari, B., 2010. Energy savings through dynamic base station switching in cellular wireless access networks. In: 2010 IEEE Global Telecommunications Conference GLOBECOM 2010. pp. 1–5.

Özyılmaz, K.R., Yurdakul, A., 2020. Iot blockchain integration: A security perspective. In: Security Analytics for the Internet of Everything. CRC Press, pp. 29–54.

Qiu, C., Hu, Y., Chen, Y., Zeng, B., 2019. Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications. IEEE Internet Things J. 6 (5), 8577–8588.

Rafsanjani, M.K., Rezaei, A., Shahraki, A., Saeid, A.B., 2014. Qarima: A new approach to prediction in queue theory. Appl. Math. Comput. 244, 514–525.

Recht, B., Re, C., Wright, S., Niu, F., 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In: Advances in Neural Information Processing Systems. pp. 693–701.

Saad, W., Bennis, M., Chen, M., 2019. A vision of 6g wireless systems: Applications, trends, technologies, and open research problems. IEEE Netw. 34 (3), 134–142.

Sana, M., De Domenico, A., Strinati, E.C., 2019. Multi-agent deep reinforcement learning based user association for dense mmwave networks. In: 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, pp. 1–6.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Shahraki, A., Abbasi, M., Haugen, Ø., 2020a. Boosting algorithms for network intrusion detection: A comparative evaluation of real adaboost, gentle adaboost and modest adaboost. Eng. Appl. Artif. Intell. 94, 103770.

Shahraki, A., Abbasi, M., Piran, M., Chen, M., Cui, S., et al., 2021. A comprehensive survey on 6g networks: Applications, core services, enabling technologies, and future challenges. arXiv preprint arXiv:2101.12475.

Shahraki, A., Geitle, M., Haugen, Ø., 2020b. A comparative node evaluation model for highly heterogeneous massive-scale internet of things-mist networks. Trans. Emerg. Telecommun. Technol. 31 (12), e3924.

Shahraki, A., Haugen, Ø., 2018. Social ethics in internet of things: An outline and review. In: 2018 IEEE Industrial Cyber-Physical Systems (ICPS). IEEE, pp. 509–516.

Shahraki, A., Haugen, Ø., 2019. An outlier detection method to improve gathered datasets for network behavior analysis in IoT. J. Commun..

Shahraki, A., Rafsanjani, M.K., Saeid, A.B., 2017a. Hierarchical distributed management clustering protocol for wireless sensor networks. Telecommun. Syst. 65 (1), 193–214.

Shahraki, A., Taherkordi, A., Haugen, Ø., Eliassen, F., 2020c. A survey and future directions on clustering: From WSNs to IoT and modern networking paradigms. IEEE Trans. Netw. Serv. Manag..

Shahraki, A., Taherkordi, A., Haugen, Ø., Eliassen, F., 2020d. Clustering objectives in wireless sensor networks: A survey and research direction analysis. Comput. Netw. 107376.

Shahraki, A., Taherzadeh, H., Haugen, Ø., 2017b. Last significant trend change detection method for offline poisson distribution datasets. In: 2017 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, pp. 1–7.

Stotas, S., Nallanathan, A., 2011. Enhancing the capacity of spectrum sharing cognitive radio networks. IEEE Trans. Veh. Technol. 60 (8), 3768–3779.

Stoyanova, M., Nikoloudakis, Y., Panagiotakis, S., Pallis, E., Markakis, E.K., 2020. A survey on the internet of things (IoT) forensics: Challenges, approaches and open issues. IEEE Commun. Surv. Tutor..

Sun, Y., Peng, M., Mao, S., 2018. Deep reinforcement learning-based mode selection and resource management for green fog radio access networks. IEEE Internet Things J. 6 (2), 1960–1971.

Sutton, R.S., Barto, A.G., 2011. Reinforcement Learning: an Introduction. MIT Press, Cambridge, MA.

Tao, X., Hafid, A.S., 2020. Deepsensing: A novel mobile crowdsensing framework with double deep Q-network and prioritized experience replay. IEEE Internet Things J..

Tello-Oquendo, L., Pacheco-Paramo, D., Pla, V., Martinez-Bauset, J., 2018. Reinforcement learning-based ACB in LTE-a networks for handling massive M2m and H2h communications. In: 2018 IEEE International Conference on Communications (ICC). pp. 1–7.

Tian, X., Tian, Z., Pham, K., Blasch, E., Chen, G., 2013. Qos-aware dynamic spectrum access for cognitive radio networks. In: Sensors and Systems for Space Applications VI, Vol. 8739. International Society for Optics and Photonics, p. 87390P.

Tseng, S.-C., Liu, Z.-W., Chou, Y.-C., Huang, C.-W., 2019. Radio resource scheduling for 5g NR via deep deterministic policy gradient. In: 2019 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, pp. 1–6.

Van Huynh, N., Hoang, D.T., Nguyen, D.N., Dutkiewicz, E., 2019. Optimal and fast real-time resource slicing with deep dueling neural networks. IEEE J. Sel. Areas Commun. 37 (6), 1455–1470.

Wang, Y., He, H., Tan, X., 2020. Truly proximal policy optimization. In: Uncertainty in Artificial Intelligence. PMLR, pp. 113–122.

Wang, Z., Li, L., Xu, Y., Tian, H., Cui, S., 2018. Handover control in wireless systems via asynchronous multiuser deep reinforcement learning. IEEE Internet Things J. 5 (6), 4296–4307.

Wang, S., Liu, H., Gomes, P.H., Krishnamachari, B., 2018. Deep reinforcement learning for dynamic multichannel access in wireless networks. arXiv:1802.06958.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., Freitas, N., 2016. Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning. pp. 1995–2003.

