IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 11, 2020, accepted July 27, 2020, date of publication August 5, 2020, date of current version August 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014427

Deep Q-Network Learning Based Downlink Resource Allocation for Hybrid RF/VLC Systems

SHIVANSHU SHRIVASTAVA[1], BIN CHEN[1], (Senior Member, IEEE), CHEN CHEN[2], (Member, IEEE), HUI WANG[1], (Member, IEEE), AND MINGJUN DAI[1], (Member, IEEE)

[1]College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China
[2]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Corresponding author: Bin Chen (bchen@szu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61575126 and Grant 61901065, in part by the Natural Science Foundation of Guangdong under Grant 2018A0303130131 and Grant 2020A151501381, in part by the Basic Research Foundation of Shenzhen City under Grant JCYJ20170818091801577, and in part by the Industry-University-Research Innovation Fund of Science and Technology Development Center of Ministry of Education-New Generation Information Technology Innovation under Grant 2019J02002.

ABSTRACT Developing high data rate systems to meet the requirements of fifth generation mobile systems has become crucial. Hybrid radio frequency/visible light communication (RF/VLC) has appeared as a promising mechanism for achieving this objective. In hybrid RF/VLC, data rate maximization is subject to constraints on bandwidth, power and the user association. The joint optimization problem of bandwidth, power and user association to maximize the data rate is non-concave and obtaining an optimal solution is difficult with conventional optimization algorithms. The existing solutions are based on a presumption of at least one optimization variable. In this article, this issue has been overcome by solving the joint optimization problem in hybrid RF/VLC with a deep Q-network (DQN) learning based algorithm, which has been recognized as an efficient learning based mechanism for optimization. Our system model considers one RF and multiple VLC access points (APs). The idle APs are also incorporated in the system model. The application of DQN learning based algorithm is carried out by finding an optimal policy with the help of an action-value function. As the data sets for the considered system are large, a multi-layered network is used for approximating the action-value function estimator. Finally, a transfer learning based algorithm has been proposed for maximizing the total data rate of the system for the case of a newly entering user equipment (UE) that uses the information of the environment before the arrival of the new UE. Through simulations, it is found that our proposed algorithms can lead to an improvement of more than 10% and 54% in the achievable sum-rate and number of iterations for convergence respectively as compared to that obtained with existing conventional optimization algorithms.

INDEX TERMS Achievable sum-rate, access points, bandwidth, radio frequency (RF), visible light communication (VLC), hybrid RF/VLC, power, user equipment, user association, deep Q-network (DQN) learning.

I. INTRODUCTION

With the growing population of mobile internet users, the requirement for data rate has seen an exponential growth in the recent years. The use of conventional only-radio frequency (RF) systems may fail to fulfill it satisfactorily in the near future [1]. Telecommunication community is searching for alternative techniques to fulfill it. Visible Light Communication (VLC) has emerged as an efficient candidate in this regard [1]–[3]. It uses the deployed light emitting diode (LED) based light sources to transmit data through dimming of light, which is invisible to the eyes. VLC offers several advantages like high data rate, lesser interference with the co-existing RF devices, providing communication and illumination simultaneously, efficient unregulated spectrum usage, and efficient frequency reuse [3]. However, it has some disadvantages like inefficiency of non-line-of-sight (NLOS) components, which prevent it's stand-alone deployment [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Martalo.

As a solution to this problem, hybrid RF/VLC has been proposed in the literature [5], [6].

Hybrid RF/VLC merges the RF and the VLC networks into a single hybrid system. A typical hybrid RF/VLC architecture consists of some light sources, with each light source acting as a VLC access point (AP) in an indoor set-up. This set-up is supported by one or multiple RF APs. A user equipment (UE) present in the indoor set-up is associated either with a VLC AP or an RF AP for receiving data. The VLC AP offers high data rate while the RF AP ensures uninterrupted communication during blockage of LOS VLC signals to a UE, or when a UE is out of the coverage area of any of the VLC APs and fails to maintain the minimum needed signal-to-noise-ratio (SNR). In this manner, both the networks compensate for the limitations of each other.

Apparently, hybrid RF/VLC systems belong to the class of heterogenous networks (HetNets). In general HetNets, the joint optimization of resource allocation and association remains a significant research problem [7]–[13]. Similarly, resource allocation is a significant research issue in hybrid RF/VLC systems. Along with deciding the association of the UEs with the APs to receive the downlink data, the allocation of downlink bandwidth and transmission power to the APs for data transmission affects the achievable sum-rate of the system significantly. The study of optimal resource allocation for achievable sum-rate maximization in hybrid RF/VLC has received tremendous focus in research [14]–[28]. A common issue faced in these research works when the joint optimization of the downlink bandwidth, transmission power of the APs, and the association parameter are involved, is the non-concavity of the downlink resource allocation problem. Generally, this issue is solved by presuming values for at least one of these parameters and then obtaining the optimal values for the other parameters with conventional convex optimization algorithms. However, performing the joint optimization of all the three parameters without such presumptions is needed, as the association of UEs depends on their signal-to-interference-plus-noise ratios (SINRs) with the APs. Hence, it is directly affected by the allocation of downlink bandwidth and power and vice versa. Presuming a value for downlink bandwidth, transmit power of APs, or association parameter may not give the most optimal solution for maximizing the achievable sum-rate of the system. A comprehensive joint optimization problem incorporates the effects of each optimization parameter on the other one and on the objective function, which ensures a robust solution.

The above issue is the motivation behind the present study. We aim at jointly optimizing the downlink bandwidth, power and association parameter for maximizing the achievable sum-rate of a downlink hybrid RF/VLC system. The problem is subject to constraints pertaining to the availability of resources. Attempting conventional optimization approaches to solve this problem may lead to rigid bandwidth and power allocation as these approaches are less adaptive to the dynamics of the network, and result into an inefficient exploitation of resources [29], [30]. In contrast to this, a moment-to-moment optimal usage of the resources would result into a better output. It is also necessary that the optimal design for association and resource allocation in hybrid RF/VLC should not depend on prior knowledge of the environment. Some model based optimal solutions have been developed in [31]–[33] for specific models in general HetNets. However, the primary concern with these methods is that the incomplete information on the system makes the solutions intractable. Also, obtaining global maximum for a resource allocation problem is challenging with model based optimization methods due to it's non-concavity. The solutions based on game theory, linear programming, Markov approximation, college admission model, and dynamic programming proposed in [5], [7]–[9], [15], [18], [24], need almost accurate information which is not always possible to achieve practically, even when localization in VLC is relatively accurate.

In this article, a deep Q-network (DQN) learning based algorithm is developed for jointly optimizing the downlink bandwidth allocation, power allocation for APs, and association parameter, which maximizes the achievable data rate in a downlink hybrid RF/VLC system. The UEs can be associated with any of the APs lying within their field of views (FOVs). Unlike [34] where DQN was carried out at each AP, DQN is trained at a central unit (CU) which controls the association, allows all the APs to set their transmit powers and allocate bandwidths to the UEs associated with them [35].

Our contributions in this article can be summarized as follows:

1) *Comprehensiveness of the problem*: To the best of our knowledge, a comprehensive problem incorporating the joint optimization of association, power and bandwidth in a downlink hybrid RF/VLC has been considered for the first time in this article. Such a problem is neither convex nor concave. Making it convex requires prior assumption of at least one optimization parameter. Thus, it is difficult to solve with conventional optimization methods in the existing works. Here, this limitation is overcome by solving it with DQN based learning. The optimal solution obtained with the help of DQN based learning is not dependent on modeling errors and works on a moment-to-moment update.
2) *Considering idle APs*: Some APs can be switched off due to hardware malfunction while some APs may not take part in communication as only selected APs have been designed for VLC. Such APs do not cause interference to a UE. Considering interference from these APs can affect the robustness of the analysis. Our mathematical model considers idle APs in the SINR expression. Such formulation improves the practicality of the system model.
3) *Novel DQN based resource allocation in hybrid RF/VLC*: For the first time, a DQN based learning algorithm is being used for solving the optimal resource allocation and association problem in hybrid RF/VLC. Our developed algorithm allows the CU to adaptively allocate the downlink bandwidth, power and the

association parameter to the APs to maximize the achievable sum-rate of the system. It is not dependent on interaction among the UEs. A DQN based learning algorithm is trained at the CU, instead of training the DQN at each AP. This helps in providing an efficiently coordinated association. As the state and action vectors are very large in such problems, the application of DQN outperforms the existing algorithms in terms of achievable sum-rate and the number of iterations needed for convergence.

4) *Study the application of DQN with transfer learning*: The successful application of DQN with transfer learning has been shown for a newly entering UE in the hybrid RF/VLC set-up, where the experience of UEs already present in the set-up is transferred to a new UE entering into the set-up. It is found that DQN with transfer learning reduces the number of iterations for convergence by approximately 54% compared to when DQN without transfer learning is used for a newly joined UE.

The rest of the paper is organized as follows: In Section II, a literature review of the existing works that have led to the present work is performed. Section III explains the system model, where the light propagation model, the RF signal propagation model, the achievable data rate formulation, and the communication model obtained after the mixing of the RF and VLC networks is discussed, and the resource allocation problem is formulated. In Section IV, the solution for the resource allocation problem is designed, where the framework for learning has been formed and the layout of the proposed algorithm has been written. Section V illustrates the transfer learning based algorithm for the newly entering UEs. In Section VI, the simulation results to verify our proposed algorithms have been studied. The computational complexity and the NP hardness of the proposed schemes have also been studied in this section. Section VIII concludes the paper.

II. RELATED WORK AND THE SIGNIFICANCE OF THE APPLICATION OF DQN LEARNING IN HYBRID RF/VLC

Efficient resource allocation and association can lead to a higher achievable sum-rate in HetNets. The problem of resource allocation becomes crucial in hybrid RF/VLC as RF and VLC networks have completely different communication models. Several resource allocation schemes exist for performing achievable sum-rate maximization and related issues like energy efficiency maximization or packet loss probability minimization [14]–[28]. In [14], the total achievable data rate of a hybrid RF/VLC system is maximized by optimizing the association parameter, with the help of *minimum distance* condition. The focus of this work is on user association where each AP allocates equal bandwidth among the UEs associated to it. A fixed allocation of the transmit power of the AP has been considered here. In [15], achievable data rate maximization is performed with joint load balancing and optimal power allocation in hybrid RF/VLC. A fixed configuration of bandwidth has been taken here. In [16], the effect of bandwidth allocation on the overall sum-rate of the hybrid RF/VLC system has been studied. A bandwidth aggregation protocol to use VLC for increasing the bandwidth of the overall hybrid RF/VLC has been proposed. An optimal packet scheduling scheme is also proposed for the data packets which arrive at the system for transmitting to the UEs via. VLC or RF networks. The scheduling scheme has an impact on the overall sum-rate of the system, as the final objective of the work is throughput optimization. A fixed configuration of power allocation and association parameter has been considered here. In [17], maximization of the total sum-rate in hybrid RF/VLC has been carried out with joint balancing of the individual achievable sum-rates of the information UEs and the energy harvesting UEs. The power and the DC-bias of the UEs are optimized while a constant bandwidth and power allocation is considered for the APs. In [18], the focus is on maximizing energy efficiency of a hybrid RF/VLC system, which is defined as the ratio of the sum-rate and the total operational power, to optimize the bandwidth and power allocation. The system model of this work considers a single RF and VLC AP each. The association parameter and the bandwidth are kept fixed here. In [19], similar to energy efficiency, power efficiency maximization of a hybrid multiple access scheme for visible light communication systems has been studied which offers a better bandwidth allocation. The fundamental objective here is to fill the odd subcarriers optimally. Once again, the association parameter and the bandwidth are kept fixed here. In [20], power efficiency maximization for situations when illumination is not needed and the light source is kept on only for the transmission of data has been performed. In this situation, a VLC AP consumes more power than the RF AP. First, the number of APs needed to be switched on for satisfying the illumination requirements has been determined. Subsequently, the UEs request for real-time communication. Resource allocation remains outside the realm of this work as fixed bandwidth and association parameter have been considered here. The study in [18] has been further extended for multiple VLC APs in [21], but for only optimal power allocation. A fixed configuration of bandwidth and association are considered here. The authors in [22] perform minimization of packet-loss-probability in a fractional association time based dual-hop hybrid RF/VLC system enabled with energy harvesting. The fractional association time is based on time division multiplexing principle, where the entire bandwidth is allocated to each UE for a specific time fraction. The objective of [22] is to obtain the optimal fraction of association time allocated to a UE. Further, in [23], the joint optimization of the fractional association and power allocation has been studied in dual-hop hybrid RF/VLC. In [24], the sum-rate has been maximized with a mobility aware load balancing scheme, using the location-sensitive feature of VLC systems. The solution is based on a *college admission model* in a matching theory. A fixed configuration of association parameter and power allocation has been considered here. In [25], the focus of the work is on the achievable sum-rate maximization with

an intelligent selection of the network among the RF and the VLC networks depending on the dynamics of the environment. The study is performed on an uplink-downlink system with main focus on the non-similarity in the uplink-downlink parameters. The association parameter is optimized in terms of weighted proportional fairness while the bandwidth and the power allocation are kept fixed. In [26], the achievable sum-rate maximization for hybrid RF/VLC system has been investigated in a cross layer domain to provide optimal association. The solution depends on the effective capacity of the network obtained after imposing constraints on the buffer length of the data at the AP which holds the data before transmitting it over the selected link. A fixed allocation of bandwidth and power have been assumed here. In [27], the user association problem with lighting constraints for a VLC only system has been studied and a greedy algorithm for maximizing the SINR based utility function has been proposed. Further, in [28], *anticipatory association scheme* was proposed to anticipate the future locations of the UEs with the aim of maximizing the achievable sum-rate maximization. The association was performed as per the locations of the UEs. A fixed bandwidth and power allocation has been considered in [27] and [28].

As mentioned earlier, the existing works mentioned above presume a value for at least one parameter among bandwidth, power, and association parameter to address the issue of non-concavity in their respective joint optimization problems. However, such a presumption affects the robustness of the solution. To address this issue, we explore into learning based solutions. Reinforcement Learning (RL) [36] has been realized as an efficient learning mechanism. It is based on interaction with the environment and requires lesser prior information. It is an online learning method and has been extensively studied in artificial intelligence researches [37]. The most popular RL technique is Q-learning which was proposed in [38]. The convergence theorem for Q-learning was later proved in [39]. In [40], an autonomous Q-learning algorithm in HetNets for optimal resource allocation for device-to-device (D2D) communication has been proposed. A utility function defined as the difference between the achievable throughput and the cost of power consumption is formulated as a stochastic non-cooperative game. Each D2D pair is a player which becomes a learning agent with a task to learn it's best strategy. In [41], the association problem in vehicular networks was solved by using an online reinforcement learning approach. The authors take the advantage of the regularities in the features of vehicular networks. Ghadimi *et al.* proposed a reinforcement learning method to obtain rate adaptation in cellular networks in [42]. However, it should be noted that obtaining an optimal solution with Q-learning method is difficult when the state and action vectors of the joint optimization problem are large. In this regard, deep learning [43] has emerged as a promising technique to solve problems with large state and action vectors. Recently, deep learning-based methods have been used in many areas, such as dynamic channel access [44], power allocation [45], mobile offloading [46], cloud radio access networks [47], interference management [48], mobile edge computing and caching [49]. We first discuss the usability of deep learning in communication networks.

As the use of machine intelligence into future mobile communication networks is drawing tremendous research interest [50], [51], a flag ship of machine learning, deep learning is drawing tremendous research interest of communication networking researchers. In [52] and [53], it's potential to solve problems in the mobile networking domain has been explored. This encourages the use of deep learning in 5G mobile communication systems, which are largely heterogeneous. The data generated in these systems are also heterogeneous to a large extent, as they are received from sources of different formats having complex correlations [54]. Solving these problems with traditional machine learning tools is quite difficult, mainly because of no improvement in performance with more data [55] and inability to handle high dimensional state/action spaces [43]. In contrast, big data fuels the performance of deep learning, as it eliminates domain expertise and employs hierarchical feature extraction. Thus, deep learning has become an efficient candidate for solving problems in communication networks, particularly in heterogeneous systems. In this regard, a detailed account of researches on the applications of deep learning in communication systems can be found in works like [56], where deep learning approaches for network cybersecurity have been discussed, [57] which reviews deep learning approaches for network traffic control, [58] which presents deep learning approaches for physical layer modulation, network access/resource allocation, and network routing, and [59] which presents deep learning approaches for emerging issues including edge caching and computing, multiple radio access and interference management. The most significant advantage offerred by DRL is that it can obtain the solution of sophisticated network optimizations, enabling network controllers like base stations to solve non-convex and complex problems like joint user association, computation, and transmission schedule, and achieve optimal solutions without complete and accurate network information. Some of the major advantages offered by deep learning in communications are as follows:

- Deep learning ensures network entities to learn and create knowledge about the communication environment. For instance, by using deep learning, network entities like UEs can learn optimal policies, like AP selection, channel selection, handover decision, caching and offloading decisions, without knowing the channel model and mobility pattern.
- Deep learning enables autonomous decision-making. It enables the network entities to observe and obtain the best policy locally with minimum or without information exchange among each other. This significantly reduces communication overheads. It also improves security and robustness of the networks considerably.
- Deep learning improves the learning speed significantly, particularly where large state and action spaces are

involved. Hence, in large-scale networks, deep learning allows network controllers like base stations or APs to control dynamic user association, spectrum access, and transmit power for a massive number of devices and UEs.

- Deep learning has also been found efficient in solving game theory problems also. Several crucial problems in communications and networking such as cyber-physical attacks, interference management, and data offloading can be modeled as non-cooperative games. Deep learning has been recently used as an efficient tool in finding the Nash equilibrium, without complete information.

In a major development in this direction, it was found that combining deep neural network (DNN) with Q-learning can improve the learning performance and learning speed [60]. This system is called DQN. Developing a DRL- or DQN-based learning method for joint resource optimization is a new research direction in HetNets. For example, recently DQN based learning has been used specifically for base station activation in [62]. Further in [63], a post decision state based experience replay and transfer RL algorithm for low latency and high reliability has been proposed, for maximizing energy efficiency in hybrid RF/VLC networks. Some learning based works in hybrid RF/VLC have also been proposed in [25], which use RL with knowledge transfer based scheme for the selection of the network among the RF and the VLC networks, depending on the dynamics of the environment. However, hybrid RF/VLC systems generally involve large state and action spaces. RL algrithms perform well for small-size models but perform poorly for large-scale models. For such cases, DQN learning can efficiently maximize the Q-value by approximating the action-value function from the current state. However, it has still remained unexplored for finding optimal resource allocation in hybrid RF/VLC systems.

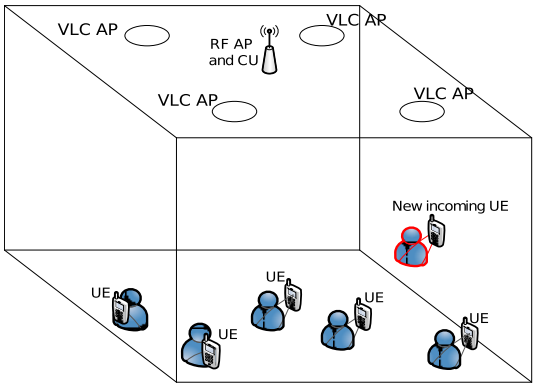

FIGURE 1. Hybrid RF/VLC system.

III. SYSTEM MODEL

Fig.1 shows the system model considered in this investigation. The set-up contains multiple VLC APs (light sources) and a single RF AP deployed on the ceiling of a typical room as shown in the figure. The CU is co-located with the RF AP system, which is responsible for controlling the network, viz. bandwidth allocation for the APs, transmit power control of the APs, and association of the UEs with the APs, with the help of the DQN algorithm. The users carrying the UEs are shown arbitrarily present on the floor of the room. A newly entering user carrying a UE is also shown at the border-line of the floor of the room. Let $\mathcal{N}$ be the set of APs indexed as $i = 0, 1, 2, \ldots, |\mathcal{N}|$. Index $i = 0$ denotes the RF AP while indices $i = 1, 2, \ldots, |\mathcal{N}| - 1$ denote the VLC APs. Let $\mathcal{M}$ be the set of UEs present inside the room indexed as $j = 1, 2, \ldots, |\mathcal{M}|$. The UEs are considered to be at height h from the floor. The downlink communication to a UE is done through the VLC and the RF networks. Each UE is associated to the RF AP or a VLC AP. VLC APs reuse the same bandwidth. Thus, inter-cell interference (ICI) is present in the VLC network. The investigations will be performed on a reference AP i -UE j pair for downlink communication. The data communication between VLC APs and the RF AP is done through a backhaul circuit [64]. The backhaul circuit also performs the underlying circuitry operations. A non-coordinated transmission has been considered in this set-up. When associated with a VLC AP, a UE receives data with LOS and reflected light ray components.

A. LIGHT PROPAGATION MODEL

The VLC APs transmit data to UEs on the donwlink. The light propagation in the VLC is modeled with diffused reflection, where the light ray incident on a surface is scattered at multiple angles. The optical power of light after undergoing diffused reflection is modeled by the Lambertian law [65] and is given as

$$P_o(\phi) = \frac{m+1}{2\pi} \cos^{m}(\phi) P_i, \quad \text{for } i \in \mathcal{N} \backslash \{0\}, \tag{1}$$

where P_i is the total LED power, ϕ is the angle of irradiance, m denotes the order of Lambertian radiation profile expressed as

$$m = -\frac{\ln 2}{\ln \cos \psi_{1/2}}, \tag{2}$$

where $\psi_{1/2}$ is the semi-angle at half illuminance of the LED. Let the LED emit light with wavelength λ and spectral power distribution $P_i(\lambda)$, P_i can be expressed as

$$P_i = \int_{\lambda} P_i(\lambda) d\lambda. \tag{3}$$

From (1), the LOS DC channel gain G_{ij}^{v} for the downlink communication from the ith VLC AP to UE j is obtained as

$$G_{ij}^{v} = \frac{(m+1) A_{\text{pd}} \cos^{m}\phi_{ij} \cos\psi_{ij} T_{\text{opt}}(\psi_{ij}) g(\psi_{ij})}{2\pi d_{ij}^{2}}, \tag{4}$$

where $T_{\text{opt}}(\psi_{ij})$ is the gain of the optical receiver filter and is unity or a constant value within the FOV of a receiver, ϕ_{ij} is the angle of irradiance at AP i, ψ_{ij} is the angle of incidence at UE j, and d_{ij} is the distance between AP i and the UE j. $g(\psi_{ij})$

TABLE 1. Meanings of important notations.

Notation	Meaning
i	The desired AP index
j	The desired UE index
k	The interferer AP index
r_{ij}	The achievable data rate between the AP i and the UE j
r_i	Downlink data rate of the AP i
r	Total achievable sum-rate
$B^v_{\max}$	Maximum bandwidth allotted to VLC AP
$B^r_{\max}$	Maximum bandwidth allotted to RF AP
P_i	Transmit power of the AP i
$\mathcal{N}$	Set of APs
$\mathcal{M}$	Set of UEs
m	Lambertian coefficient
ϕ	Angle of irradiance
ϕ_1, ϕ_2	Angle of irradiance at transmitter and reflecting point
ψ	Angle of incidence
ψ_1, ψ_2	Angle of incidence at reflecting point and the receiver
$G^{(p)}$	DC Channel gain after pth reflection
A_s	Incidence area
G_{EffRef}	Effective channel gain after reflection
$P_q^{(p)}$	The optical power of the reflected light wave at the pth reflecting point emitted from the qth transmitting AP
C	Lower bound on channel capacity for VLC networks
B_{ij}	bandwidth allotted to the UE j by the AP i
G_{ij}	Channel gain between AP i and UE j
n^v_j	Noise at the receiver of the UE while receiving VLC signals
n^r_j	Noise at the receiver of the UE while receiving RF signals
ρ_j	Responsivity of the receiver photo diode (PD)
pl_{0j}	Path-loss exponent for RF network
α_{ij}	Indicator function showing association of AP i-UE j
N^r_0	RF noise power
N^v_0	VLC noise power
$P^r_{\max}$	Maximum transmit power of the RF AP
$P^v_{\max}$	Maximum transmit power of the VLC AP

is the concentrator gain given as

$$g(\psi_{ij}) = \begin{cases} \dfrac{n^2}{\sin^2\psi_{\text{FOV}}} & \text{if } 0 \le \psi_{ij} \le \psi_{\text{FOV}} \\ 0 & \text{if } \psi_{ij} > \psi_{\text{FOV}}, \end{cases} \tag{5}$$

where n is the refractive index given as $n = \frac{\text{speed of light in vaccum}}{\text{speed of light in that optical material}}$, and ψ_{FOV} is the angle of FOV of the receiver UE.

Next, the channel gains of NLOS reflected light components received by the photo diode (PD) at a UE have been computed. The lth reflected light ray component is a light ray coming from the $(l-1)$th reflecting point. The $(l-1)$th reflecting point acts as a virtual light source and the lth reflecting point becomes the virtual receiver. Investigations in [65] find that the effective DC channel gain of the light ray undergoing various reflections G_{EffRef}, is the cummulative of the channel gains between all the pairs of reflecting points. Mathematically,

$$G_{\text{EffRef}} = \sum_{p=0}^{\infty} G^{(p)}, \tag{6}$$

where p denotes the index of reflection, $G^{(p)}$ is the DC channel gain after the pth reflection from the source LED which can be further expressed as

$$G^{(p)} = \int_S G_1 G_2 \ldots G_{p+1} P_q^{(p)} dA_s, \tag{7}$$

where dA_s is the infinitesimally small reflection surface area, $P_q^{(p)}$ is the optical power of the reflected light ray component after p reflections emitted from the qth transmitting VLC AP. The infinitesimally small area of the wall surface is considered as the variable for the above integration. $G_1, G_2, \ldots, G_{p+1}$ are DC channel gains of the path traced by each reflected component and are expressed as [65]

$$\begin{aligned} G_1 &= \frac{(m+1)A_s}{2\pi d_1^2}\cos^m(\phi_1)\cos(\psi_1), \\ G_2 &= \frac{A_s}{\pi d_2^2}\cos^m(\phi_2)\cos(\psi_2), \\ &\vdots \\ G_{p+1} &= \frac{A_s}{\pi d_{p+1}^2}\cos^m(\phi_{p+1})\cos(\psi_{p+1})T_{\text{opt}}(\psi_{p+1})g(\psi_{p+1}), \end{aligned} \tag{8}$$

where A_s is the incidence surface area, ϕ_b and ψ_b $(b = 1, 2, \ldots, p+1)$ are the irradiance and incidence angles at the bth reflection (b is a dummy variable). In (4), G^v_{ij} is the DC channel gain between ith VLC AP and the PD based jth receiver. On the other hand in (8), G_1 is the channel gain between the ith VLC AP and the first reflecting point, G_2 is the channel gain between the second and the third reflecting points, and similarly G_{p+1} is the channel gain between the pth reflecting point and the receiver PD. The channel gains at all the reflecting points are nearly in the same mathematical form. G_{p+1} is the function of T_{opt} and $g(\psi_{p+1})$ as T_{opt} and $g(\psi_{p+1})$ are properties of the receiving PD and G_{p+1} is the gain relating the last reflection point and the receiving PD.