Wang, H., Wu, Y., Min, G., Xu, J., Tang, P., 2019. Data-driven dynamic resource scheduling for network slicing: A deep reinforcement learning approach. Inform. Sci. 498, 106–116.

Watkins, C.J., Dayan, P., 1992. Q-learning. Mach. Learn. 8 (3–4), 279–292.

Wei, Y., Yu, F.R., Song, M., Han, Z., 2018. User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach. IEEE Trans. Wireless Commun. 17 (1), 680–692.

White, G., Nallur, V., Clarke, S., 2017. Quality of service approaches in IoT: A systematic mapping. J. Syst. Softw. 132, 186–203.

Wu, D., Chen, X., Yang, X., Wang, H., Tan, Q., Zhang, X., Xu, J., Gai, K., 2018. Budget constrained bidding by model-free reinforcement learning in display advertising. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1443–1451.

Wu, D., Dong, X., Shen, J., Hoi, S.C., 2020. Reducing estimation bias via triplet-average deep deterministic policy gradient. IEEE Trans. Neural Netw. Learn. Syst..

Xiao, X., 2008. Technical, Commercial and Regulatory Challenges of Qos: an Internet Service Model Perspective. Morgan Kaufmann.

Xie, J., Yu, F.R., Huang, T., Xie, R., Liu, J., Wang, C., Liu, Y., 2018. A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges. IEEE Commun. Surv. Tutor. 21 (1), 393–430.

Xu, Z., Wang, Y., Tang, J., Wang, J., Gursoy, M.C., 2017. A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs. In: 2017 IEEE International Conference on Communications (ICC). IEEE, pp. 1–6.

Xu, Y.-H., Yang, C.-C., Hua, M., Zhou, W., 2020. Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications. IEEE Access 8, 18797–18807.

Yang, Q., Liu, Y., Chen, T., Tong, Y., 2019. Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) 10 (2), 1–19.

Yang, H., Zhong, W.-D., Chen, C., Alphones, A., Xie, X., 2020. Deep reinforcement learning based energy-efficient resource management for social and cognitive internet of things. IEEE Internet Things J..

Ye, H., Li, G.Y., Juang, B.-H.F., 2019. Deep reinforcement learning based resource allocation for V2v communications. IEEE Trans. Veh. Technol. 68 (4), 3163–3173.

Yi, W., Zhang, X., Wang, W., Li, J., 2018. Multi-agent deep reinforcement learning based adaptive user association in heterogeneous networks. In: International Conference on Communications and Networking in China. Springer, pp. 57–67.

Yu, Y., Wang, T., Liew, S.C., 2019. Deep-reinforcement learning multiple access for heterogeneous wireless networks. IEEE J. Sel. Areas Commun. 37 (6), 1277–1290. http://dx.doi.org/10.1109/JSAC.2019.2904329.

Zhang, P., Hao, J., Wang, W., Tang, H., Ma, Y., Duan, Y., Zheng, Y., 2020a. Kogun: Accelerating deep reinforcement learning via integrating human suboptimal knowledge. arXiv preprint arXiv:2002.07418.

Zhang, Q., Liang, Y.-C., Poor, H.V., 2019. Intelligent user association for symbiotic radio networks using deep reinforcement learning. arXiv preprint arXiv:1905.04041.

Zhang, C., Patras, P., Haddadi, H., 2019. Deep learning in mobile and wireless networking: A survey. IEEE Commun. Surv. Tutor. 21 (3), 2224–2287. http://dx.doi.org/10.1109/COMST.2019.2904897.

Zhang, L., Tan, J., Liang, Y.-C., Feng, G., Niyato, D., 2019. Deep reinforcement learning-based modulation and coding scheme selection in cognitive heterogeneous networks. IEEE Trans. Wireless Commun. 18 (6), 3281–3294.

Zhang, H., Yang, N., Huangfu, W., Long, K., Leung, V.C., 2020b. Power control based on deep reinforcement learning for spectrum sharing. IEEE Trans. Wireless Commun..

Zhang, Z., Zheng, Y., Hua, M., Huang, Y., Yang, L., 2018. Cache-enabled dynamic rate allocation via deep self-transfer reinforcement learning. arXiv preprint arXiv: 1803.11334.

Zhao, N., Liang, Y.-C., Niyato, D., Pei, Y., Jiang, Y., 2018. Deep reinforcement learning for user association and resource allocation in heterogeneous networks. In: 2018 IEEE Global Communications Conference (GLOBECOM). IEEE, pp. 1–6.

Zhao, N., Liang, Y.-C., Niyato, D., Pei, Y., Wu, M., Jiang, Y., 2019. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks. IEEE Trans. Wireless Commun. 18 (11), 5141–5152.

Zheng, K., Zhao, L., Mei, J., Dohler, M., Xiang, W., Peng, Y., 2015. 10 gb/s hetsnets with millimeter-wave communications: access and networking-challenges and protocols. IEEE Commun. Mag. 53 (1), 222–231.

Zhong, C., Lu, Z., Gursoy, M.C., Velipasalar, S., 2019. A deep actor-critic reinforcement learning framework for dynamic multichannel access. IEEE Trans. Cogn. Commun. Netw. 5 (4), 1125–1139.

Zhou, X., Sun, M., Li, G.Y., Juang, B.-H.F., 2018. Intelligent wireless communications enabled by cognitive radio and machine learning. China Commun. 15 (12), 16–48.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A., 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3357–3364.

Zhu, J., Song, Y., Jiang, D., Song, H., 2018. A new deep-q-learning-based transmission scheduling mechanism for the cognitive internet of things. IEEE Internet Things J. 5 (4), 2375–2385. http://dx.doi.org/10.1109/JIOT.2017.2759728.