Let $\Gamma_p(\lambda)$ be the spectral reflectance of the material at the pth reflecting point, then $P_q^{(p)}$ is given as

$$P_q^p = \int_\lambda P_i(\lambda)\Gamma_1(\lambda)\Gamma_2(\lambda)\ldots\Gamma_p(\lambda)d\lambda. \tag{9}$$

All the sufaces of all the reflecting points are assumed to be composed of the same material. As Γ_p is a function of λ, thus, it is assumed that $\Gamma_1(\lambda) = \Gamma_2(\lambda) = \ldots = \Gamma_p(\lambda) = \Gamma$.

The effective recieved optical power P_{eff} from a single LED will be the sum of the LOS and the NLOS components and is expressed as

$$P_{eff} = G_{\text{EffRef}}P_i + G^v_{ij}P_i = G_{ij}P_i \quad \text{for } i \in \mathcal{N}\backslash\{0\}, \tag{10}$$

where P_i is the power transmitted by VLC i and $G_{ij} = G_{\text{EffRef}} + G^v_{ij}$ is the effective channel gain between AP i and UE j for $i \in \mathcal{N}\backslash\{0\}$.

B. RF SIGNAL PROPAGATION MODEL

The signal received by UE j from the RF AP follows the RF signal propagation model, where the power channel gain includes fading as well as path loss. The received RF signal power is modeled using the WINNER-II channel model [66] given as

$$G_{0j} = L d_{0j}^{-pl_{0j}} \chi_{0j}, \tag{11}$$

where χ_{0j} is the Nakagami fading channel, pl_{0j} is the path-loss exponent and d_{0j} is the distance of UE j with the RF AP indexed as $i = 0$. Here, $L = 10^{X/10}, X = M + N \log_{10}\left(\frac{f_c}{5}\right), f_c$ is the carrier frequency in GHz, M and N are the propagation constants depending on the propagation model. In an LOS environment, $M = 46.8$ and $N = 20$ while in a non-LOS environment, $M = 43.8$ and $N = 20$. The Nakagami fading channel χ_{0j} has a gamma distribution fading power. It is a general fading distribution. It approximates to Rayleigh distribution when $\kappa = 1$ and Rician fading distribution when $1 \geq \kappa \leq \infty$.

C. ACHIEVABLE DATA RATE

As the objective of our work is maximization of the achievable sum-rate of hybrid RF/VLC systems, developing insight on the achievable data rate of a UE, when it is associated with RF or a VLC AP, is significant. During a UE's association with the RF AP, it's achievable data rate will be expressed by the Shannon's capacity formula. On the other hand, when a UE is associated with a VLC AP, it's communication is based on intensity modulation/direct detection (IM/DD) of light. In this scheme, the signal amplitude depicts the instantaneous optical power. Consequently, there are constraints on the signal to be real-valued and non-negative. Due to these constraints, direct application of Shannon capacity formula may not fulfil the purpose of obtaining the achievable data rate.

The authors in [67]–[70] have investigated the capacity of an IM/DD channel corrupted by the Gaussian noise. In [68], investigations show that the channel capacity in VLC networks can be approximated by it's lower bound as

$$C = \frac{1}{2} B \log_2\left(1 + w\frac{\rho^2 P_{eff}^2}{\sigma^2}\right), \tag{12}$$

where w is a constant and is given as $w = e/2\pi$ (e is the Euler's number), ρ is the responsivity of the PD, B is the modulation bandwidth, P_{eff} is the received optical power and σ^2 is the Gaussian noise power. It was found in [68] that a factor of $\frac{1}{2}$ appears as a result of various constraints in VLC. It was also found that the expression (12) is accurate and for a high SNR, it is found to concur with the upper bound also.

D. COMMUNICATION MODEL

Each UE will receive data from the RF AP or from one of the VLC APs. It's association will be decided with the help of the DQN based learning algorithm proposed ahead. Some APs are likely to be idle and no UE will be associated with them. For UE j associated with AP i for $i \in \mathcal{N}$, the channel gain vector will be written as $G_j = [G_{0j}, G_{1j}, G_{2j}, \ldots, G_{|\mathcal{N}|j}]$ where $G_{0j} \in \mathbb{C}$, $[G_{1j}, G_{2j}, \ldots, G_{|\mathcal{N}|j}] \in \mathbb{R}_{\geq 0}^{(|\mathcal{N}|-1)\times 1}$ and $G_{ij} \in G_j$ denotes the channel gain between UE j and AP i. The signal transmitted by the APs will be represented in the vector form as $x = [x_0, x_1, x_2, \ldots, x_{|\mathcal{N}|}]$, where $x_0 \in \mathbb{R}$ and $[x_1, x_2, \ldots, x_{|\mathcal{N}|}] \in \mathbb{R}_{\geq 0}^{(|\mathcal{N}|-1)\times 1}$. Remember that index $i = 0$ in the above sets denotes the RF AP. Let us consider that the UE j is associated to AP i. When UE j is associated to AP $i = 0$ i.e., the RF AP, it will receive signal y_j expressed as

$$y_j = \sqrt{G_{0j} P_0} \times x_0 + n_j^r, \quad \text{for } i = 0, \tag{13}$$

where n_j^r is the additive white Gaussian noise (AWGN). On the other hand, when UE j is associated with ith VLC AP, y_j will be expressed as

$$y_j = \rho_j G_{ij} P_i x_i + \sum_{k \in N\backslash\{i\}} \rho_j G_{kj} P_k x_k D_k(\alpha_{kj'}) + n_j^v, \quad \text{for } i \in \mathcal{N}\backslash\{0\}, \tag{14}$$

where ρ_j is as mentioned in (12), the responsivity of the receiving PD at the UE j, n_j^v includes the shot noise and thermal noise, and

$$D_k(\alpha_{kj'}) = \left(1 - \prod_{j' \in \mathcal{M}\backslash\{j\}} (1 - \alpha_{kj'})\right), \tag{15}$$

where $\alpha_{kj'}$ is an indicator function denoting the association of the AP k with UE j' such that

$$\alpha_{kj'} = \begin{cases} 1 & \text{if AP } k \text{ is associated to UE } j' \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

In (14), $\alpha_{ij} = 1$ means UE j is associated to AP i. AP i - UE j are the desired AP-UE pair while AP k is causing interference at the jth receiving UE. The first term in (14) represents the desired signal whereas the second term denotes interference. Note that a conventional form of the expression does not have $D_k(\alpha_{kj'})$ in the interference term. We multiply $D_k(\alpha_{kj'})$ in the interference term to include the case of idle APs which are not transmitting. It ensures that AP k is considered as the interferer only if it is transmitting to at least one UE j', where $j' \neq j$. The parameter $\alpha_{kj'}$ signifies the association of UE j' with AP k. $D_k(\alpha_{kj'}) = 0$ and 1 if AP k is not transmitting and transmitting to UE j' respectively. This factor incorporates the situation when an AP is momentarily switched off due to hardware failure.

Following (12), (13), and (14), the instantaneous achievable data rate at UE j for the input signal which is continous and follows negative exponential distribution is expressed

as [13]

$$r_{ij} = \begin{cases} B_{0j}\log_2\left(1+w\gamma_{0j}\right), & \text{for } i=0 \text{ and} \\ \frac{1}{2}B_{ij}\log_2\left(1+w\gamma_{ij}\right), & \text{for } i\in\mathcal{N}\backslash\{0\}, \end{cases} \tag{17}$$

where γ_{0j} and γ_{ij} are the lower bounds of $SINR_{0j}$ and $SINR_{ij}$ which are given as

$$SINR_{0j} = \frac{P_0 G_{0j}}{N_0^r B_{0j}}, \text{ and}$$

$$SINR_{ij} = \frac{\rho_j^2 G_{ij}^2 P_i^2}{N_0^v B_{ij} + \sum_{k\in N\backslash\{i\}} \rho_j^2 G_{kj}^2 P_k^2 \left(1-\prod_{j'\in M\backslash\{j\}}(1-\alpha_{kj'})\right)^2}, \tag{18}$$

where B_{0j} is the bandwidth of the RF AP ($i = 0$) - UE j link and B_{ij} is the bandwidth of VLC AP i - UE j link such that $i \in \mathcal{N}\backslash\{0\}$. As only one RF AP has been considered in the model, it is assumed that the RF signals suffer negligible interference. Thus, when the UE j is connected to the RF AP, we are interested in the SNR. However, for the sake of consistency in notations, the SNR for RF AP-UE j link is expressed as $SINR_{0j}$. When UE j is connected to a VLC AP, SINR will be of interest. Note that, any general mention of SINR further will mean SNR in the case of RF signals. Based on the above expression for instantaneous data rate, the throughput of AP i can be formulated as

$$r_i = \sum_{j\in\mathcal{M}} \alpha_{ij} r_{ij}, \quad \text{for } i\in\mathcal{N}. \tag{19}$$

E. THE RESOURCE ALLOCATION PROBLEM

This article aims for finding the optimal user association, transmit power allocation for APs, and the optimal downlink bandwidth allocation done by an AP for the UEs associated with it. The resource allocation will be done for maximizing r_i obtained in (19). The resource allocation problem is formulated as

$$\mathcal{P}: \max_{B_{ij},P_i,\alpha_{ij}} r_i, \quad \text{for } i\in\mathcal{N}, j\in\mathcal{M}, \tag{20}$$

subject to the following constraints:

$$\mathcal{C}_1: \sum_{j\in\mathcal{M}} \alpha_{ij} B_{ij} \le B_{\max}^v, \quad \text{for } i\in\mathcal{N}\backslash\{0\}, \tag{21}$$

where $B_{\max}^v$ is the total bandwidth that can be allocated to a VLC AP. The constraint in (21) illustrates that the sum of bandwidths allocated to the UEs associated to VLC AP i for $i \in \mathcal{N}\backslash\{0\}$ cannot exceed $B_{\max}^v$. Similar constaint is imposed on the RF AP formulated as follows:

$$\mathcal{C}_2: \sum_{j\in\mathcal{M}} \alpha_{0j} B_{0j} \le B_{\max}^r, \quad \text{for } i=0, \tag{22}$$

The constraint in (22) shows that the sum of bandwidths allocated to UEs associated with the RF AP cannot exceed $B_{\max}^r$, which is the total bandwidth allocated to the RF AP. The next constraint is imposed on the transmission power to ensure the power budget saving and safety considerations for the eyes. The transmission power of a VLC AP cannot exceed it's maximum available power $P_{\max}^v$, formulated as

$$\mathcal{C}_3: 0 \le P_i \le P_{\max}^v, \quad \text{for } i\in\mathcal{N}\backslash\{0\}, \tag{23}$$

Similarly, the transmission power of an RF AP cannot exceed it's maximum available power $P_{\max}^r$ formulated as:

$$\mathcal{C}_4: 0 \le P_0 \le P_{\max}^r, \quad \text{for } i=0, \tag{24}$$

Additional constraints have been imposed on $SINR_{ij}$ for $i \in \mathcal{N}, j \in \mathcal{M}$ for achieving reliable communication. Let the minimum level for SINR required by the jth UE from the ith AP for successful communication be γ_{ij}. Thus, the constraint on the SINR is as follows:

$$\mathcal{C}_5: SINR_{ij} \ge \gamma_{ij}, \quad \text{for } i\in\mathcal{N}, j\in\mathcal{M}, \tag{25}$$

In the constraint $\mathcal{C}_5$, γ_{ij} is the minimum threshold for $SINR_{ij}$. For the calculations in this work, we consider $SINR_{ij} = \gamma_{ij}$. Equality is assumed here to carry out the optimization of the variables B_{ij}, P_i and α_{ij}. As the optimization of B_{ji}, P_i and α_{ij} will lead to the optimization of γ_{ij}, taking equality as

$$SINR_{ij} = \gamma_{ij} \tag{26}$$

facilitates the solution.

When the constraint $\mathcal{C}_5$ in (25) holds with equality, the following conditions are obtained for preventing SINR constraint voilation [11], [12]

$$1 - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} \xi_{ij} > 0, \text{ and}$$

$$\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} \beta_i \xi_i \le 1, \tag{27}$$

where

$$\xi_{ij} = \left(1+\frac{1}{\gamma_{ij}}\right)^{-1}, \text{ and} \tag{28}$$

$$\beta_{ij} = \frac{N_0 B_{ij}}{(G_{ij}P_i/\gamma_{ij}) - N_0 B_{ij}} + 1. \tag{29}$$

Constraints (25), (27), (28), and (29) are significant for controlling interference in the system. It is possible that the maximization of the achievable data rates for different APs, namely r_i, interfere with each other due to the inter-AP interference. Thus, maximizing the achievable data rates for different APs at the same time will be difficult. The constraint (25) ensures that a minimum SINR threshold for every AP - UE pair is maintained. The minimum SINR threshold has been denoted as γ_{ij}. Putting a minimum SINR constraint on each AP-UE pair ensures a cap on the interference caused by the APs. When the interference from an AP increases to a level that violates this SINR constraint at some UE, the DQN learning mechanism will regulate the transmission power of

the interfering AP in a manner that the SINR constraint is satisfied. The constraint in (25) leads to the constraints in (27) - (29), which are used to incorporate (25) in the algorithm while solving the optimization problem. An AP will interfere with the signals of another AP if the constraints in (27) - (29) are violated. This process will be accomplished with the help of the *state space vector* $\mathcal{S}_{ij}$ formulated in the subsequent section.

Note that the sum of logarithmic functions is concave in nature. However, problem $\mathcal{P}$ in (20) is jointly non-concave in B_{ij}, P_i and α_{ij} (please refer section Appendix for proof).

IV. DQN-BASED LEARNING ALGORITHM FOR RESOURCE ALLOCATION IN HYBRID RF/VLC

Now, a DQN-based learning algorithm to maximize the network throughput in (20) is developed.

A. FRAME WORK FOR LEARNING

In this section, a DQN-based learning algorithm for the resource allocation problem in (20) has been formulated. The proposed algorithm maximizes the achievable data rate of AP i in (20) while satisfying the constraints in (21)-(27). Learning based algorithms run with the help of three vector variables: *state*, *action*, and *reward*. The state vector defines the present status of the environment. The action vector defines the action taken after observing the present status of the environment. The reward vector defines the reward received by the system after an action is taken by the system. Let $\mathcal{S}_{ij} = \{s_{ij}^1, s_{ij}^2, \ldots, s_{ij}^l\}$ be the state vector and $\mathcal{A}_{ij} = \{a_{ij}^1, a_{ij}^2, \ldots, a_{ij}^m\}$ be the action vector. l and m depend on the formulations of $\mathcal{S}_{ij}$ and $\mathcal{A}_{ij}$. At time t, the system is in the state $s_{ij}(t) \in \mathcal{S}_{ij}$ and it receives reward $R_i(s, a)$. When action $a_{ij}(t) \in \mathcal{A}_{ij}$ is taken on the system, it moves to state $s_{ij}(t+1) \in \mathcal{S}_{ij}$. The outcome of action $a_{ij}(t)$ is received in terms of the reward. The CU trains the learning algorithm to perform the association of UEs and communicate the power and bandwidth allocation with APs. This process is repeated iteratively. With each iteration, the system moves towards receiving the maximum reward. The action vector, state vector and reward are formulated as follows:

1) ACTION SPACE ($\mathcal{A}_{ij}$)

As it can be seen in (20), the association and resource allocation variables are α_{ij}, B_{ij}, and P_i, the action space $\mathcal{A}_{ij}$ will be formulated with α_{ij}, B_{ij}, and P_i for $i \in \mathcal{N}$ and $j \in \mathcal{M}$. Let $\mathbb{B}_{ij}$ and $\mathbb{P}_i$ be the discretized sets of B_{ij} and P_i respectively, for $i \in \mathcal{N}$ and $j \in \mathcal{M}$. The following formulation is made for $\mathbb{B}_{ij}$ and $\mathbb{P}_i$:

$$\mathbb{B}_{ij} = \left\{ 0, B_{\min}^{r/v} \left(\frac{B_{\max}^{r/v}}{B_{\min}^{r/v}} \right)^{\frac{u}{(|\mathbb{B}_{ij}|-2)}}, u = 0, 1, 2, \ldots, |\mathbb{B}_{ij}| - 2, \right. \tag{30}$$

where $B_{\min}^{r/v}$ and $B_{\max}^{r/v}$ are minimum and maximum values of B_{ij} for RF and VLC APs respectively. Similarly, $\mathbb{P}_i$ is obtained as

$$\mathbb{P}_i = \left\{ 0, P_{\min}^{r/v} \left(\frac{P_{\max}^{r/v}}{P_{\min}^{r/v}} \right)^{\frac{u}{(|\mathbb{P}_i|-2)}}, u = 0, 1, 2, \ldots, |\mathbb{P}_i| - 2, \right. \tag{31}$$

where $P_{\min}^{r/v}$ and $P_{\max}^{r/v}$ are minimum and maximum levels of the transmit power for the RF and the VLC APs respectively. Note that the cardinality of α_{ij} will be $2^{|\mathcal{N}|\times|\mathcal{M}|}$ for $i \in \mathcal{N}$ and $j \in \mathcal{M}$. The design of $\mathcal{A}_{ij}$ involves $2^{|\mathcal{N}|\times|\mathcal{M}|}$ values of α_{ij} because of the presence of the interference term in (18). Without the loss of generality, $|\mathbb{B}_{ij}| = |\mathbb{P}_i| = 2^{|\mathcal{N}|\times|\mathcal{M}|}$ has been considered for the formulation of $\mathcal{A}_{ij}$. The discretized values $\mathbb{B}_{ij}$, $\mathbb{P}_i$ and α_{ij} will be used to compute the threshold γ_{ij} according to (18) on each link from the ith AP to the jth UE and the action state vector will be formulated as

$$\mathcal{A}_{ij} = \{\gamma_{ij}^1, \gamma_{ij}^2, \ldots, \gamma_{ij}^{|\mathbb{B}_{ij}|\times|\mathbb{P}_i|\times|\alpha_{ij}|}\}. \tag{32}$$

At every iteration, the CU will chose one value from the set $\mathcal{A}_{ij}$ for each AP. While choosing a strategy from $\mathcal{A}_{ij}$, the CU adapts the transmit power $\mathbb{P}_i$ and the bandwidth allocation $\mathbb{B}_{ij}$ for the ith AP (such that $j \in \mathcal{M} \backslash \{\alpha_{ij} = 0\}$), and observes the changes in the environment and it's own transmission. Thus, the action is the selection of $\mathbb{B}_{ij}$ and $\mathbb{P}_i$ to achieve a minimum SINR (γ_{ij}). Next, the state space vector has been designed.

2) STATE SPACE ($\mathcal{S}_{ij}$)

The state space vector is based on constraints defined in (21)- (29) and is defined with binary variables as $\mathcal{S}_{ij} = \{I_1^{ij}, I_2^{ij}, \ldots, I_6^{ij}\}$, where

$$I_1^{ij} = \begin{cases} 0 & \text{if } \sum_{j\in\mathcal{M}} \alpha_{ij} B_{ij} \le B_{\max}^{v}, \text{ for } i \in \mathcal{N}\backslash\{0\}, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases}$$

$$I_2^{ij} = \begin{cases} 0 & \text{if } \sum_{j\in\mathcal{M}} \alpha_{0j} B_{0j} \le B_{\max}^{r}, \text{ for } i = 0, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases}$$

$$I_3^{ij} = \begin{cases} 0 & \text{if } 0 \le P_i \le P_{\max}^{v}, \text{ for } i \in \mathcal{N}\backslash\{0\}, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases}$$

$$I_4^{ij} = \begin{cases} 0 & \text{if } 0 \le P_0 \le P_{\max}^{r}, \text{ for } i = 0, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases}$$

$$I_5^{ij} = \begin{cases} 0 & \text{if } \sum_{i\in\mathcal{N}, j\in\mathcal{M}} \xi_{ij}(\gamma_{ij}) < 1, \text{ for } i \in \mathcal{N}, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases}$$

$$I_6^{ij} = \begin{cases} 0 & \text{if } \sum_{i\in\mathcal{N}, j\in\mathcal{M}} \beta_{ij}\xi_{ij}(\gamma_{ij}) < 1, \text{ for } i \in \mathcal{N}, j \in \mathcal{M}, \\ 1 & \text{otherwise.} \end{cases} \tag{33}$$

It can be seen that the formulations in (25)-(29) help in creating the state vector, so as to help the proposed DQN in maintaining a tradeoff between the desired signal power

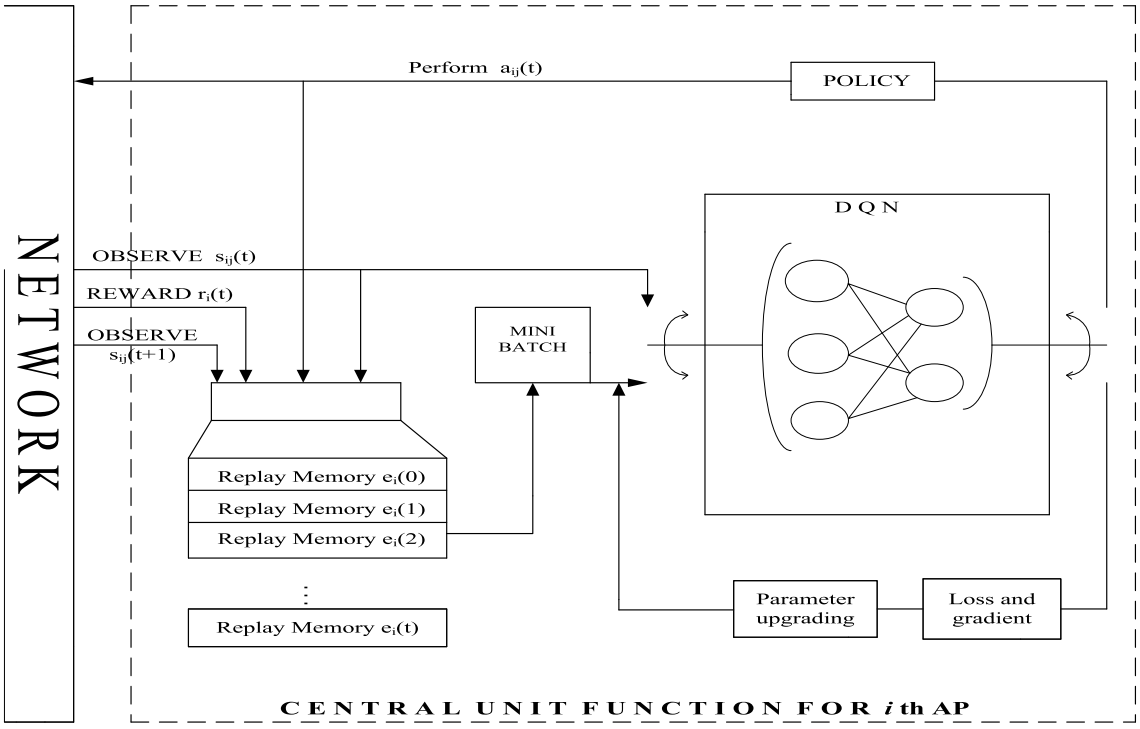

FIGURE 2. The DQN learning for the *i*th AP at the CU.

and the interference suffered. Note that the total number of possible states will be 2^6.

3) REWARD (r_i)

As mentioned before, AP i receives an immediate reward depending on the action taken in a particular state. For each $i \in \mathcal{N}$ and $j \in \mathcal{M}$ at time iteration t, the CU decides actions $a_{ij}(t) \in \mathcal{A}_{ij}$ for the $i-j$ link after observing the state $s_{ij}(t)$. The CU communicates $\alpha_{ij}(t)$ through a backhaul link to AP i for all $j \in \mathcal{M} \backslash \{\alpha_{ij} = 0\}$. In the explanation ahead, the subscripts i and j in s_{ij}, a_{ij} and $\mathcal{A}_{ij}$ have been dropped for simplicity. The immediate reward $R_i(s, a)$ is received in the form of the data rate of the AP i and is defined as

$$R_i(s,a) = \begin{cases} r_{\text{fix}}, & \text{if } \sum_{c=1}^{6} I_c^i > 0, \\ r_i, & \text{otherwise}, \end{cases} \tag{34}$$

where r_{fix} is a reward smaller than the reward obtained after applying any action violating the interference constraints. When the constraints are satisfied, the reward received by AP i is r_i. The CU will seek to find an optimal policy for each AP to maximize it's own r_i. The CU repeatedly makes the decision and finally obtains the optimal policies for the APs to maximize their respective r_is for constraints (21) to (29). Since, r_is are always non-negative, maximization of $\sum_{i \in \mathcal{N}} r_i$ can be achieved by maximizing individual r_i for each AP i. Therefore, the CU will seek to find an optimal policy through DQN learning algorithm to maximize the reward for AP i.

The action vector, state vector, and reward have been used for performing DQN learning as shown in Fig. 2. The CU is shown to be equipped with a replay memory to store the experience $e_i(t) = \{a_{ij}(t), s_{ij}(t), r_i(t), s_{ij}(t+1)\}$, which was gathered at the transition of two consecutive time instants t and $t+1$. The replay memory gets $s_{ij}(t)$, $r_i(t)$, and $s_{ij}(t+1)$ from the network and $a_{ij}(t)$ from the DQN learning output. A mini batch is present which takes training samples from the replay memory at each iteration. Each iteration consists of fixed number of episodes EP_N such that each episode uses one training sample and runs for T time slots as shown in Algorithm 1. Further, a DQN block is shown where DQN learning is performed. The input switch of the DQN block switches it's connection alternately with the output of the mini batch and with a link to the network. When connected with the output of the mini batch, it receives the training samples while when connected with the link to the network, it gathers knowledge about the state $s_{ij}(t)$. The DQN learning output is produced in the form of the selected action $a_{ij}(t)$. The output port of the DQN block switches it's connection alternately between two input ports ahead. The first input port feeds $a_{ij}(t)$ to the replay memory. The second input port feeds $a_{ij}(t)$ to the loss and gradient and parameter upgrading blocks, where the upgraded θ is obtained. The output of the parameter upgrading block is fedback to the input of the DQN block with the mini batch ouput.

To accomplish the DQN based learning algorithm for AP i, the CU finds an optimal policy π for it with the help of state-value function $V^\pi(s)$ [43]. It is the maximum discounted sum of immediate rewards $R_i(s, a)$ over a long span of time while the optimal policy π is being followed. Mathematically, it is written as

$$V^\pi(s,a) = \max_\pi \{\sum_{t=0}^{\infty} \zeta^t E(R(s,a))_t | s_t = s, a_t = a, \pi\}. \tag{35}$$

The optimal action-value function $Q^*(s,a) \triangleq \max_{\pi} V^{\pi}(s,a)$ is obtained with the help of Bellman's equation as

$$Q^*(s,a) = \max_{a \in \mathcal{A}} \{r(s,a) + \zeta Q^*(s',a')\}. \quad (36)$$

where ζ is the learning rate at which $Q^*(s,a)$ is updated. In (35), $Q^*(s,a)$ iteratively converges to it's optimal value for $t \to \infty$.

The maximization of $Q(s,a)$ leads to the maximization of r_i as the objective of DQN learning is to define an environment for the agent to perform certain actions to maximize the reward. In this work, the reward is the achievable data rate of the ith AP, r_i. First, a state value function $V^{\pi}(s)$ is calculated. The state value function $V^{\pi}(s)$ tells which state gives the highest reward, i.e., the achievable data rate r_i, and is given as where

$$R_i(s,a) = \begin{cases} r_{\text{fix}}, & \text{if} \sum_{c=1}^{6} I_c^i > 0, \\ r_i, & \text{otherwise}, \end{cases} \quad (37)$$

The next step is the calculation of the action-value function $Q(s,a)$, which signifies the action or the policy that the agent should take so that the maximum state value is achieved. Mathematically, $Q(s,a) = \max_{\pi} V^{\pi}(s)$. Thus, maximizing the action - value function leads to the maximization of the reward r_i.

If vectors are large, obtaining optimal $Q^*(s,a)$ becomes challenging. Thus, the optimal action-value function is estimated with the help of a function estimator. In this regard, [43] has been followed, where a neural network for this estimation as $Q(s,a;\theta) \approx Q^*(s,a)$ has been proposed. In this article, a fully connected feed-forward multilayer perception (MLP) network is used for this approximation. Since it is a neural network acting as the action-value approximator, it also brings advantage to the DQN based algorithm. In this approximation, it includes *experience replay* to improve the performance of learning, in which the CU stores the experience of the environment at each time step for AP i as $e_i(t) = \{a_{ij}(t), s_{ij}(t), r_i(t), s_{ij}(t+1)\}$ into a replay memory. The replay memory at different time instants is written as $D_i(t) = \{e_i(1), \ldots, e_i(t)\}$. The two different MLP networks used as Q-network approximators are action-value function approximator $Q(s,a)$ and the target action-value function approximator $Q(s,a;\theta)$. Here, θ and θ^- are the parameters of the present and previous iterations respectively. With each iteration, the present iteration parameter θ of the action-state function is updated. This is done with the help of the display memory D_i where a random sample $(a, s, r, \hat{s})$ is chosen. The updation of θ^- is done after a fixed number of iterations, where the parameters of the target value function are replaced with the updated θ of the action value function. The update procedure is done with the help of gradient descent algorithm based on the following cost function:

$$L(\theta_i) = E\left[\left(r_i(s,a) + \zeta \max_{\hat{a} \in A} \left(\widehat{Q}_i(\hat{s},\hat{a},\theta_i^-)\right) - Q_i(s,a,\theta_i)\right)^2\right]. \quad (38)$$

Algorithm 1 Achievable Data Rate Maximization in Hybrid RF/VLC Systems

for $i = 0, 1, 2, \ldots, |\mathcal{N}|$ **do**
  Initialization
  Initialize the replay memory
  Initialize the policy $\pi(a_{ij}|s_{ij};\theta_i)$ parameter θ_i
  Initialize the neural network for action-value function Q_i with random weights θ_i
  Initialize the neural network target action-value function $\widehat{Q}_i$ with $\theta_i^- = \theta_i$
end for
for Iter =1:K **do**
  Receive the initial state
  for Episode = 1: EP_N **do**
    for $t < T$ **do**
      for $i = 0, 1, 2, \ldots, |\mathcal{N}|$ **do**
        Chose $a_{ij}^*(t)$ as per the maximizing equation for $j \in \mathcal{M}$
        Select an action

$$a_{ij}(t) = \arg\max_{a_{ij}(t)} Q(s_{ij}(t), a_{ij}(t); \theta_i) \quad (39)$$

        Otherwise select a random action with probability ϵ
        Update the state $s_{ij}(t+1)$ and the reward $r_i(t)$ according to (33) and (37)
        Store $e_i(t) = (a_{ij}(t), s_{ij}(t), r_i(t), s_{ij}(t+1)$ in the experience replay memory created for AP i, D_i.
        Update the current parameters θ_i of the action-value function $Q(s_{ij}(t), a_{ij}(t); \theta_i)$, by sampling mini-batch of transitions from $D_i(t)$
        After every fixed number of steps, update $\theta_i^- = \theta_i$
        Get mini batch samples from the replay memory
      end for
    end for
  end for
end for
Perform $r = \sum_{i \in \mathcal{N}} r_i$
As the non-negative r_i of each AP is optimized and the sum-rate r is the sum of r_is, it will lead to the optimization of the overall system

The DQN based learning algorithm for maximizing the achievable sum-rate of the hybrid RF/VLC system is given in Algorithm 1. The above application of DQN learning to solve a resource allocation problem is expected to prove efficient as the considered hybrid RF/VLC system involves large state and action vector spaces. In this regard, DQN learning takes advantage of neural networks to train the learning process and efficiently maximize the Q-value by approximating the action-value function from the current state. With such an application, a higher convergence speed of the algorithm and a better output achievable sum-rate are expected. Moreover, the solution has been achieved without complete and accurate network information. It is clear that first each r_i for the ith AP

is optimized. As r_is are non-negative and the overall sum-rate r is their sum, it will lead to the overall system optimization.

V. A NEWLY ENTERING UE

To investigate a dynamic system, a new UE entering into the scenario has been considered. Note that the DQN based learning algorihthm estimates the new Q-function on the basis of the reward of every action for each AP. The CU learns the environment of each AP respectively. Then it takes an action linked with the highest reward, which means performing the association of the UEs with each AP and allocating bandwidth and power to each AP, in a manner which gives the highest reward. Thus, the AP gets the reward pertaining to the action taken by the DQN based learning algorithm. The parameters of the Q-function are updated as per the reward received immediately. In other words, these parameters reflect the effects brought by the action parameter of each AP. Each AP causes interference to the other UEs in the hybrid RF/VLC environment. The Q-function parameters reflect the local environment of each AP and also an overall interrelationship between the different modules of the hybrid RF/VLC system.

In case when a new UE joins the environment, discarding all the already gathered information for the individual APs at the CU, the interconnection of the modules in the environment, and initiating the algorithm again for the new system will be an inefficient procedure. We propose to try the application of the *transfer learning* phenomena in such situation [61]. The already gathered information about the environment obtained through the Algorithm 1 before the new entrant UE has entered will be used immediately after it enters the environment. As the cognitive cycle proceeds further, the information will be updated according to Algorithm 2.

Algorithm 2 Transfer Learning for a Newly Entering UE

(Run as a new UE joins the network)
Add the new UE with index $|\mathcal{M}|+1$
Initialize Q for AP i with parameters of the action-value function pertaining to the UE nearest to the new-comer $|\mathcal{M}|+1$th UE {The information for the UE nearest to the $|\mathcal{M}|+1$th UE is used by the CU when it enters the scenario (transfer learning)}
for $i = 0, 1, 2, \ldots, |\mathcal{N}|$ **do**
  Algorithm 1 is started with the existing action value functions for $|\mathcal{M}|+1$ UEs and then proceeded iteratively.
end for
Perform $r = \sum_{i\in\mathcal{N}} r_i$

VI. SIMULATION RESULTS

In this section, the effectiveness of our proposed algorithms has been verified with the help of simulations.

A. PERFORMANCE ANALYSIS OF THE PROPOSED ALGORITHMS

Initially, the following set-up has been considered: The hybrid RF/VLC network consists of 1 RF AP, 4 VLC APs, and 4 UEs. At a given time instant, each AP can serve multiple UEs, while one UE can receive data from only one AP. The values for the parameters has been decided from [18] for performing the simulations. The VLC AP noise N_0^v is $10^{-21}\text{A}^2/\text{Hz}$, the average optical power per VLC AP (LED lamp) is 9.2 W, physical area of the PD A_{pd} is 1 cm^2, PD responsivity ρ_j for all $j \in \mathcal{M}$ UEs is 0.28 A/W, receiver FOV is 60°, half angle of the LED $\phi_{1/2}$ is taken as 70°, and the maximum illuminous intensity of the LED is 28 cd. A learning rate ζ of 0.01 and discount factor of 0.9 are used for all the APs. The path-loss exponent pl_{0j} is taken as 2.8. For designing the action vector $\mathcal{A}_{ij}$, $B^v_{\max}$ = 20 MHz, $B^v_{\min}$ = 12 MHz, $B^r_{\max}$ = 10 MHz, $B^r_{\min}$ = 5 MHz $P^v_{\max}$= 195 mW, and $P^v_{\min}$ = 180 mW, $P^r_{\max}$ = 15 dBm, and $P^r_{\min}$ = 8 dBm have been considered. The order of Lambertian constant m is taken as 1.2, the room dimensions are taken as 10 m ×10 m ×7 m, the height of the UE is considered as 0.9 m from the floor. The RF AP is placed at the center and the VLC APs are placed at the positions $\left[\pm\frac{10}{\sqrt{8}}, \pm\frac{10}{\sqrt{8}}\right]$ on the room ceiling. The replay memory capacity is considered as 100 and the mini-batch for buffer is kept at a size of 10 respectively. The investigations have been performed over 1000 monte-carlo simulations. The input to the neural network has 7 nodes: 6 nodes for the state and 1 node for the selected action to be taken. The structure of DQN consists two-hidden layers of fully-connected neural network with 3 and 2 neurons, respectively. The state and action vectors are functions of downlink bandwidth, power, and association parameter. Thus, passing state and action vectors through the input of the neural network means passing the downlink bandwidth, power and association parameter. The neural network trains the DQN-learning algorithm for generating action-value approximator with environmental interaction and receive the maximum reward. As the iterations proceed, the algorithm converges towards the optimal policy selection from the action vector $\mathcal{A}_{ij}$ in (32), which is choosing optimal $\mathbb{B}_{ij}$, $\mathbb{P}_i$ and α_{ij}, and the achievable sum-rate is maximized.

In the evaluations, the outcome of the proposed schemes has been compared with the exhaustive search algorithm, the received SINR based and the received power based association schemes as benchmarks which are popular resource allocation techniques in heterogeneous networks. We have also made the comparison of the proposed schemes with the Q-learning based power allocation scheme for hybrid RF/VLC proposed in [71]. First, an explanation on the received SINR and receive power based resource allocation schemes is provided. These schemes have been widely used in general heterogeneous networks.

1) RECEIVED SINR BASED SCHEME

The fundamental work in this area can be found in [10]. Further it has been followed in [72]. The fundamental problem addressed in [10] is the optimal allocation of association parameter for equal resources alloted to all the APs. The optimal association parameter association is aimed for achievable

data rate maximization of a single AP i UE j link. The problem for obtaining optimal association parameter is given as

$$\max_{\alpha} \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{M}} \alpha_{ij} \log(\frac{R_{ij}}{\text{Load}_i})$$
$$\text{st} \sum_{i=1}^{N} \alpha_{ij} = 1 \text{for} j \in \mathcal{M}$$
$$\alpha_{ij} \in \{0, 1\}, i \in \mathcal{N}, j \in \mathcal{M} \quad (40)$$

where, $\text{Load}_i = \sum_{j'=1}^{M} \alpha_{ij'}$ is the total load on the ith AP, which means that Load_i represents the number of UEs associated with the ith AP. As equal resource allocation has been followed at each of the $|\mathcal{N}|$ APs, the data rate R_{ij} will be equally divided among all the UEs associated with the ith AP. The authors propose a highest SINR based algorithm for solving this problem. The algorithm is based on the SINR which a UE has with each of the $|\mathcal{N}|$ APs. To formulate the algorithm, the problem in (40) is re-written in terms of Lagrange multiplier as

$$\max_{\alpha} D(\mu) = f_{\alpha}(\mu) + g_K(\mu) \quad (41)$$

$$f_{\alpha}(\mu) = \begin{cases} \max_{\alpha} \alpha_{ij}(\log(R_{ij}) - \mu_i) \\ \text{st} \sum_{i=1}^{N} \alpha_{ij} = 1 \\ \alpha_{ij} \in \{0, 1\} \end{cases} \quad (42)$$

$$g_K(\mu) = \max_{\text{Load}_i < \mathcal{M}} \sum_{j} \sum_{i=1}^{N} \text{Load}_i(\mu_i - \log(\text{Load}_i)) \quad (43)$$

The proposed algorithm is aimed to solve the problem (41) is as follows:
UE's algorithm:

- Each UE measures the SINR by using the pilot signals from all the APs, and receives the value of μ_i broadcast by each AP at the beginning of the iteration.
- UE j determines the AP $i*$ which satisfies the follows:

$$i* = \arg\max_{i}(\log(R_{ij}) - \mu_i(t)) \quad (44)$$

  If there are multiple maximizers, the UE will chose one of them.

AP's algorithm: Each AP updates the new value of Load_i and μ_i in the two steps and announces the new multiplier μ_i to the system.

- To obtain the maximizer of problem in (43), we set it's gradient to be 0 with the constraint $\text{Load}_i \leq |\mathcal{N}|$ i.e.,

$$\text{Load}_i(t+1) = \min\{|\mathcal{N}|, \exp(\mu_i(t) - 1)\} \quad (45)$$

- The new value of the Lagrange multiplier is updated by

$$\mu_i(t+1) = \mu_i(t) - \delta(t).\text{Load}_i(t) - \sum_{i} \alpha_{ij}(t) \quad (46)$$

where $\delta(t)$ is a dynamically chosen stepsize sequence based on some suitable estimates.

2) RECEIVED POWER BASED ASSOCIATION SCHEME

The next comparison of the proposed DQN-learning based resource allocation scheme has been made with received power based association technique. The most significant received power based association technique has been shown by Lin *et al.* in [73]. The association of a UE is decided according to the signal power it receives from different APs. A UE will be associated with an AP if it provides signals at the highest power. Suppose UE j is at a position y_j, VLC AP i at a position $x_{v,i}$, $i \in \mathcal{N}$ in a hybrid RF/VLC system. If the position of the RF AP is x_0, a UE will be associated with the VLC AP if

$$\min_{i} \left(P_i(m+1)A_{pd} \cos^m \phi_{ij} \cos \psi_{ij} T_{\text{opt}}(\psi_{ij}) g(\psi_{ij}) \right.$$
$$\left. \left(|x_i - y_j|\right)^{-2}\right) \geq P_0 L \chi_{0j} |x_0 - y_j|\Big)^{-pl_{0j}} \quad (47)$$

The above condition is also based on the received power at the UEs from the APs. The channel losses in the RF and VLC mediums has also been taken into consideration.

3) EXHAUSTIVE SEARCH METHOD

The third benchmark considered for investigating the efficiency of our proposed schemes is the Exhaustive search method [74]. This method is highly complex. The Exhaustive search method used here involves a trellis based mechanism. For instance, let us imagine the optimization of the parameters α_{ij}, B_{ij}, and P_i, which lead to the calculation of the action variable γ_{ij} as a traverse between it's initial random value and it's final optimal value. This process involves forming a trellis between the two points. The trellis consists of a certain number of levels, with the final level having the optimal value of γ_{ij}. Each level consists of a number of possible values for γ_{ij}. The main objective here is it to determine all the possible paths from the initial random value to the final optimal value. It involves working through the trellis from level 1 to the final level which involves calculating the number of paths at each level. Let $\mathtt{R}$ be the set of trellis levels, then there will be $|\mathtt{R}|$ trellis levels. Each level has M points where optimum value could be obtained. If $Q(r_l, m)$ be the number of paths at the point m of the level r_l, where $1 \leq m \leq M$ possible from the level 1, as shown in Fig. 3, the calculation of the total number of paths possible will be $\sum_{m=1}^{M} Q(r, m)$.

4) Q-LEARNING BASED POWER ALLOCATION SCHEME IN HYBRID RF/VLC

We compare our proposed schemes with the state-of-the-art multi agent Q-learning based power allocation in hybrid RF/VLC systems proposed Kong *et al.* in [71]. Kong *et al.* have used multi agent Q-learning for optimization of the transmit power of the RF and VLC APs. Being multi-agent Q-learning, it is performed at each AP separately. On the basis of the application of Q-learning, each AP decides it's transmit power. We compare our results with [71] as it is the state-of-the-art work available on this topic.

FIGURE 3. Exhaustive Search Mechanism.

The proposed DQN learning based resource allocation algorithm is different from the work in [71] in several aspects. The work in [71] deals only with transmit power allocation for the APs. On the other hand, the proposed DQN learning based resource allocation algorithm deals with transmit power allocation for the APs, the bandwidth allocation for the APs and deciding the association of the UEs with the APs. It can be seen that the domain of the problem addressed here is much larger. The work in [71] consists of only two constraints on the transmit powers of the RF and the VLC APs. However, our work considers six constraints which consist of the two transmit power constraints, two constraints on the bandwidths of the RF and VLC APs, one-one constraint on the association parameter and the SINR each. Thus, our problem formulation is more practical. Considering constraints only on transmit power of the APs leads to presumptions on bandwidth and association parameters, which may compromise with the practicality of the system.

This process of comparing the schemes proposed in [71] with our schemes is accomplished by implementing the scheme proposed in [71] for our system and then comparing them with our results (shown ahead in Fig. 10). In [71], transmit power of the APs is the optimization variable and is optimized with Q-learning. Thus, to implement [71] in our work, the action vector in expression (32) is formulated with only power $\mathbb{P}_i$ terms as variables and presumptions are made for bandwidth $\mathbb{B}_{ij}$ and association α_{ij} parameters. The problem (20)-(27) is reduced to

$$\mathcal{P}: \max_{B_{ij},P_i,\alpha_{ij}} r_i, \qquad \text{for } i \in \mathcal{N}, j \in \mathcal{M}, \tag{48}$$

such that

A constraint is imposed on the transmission power to ensure the power budget saving and safety considerations for the eyes. The transmission power of a VLC AP cannot exceed its maximum available power $P_{\max}^{\text{VLC}}$, formulated as

$$\mathcal{C}_1 : 0 \leq P_i \leq P_{\max}^{\text{VLC}}, \qquad \text{for } i \in \mathcal{N}\backslash\{0\}, \tag{49}$$

Similarly, the transmission power of an RF AP cannot exceed its maximum available power $P_{\max}^{\text{RF}}$ formulated as:

$$\mathcal{C}_2 : 0 \leq P_0 \leq P_{\max}^{\text{RF}}, \qquad \text{for } i = 0, \tag{50}$$

Further, the optimization of action vector $\mathcal{A}_{ij}$ is carried out with Q-learning. For the allocation of bandwidth, equal allocation is considered for all the APs, while association parameter α_{ij} is allocated as per the minimum distance criteria. For power allocation, the investigation is performed with two cases, when the number of UEs is fixed and when a new UE is entering into the system. When the number of UEs is fixed, DQN learning without transfer learning serves the purpose while when a new UE is entering into the scenario, the application of transfer learning is investigated. The comparisons have been shown ahead in Fig. 10. The achievable sum-rate is compared with the increasing number of VLC APs deployed as shown in Fig. 9.

We now present the simulation results. In Fig. 4, the number of iterations needed for the maximization of the normalized achievable sum-rate with the application of the proposed algorithms has been studied. The investigation for the fixed number of UEs is done in Fig. 4a and in Fig. 4b, the investigation for the case of a new incoming UE is made. As mentioned above, $|\mathcal{N}|$ and $|\mathcal{M}|$ are conisdered as 5 and 4 respectively. In Fig. 4a, the DQN learning mechanism starts showing output achievable sum-rate of 380 Mbits/s at nearly 240 iterations which goes on increasing with minor fluctuations as the iterations are increased. In nearly 1600 iterations, the final value of the maximized achievable sum-rate is obtained as 1270 Mbits/s. On the other hand, the exhaustive search algorithm starts showing output at nearly 250 iterations and shows a final achievable sum-rate value of 1140 Mbits/s in nearly 1600 iterations. The final values of the achievable sum-rate obtained with the received power and the SINR based association schemes are 850 and 990 Mbits/s respectively, which shows that the proposed DQN based learning based mechanism outperforms exhaustive search, received power based association and received SINR based association, and leads to at least 10% increase in the achievable sum-rate.

Fig. 4b shows the performance of the transfer algorithm (Algorithm 2) for the case of a newly entering UE, which has been labelled as *DQN-transfer learning*. Fig. 4b also shows the performance of DQN-learning based method, exhaustive search, received SINR based and received power based algorithms for this case. For applying DQN-transfer learning, the CU uses the information of the already learned network for the newly joined UE while for applying DQN learning based method, the CU initiates action-value function parameters randomly for the newly joined UE. Similarly, the exhaustive search method, the receive SINR and the receive power based algorithms re-start from the beginning after the arrival of the new UE and operate till convergence. It can be seen that the DQN-transfer learning converges to a final value of nearly 1290 Mbps in just 1200 iterations, while the DQN learning based mechanism converges to it in 2600 iterations. The exhaustive search, received SINR and received power based mechanisms converge to the final values attained in Fig. 4a, but in 2650, 3200 and 2700 iterations respectively. Note that the high number of iterations needed by the received SINR and power based association schemes arises due to

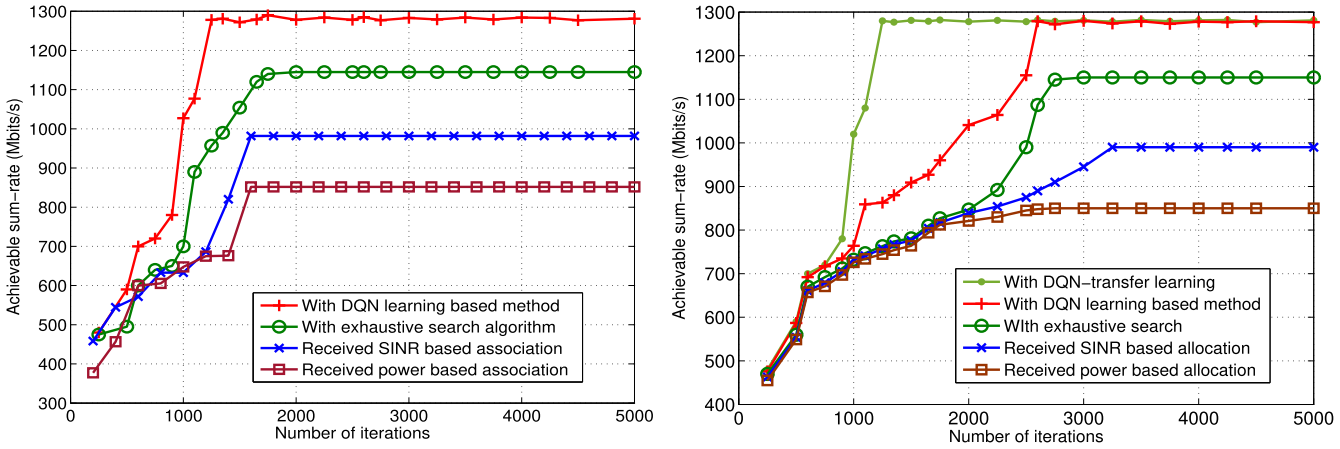

FIGURE 4. Graph depicting the convergence of the proposed algorithms by showing achievable sum-rate vs the number of iterations when (a) a fixed number of UEs are present in the room. (The performance of Algorithm 1 compared with the existing algorithms.) (b) a new incoming UE enters the room. (The performances of Algorithm 1 and Algorithm 2 (DQN with transfer learning) are compared with the existing algorithms).

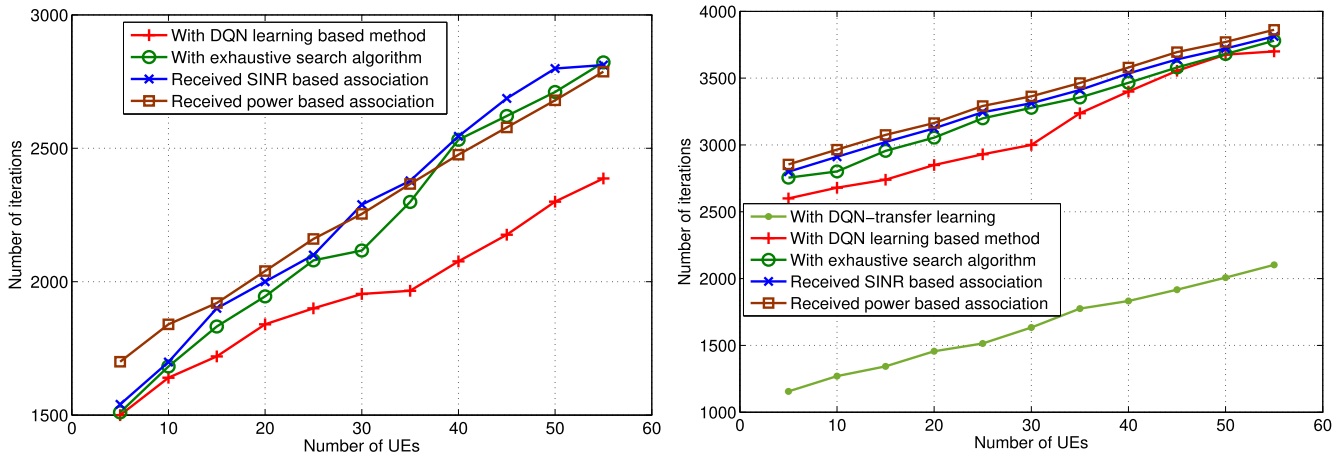

FIGURE 5. Graph depicting the behavior for number of iterations with the umber of UEs when (a) a fixed number of UEs are present in the room. (The performance of Algorithm 1 compared with the existing algorithms.) (b) a new incoming UE enters the room. (The performances of Algorithm 1 and Algorithm 2 (DQN with transfer learning) are compared with the existing algorithms).

load balancing and proportional fairness issues. Unlike the proposed DQN based approaches, these algorithms use the total sum-rate as the objective function.

In Fig. 5, the average number of iterations needed for convergence with varying number of UEs is shown. Fig. 5a shows that as the number of UEs increases, the number of iterations needed for the convergence of all the algorithms increases. It can be seen that the DQN based learning algorithm can attain a level of achievable sum-rate is much lesser number of iterations compared to the other algorithms for a given number of UEs in the network. For a network with 55 UEs, at least 10% higher achievable sum-rate can be attained by the DQN-learning based algorithm in 14.28% lesser iterations compared to the achievable sum-rate value attained by exhaustive algorithm. Further, Fig. 5b shows that when a new UE enters the network, the DQN-transfer learning achieves its maximum achievable sum-rate in nearly 54% lesser number of iterations for a given number of UEs present in the set-up.

Fig. 6 shows the plot for achievable data rate with the number of UEs. Fig. 6a shows the results for a fixed set up while Fig. 6b shows the results for the case of the arrival of the new UE. In both the figures, the number of UEs are varied from 5 to 55. It can be seen that for this entire range of the number of UEs, the DQN based learning algorithm and DQN-transfer learning outperform the exhaustive algorithm, the received power and the received SINR based association by reasonable margins. On increasing the number of UEs, the achievable sum-rate increases with the increase in the number of AP-UE links. The achievable sum-rate is found to increase at a higher rate in the 5 to 15 UEs range. As the number of UEs is increased from 15 to 25, a slight decrease in the rate of increment can be seen. For all further increments in the number of UEs till 45, a slight decrement in

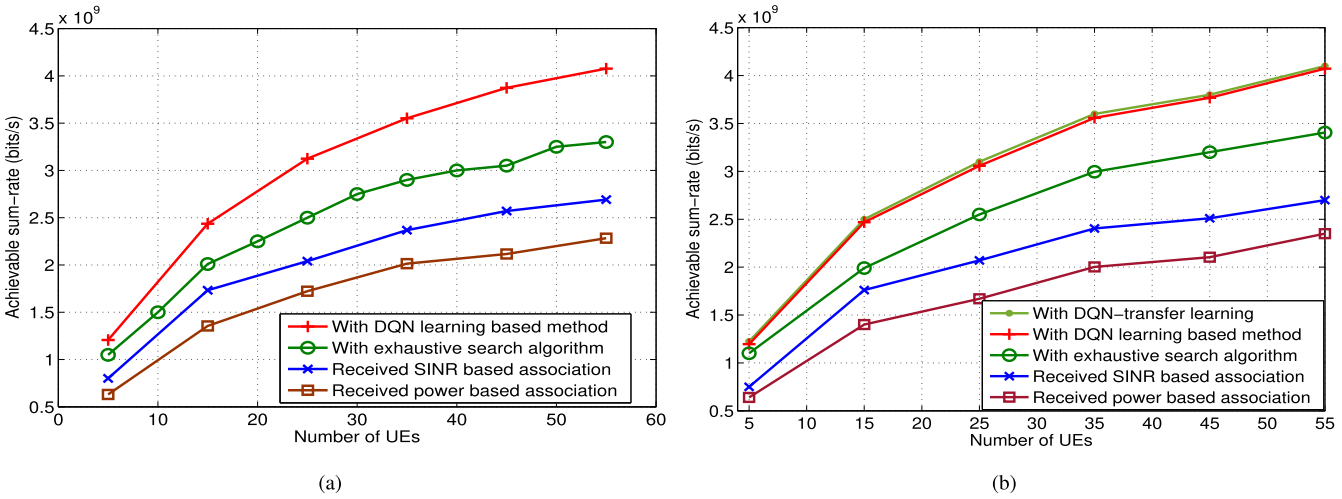

FIGURE 6. Graph depicting the behavior for achievable sum-rate vs the number of UEs when (a) a fixed number of UEs are present in the room. (The performance of Algorithm 1 compared with the existing algorithms.) (b) a new incoming UE enters the room. (The performances of Algorithm 1 and Algorithm 2 (DQN with transfer learning) are compared with the existing algorithms).

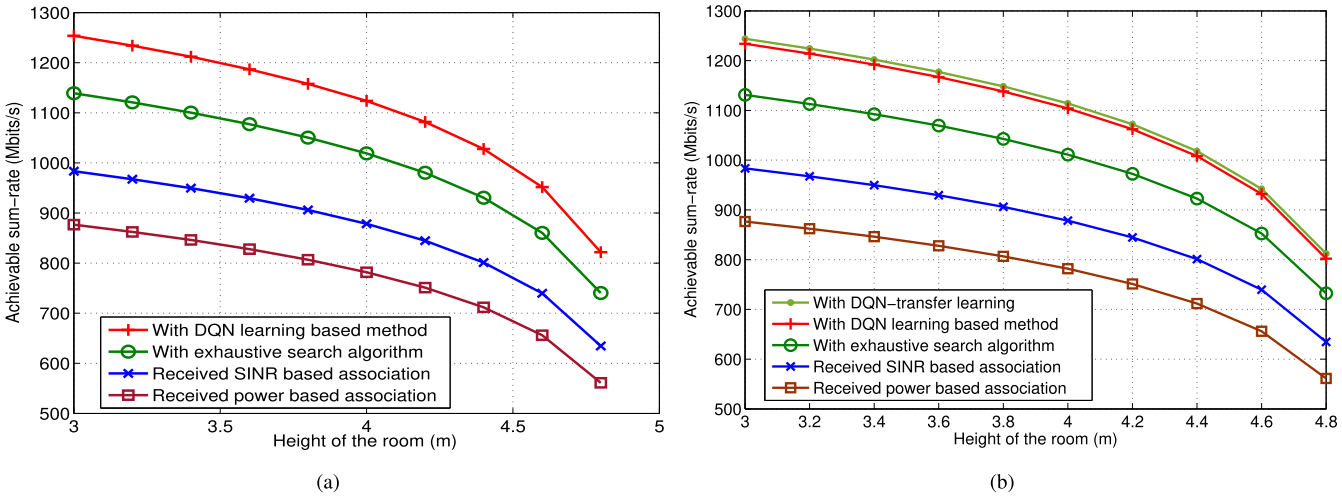

FIGURE 7. Graph depicting the behavior for achievable sum-rate vs the height of the room when (a) a fixed number of UEs are present in the room. (The performance of Algorithm 1 compared with the existing algorithms.) (b) a new incoming UE enters the room. (The performances of Algorithm 1 and Algorithm 2 (DQN with transfer learning) are compared with the existing algorithms).

the respective rates can be seen. This behavior is the same in all the algorithms investigated here. Intuitively, it is due to the fact that an increase in the number of AP-UE links also results in increased interferences. However, when the number of UEs is further increased from 45 to 55, the rate of increment again increases, which shows that for a high number of UEs, the desired signal power component becomes dominant. In Fig. 6b, the DQN-learning algorithm gives nearly the same output as the DQN-transfer algorithm. The difference between their applications is only the number of iterations needed to converge to their final outputs, as shown in Figs. 4 and 5.

In Fig. 7, the effectiveness of the proposed algorithms is investigated for a varying height of the room. As the height of the room increases, the transmitter-receiver separation increases which results into a decrease in the achievable sum-rate. This decrease is evident from (4), (8), and (11), where it is shown that the channel gains for RF and VLC networks decrease in magnitude with the increase in the transmitter-receiver separation. As the height of the room is increased, the attenuation in the signals received by the UEs increases. Similar to the previous figures, Fig. 7a shows the investigation for the fixed UEs case while Fig. 7b shows the investigation for the newly entering UE. It can be seen that the DQN learning based algorithm and DQN-transfer algorithm outperform the algorithms under consideration.

Fig. 8 shows investigations on the FOV of UE j. The FOV impacts the VLC system performance significantly. The achievable sum-rate obtained with the different schemes under consideration has been plotted over a wide range of FOV from 60° to 180°. When the FOV of the receiver is small, the effect of interfering signals is lesser on it. Thus, it gives a higher achievable sum-rate. Contrarily, when the FOV of the receiver is large, it receives more interfering

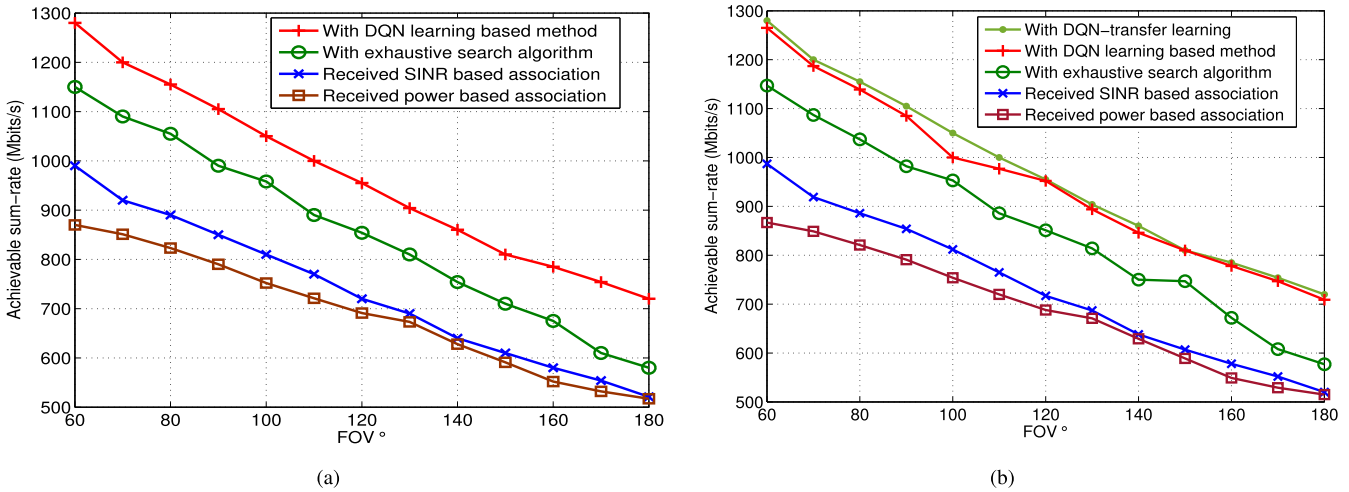

FIGURE 8. Plot showing the variation of achievable sum-rate vs the FOV of the UE j when (a) a fixed number of UEs are present in the room. (The performance of Algorithm 1 compared with the existing algorithms.) (b) a new incoming UE enters the room. (The performances of Algorithm 1 and Algorithm 2 (DQN with transfer learning) are compared with the existing algorithms).

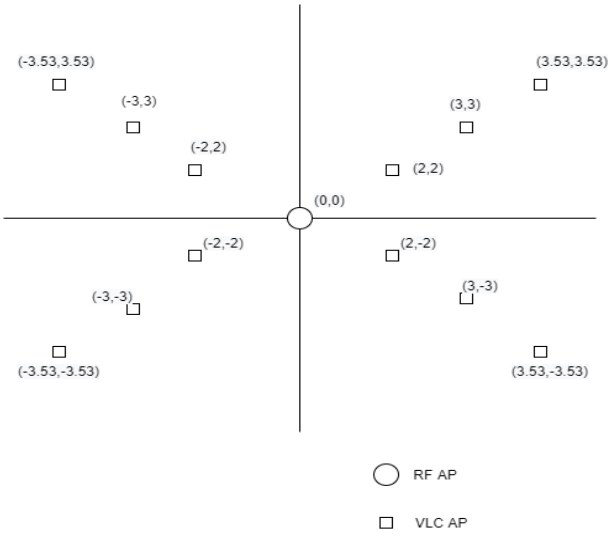

FIGURE 9. Deployment architecture of new APs.

signals from the unassociated APs. Thus, the interference increases which results into the decrement of the achievable rate. This behavior has been depicted in Fig. 8a and Fig. 8b. It can be seen that the proposed DQN learning and DQN-transfer learning based methods outperform the other schemes under consideration. It can also be concluded that for a fixed deployment of VLC APs on the corners of the room, a sharp decrease occurs in the achievable sum-rate with the increasing FOV of the receiver. Such behavior may change for a different deployment of the APs.

The results presented so far do not consider the case of dense AP deployment. To address this concern, we perform simulations for higher number of APs, as shown in Fig. 9. The deployment of the new APs is done as it was done earlier for 4 APs. The 4 APs which were deployed earlier are positioned on the same coordinates in the four corners of the room as before. The new APs are placed within the area covered by these 4 APs as shown in the Fig. 9. The coordinates of the new APs has also been shown here. Next in Fig. 10, the achievable sum-rate vs. the number of APs has been plotted for this set-up. Fig. 10a shows the performance of DQN-learning mechanism, while Fig. 10b shows the performance of the transfer learning algorithm. From Fig. 10, it can be seen that the DQN-learning and the DQN transfer learning algorithms outperform the existing algorithms for the static and the dynamic cases.

B. COMPLEXITY ANALYSIS OF THE PROPOSED DQN-LEARNING BASED ALGORITHMS

The objective of this work is the maximization of action value function $Q(s, a)$ which is achieved by bringing $Q(s, a)$ as close to the target action-value function $\widehat{Q}(s, a, \phi)$. The algorithmic complexity is the sum of the statistical and the algorithmic error in this process [75]. The total error rate is given by

$$||\widehat{Q} - Q^k|| \leq C.\frac{\phi_{\mu,\sigma}}{(a-\iota^2)}.|\mathcal{A}| .(\log n)^{1+2\zeta}.n^{(\alpha*-1)/2} + \frac{4\iota^{K+1}}{(1-\iota)^2}.R_{i\max} \quad (51)$$

where Q^k is the Q term at the kth iteration, μ and σ are the mean and standard deviation of $\mathcal{P}(\mathcal{S} \times \mathcal{A})$ where $\mathcal{P}$ denotes distribution, $\phi_{\mu,\sigma}$ is a constant such that $(1-\iota)^2 \sum_{v\geq 1} \iota^{v-1} v.K \leq \phi_{\mu,\sigma}$, n is the sample size, ζ is a constant, $R_{i\max}$ is the maximum reward value for the ith AP. The first term on the right hand side (RHS) of the equation is the statistical error while the second term on the RHS of the equation is the algorithmic error. The algorithmic error converges to zero in linear rate as the algorithm proceeds, but the statistical error represents the fundamental problem. When the following condition for the number of iterations K is satisfied, the statistical error dominates the algorithmic

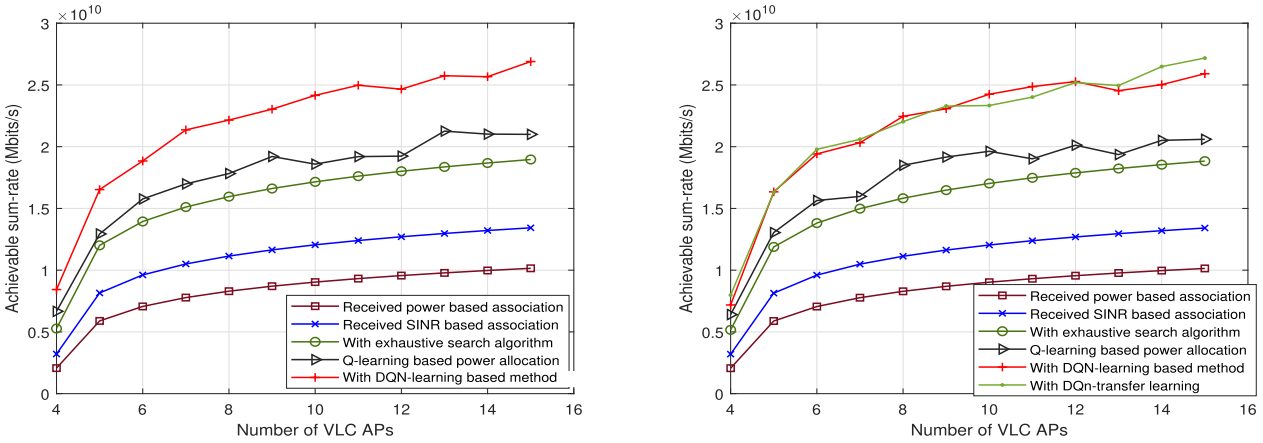

FIGURE 10. (a) Performance comparison with Q-learning based power allocation in hybrid RF/VLC. (b) Performance comparison with Q-learning based power allocation in hybrid RF/VLC for a newly entering UE.

error

$$K \geq [\log \mathcal{A} + (1-\alpha*).\log n]/\log(1/\iota) \tag{52}$$

Viewing ι and $\phi_{\mu,\sigma}$ as constants and ignoring the polygarithmic term, the proposed algorithms achieve the error rate

$$|\mathcal{A}|.n^{(\alpha*-1)/2} = |\mathcal{A}|.\max_{j\in|q|} n^{\beta_j^*/(2\beta_j^*+t_j)} \tag{53}$$

which scales linearly with the capacity of the action space and goes to zero when n goes to ∞. Here, t_j and β_j are time parameters for the jth UE. The term $n^{\beta_j^*/(2\beta_j^*+t_j)}$ in the above equation recovers the statistical rate of the non-parametric regression in l_2-norm. It is further found that the algorithm achieves an error rate of $|\mathcal{A}|.n^{-\beta_j/(2\beta_j+r)}$ when K is sufficiently large, where $r \in \mathbb{N}$, $\mathbb{N}$ denotes a natural number.

Note that π is the greedy policy with respect to $\widehat{Q}$ and Q functions. As the construction of Q is done with an iterative algorithm, the error convergence has to be related to the error in the previous steps, i.e., $\widehat{Q}_k - \widehat{Q}_{k-1}$. This relation is formulated as

$$||\widehat{Q} - Q^k|| \leq \frac{2\phi_{\mu,\sigma}.\iota}{(1-\iota)^2}.\max_{k\in[K]} ||\widehat{Q}_k - \widehat{Q}_{k-1}||_\sigma + \frac{4\iota^{K+1}}{(1-\iota)^2}.R_{i\max} \tag{54}$$

where $\phi_{\mu,\sigma}$ is a constant that depends only on the distributions of μ and σ.

Thus, as mentioned above, the total error is the sum of algorithmic and statistical errors, where $\max_{k\in[K]} ||\widehat{Q}_k - \widehat{Q}_{k-1}||_\sigma$ is the statistical error and the second term on the RHS of the equation is the algorithmic error. The statistical error goes to zero as n increases to a large number. The algorithmic error goes to zero as the number of iterations K increases. The fundamental difficulty of DQN is the error incurred in the single step. The bound on $||\widehat{Q}_k - \widehat{Q}_{k-1}||_\sigma$ is obtained as

$$||\widehat{Q}_k - \widehat{Q}_{k-1}||_\sigma^2 \leq 4.[\text{dist}_\infty(\mathcal{F}_0, \mathcal{G}_0)]^2 + C.V_{\max}^2/n.\log N_\delta + C.V_{\max}.\delta \tag{55}$$

where $V_{\max} = R_{\max}/1-\iota$ $\mathcal{F}_0 = \{f : \mathcal{S}\times\mathcal{A} \to \mathbb{R} : f \in \mathcal{F}\}$, $\mathcal{G}_0 = \{f : \mathcal{S}\times\mathcal{A} \to \mathbb{R} : f \in \mathcal{G}\}$, $\mathcal{F}$ is the family of DQN defined on the state vector $\mathcal{S}$ and $\mathcal{G}$ is the set of composition of smooth functions defined on $\mathcal{S} \subseteq \mathbb{R}$. At the fundamental level, both the proposed algorithms use DQN learning. Thus the above expressions depict the complexity of the proposed algorithms.

C. ON NP HARDNESS OF THE RESOURCE ALLOCATION PROBLEM OF HYBRID RF/VLC

In this section, we make an analysis on the NP hardness of the proposed algorithms. The proposed algorithms are based on the maximization of action value function $Q(s, a)$. The maximization is carried out by bringing $Q(s, a)$ as close to the target $Q(s, a, \phi)$ as possible. It is performed with the help of a neural network based multi layer perception (MLP) network. The action value and the target action value functions need one - one neural network based MLP networks each. Thus, the core function of the proposed algorithm is based on a neural network as shown in Fig. 11. The network considered here has 2 hidden layers of 3 and 2 neurons respectively. This neural network has 7 inputs, 6 from the state vector and 1 from the action vector as shown in the figure.

It can be seen that the state vector has 6 binary input values. First, the input node decides 1 or 0 to be given into the neural network. The bit 1 or 0 is decided according to an $\mathcal{M}$ dimensional linear equation as shown in (33) for the constraints in the problem. For an $i-j$ link,

$$\begin{aligned} I_1^{ij} &= 1 \text{ for } \alpha_{i1}B_{i1} + \alpha_{i1}B_{i1}\alpha_{i1} + \ldots + \alpha_{iM}B_{iM} \leq B_{\max}^v \\ I_1^{ij} &= 0 \text{ for } \alpha_{i1}B_{i1} + \alpha_{i1}B_{i1}\alpha_{i1} + \ldots + \alpha_{iM} \text{ for } B_{iM} > B_{\max}^v \end{aligned} \tag{56}$$

$$\begin{aligned} I_2^{ij} &= 1 \text{ for } \alpha_{i1}B_{i1} + \alpha_{i1}B_{i1}\alpha_{i1} + \ldots + \alpha_{iM}B_{iM} \leq B_{\max}^r \\ I_2^{ij} &= 0 \text{ for } \alpha_{i1}B_{i1} + \alpha_{i1}B_{i1}\alpha_{i1} + \ldots + \alpha_{iM}B_{iM} > B_{\max}^r \end{aligned} \tag{57}$$

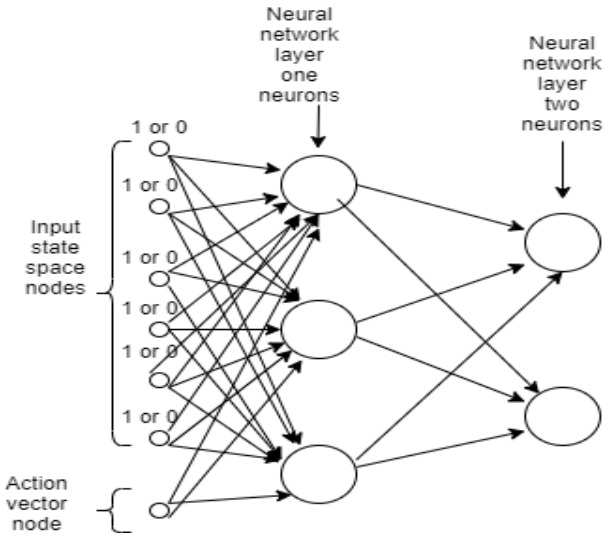

FIGURE 11. Depiction of the Neural Network for performing MLP.

Similar is the case for the other state variables for maintaining the minimum SINR values as

$$I_5^{ij} = 1 \text{ for } \zeta_{i1} + \zeta_{i2} + \ldots + \zeta_{iM} \leq 1$$
$$I_5^{ij} = 0 \text{ for } \zeta_{i1} + \zeta_{i2} + \ldots + \zeta_{iM} > 1 \quad (58)$$
$$I_6^{ij} = 0 \text{ for } \beta_{i1}\zeta_{i1} + \beta_{i1}\zeta_{i1} + \ldots + \beta_{iM}\zeta_{iM} > 1$$
$$I_6^{ij} = 0 \text{ for } \beta_{i1}\zeta_{i1} + \beta_{i1}\zeta_{i1} + \ldots + \beta_{iM}\zeta_{iM} > 1 \quad (59)$$

The state variables pertaining to power constraints involve selection of power P_i within the constraints. Thus, the input nodes involve solving an M dimensional hyperplane. Once the decisions regarding 1 or 0 are formed at the input nodes, each node gives it's decision to each of the neurons of the first layer. Let $I_1^{ij}, \ldots, I_6^{ij}$ be the inputs from the 6 input nodes, the neural network checks $\sum_{r=1}^{6} I_r^{ij}$ is $>$ or $=$ 0, which accounts to solving a 6 dimensional hyperplane.

The neural network needed to solve a hyperplane is NP hard [77]. Therefore, as both the proposed algorithms involve solving hyperplanes with neural network, both of them are NP hard.

D. DISCUSSIONS ON THE BETTER PERFORMANCE OF DQN LEARNING OVER EXHAUSTIVE SEARCH METHOD

A reasonable question comes here that why do the proposed algorithms perform better than the Exhaustive search mechanisms. As mentioned earlier, the Exhaustive search method used here involves a trellis based mechanism. For instance, let us imagine the optimization of the parameters α_{ij}, B_{ij}, and P_i, which lead to the calculation of the action variable γ_{ij} as a traverse between it's initial random value and it's final optimal value. This process involves forming a trellis between the two points. The trellis consists of a certain number of levels, with the final level denoting the optimal value of γ_{ij}. Each level consists of a number of possible values for γ_{ij}. The main objective here is it to determine all the possible paths from the initial random value to the final optimal value. It involves working through the trellis from level 1 to the final level. It involves calculating the number of paths at each level. Let there be $|\mathbb{R}|$ trellis levels with M points at each level. Let $Q(r, m)$ be the number of paths at the point m of the level r, where $1 \leq m \leq M$ possible from the level 1, as shown in Fig. 3. The calculation of the total number of paths possible will be $\sum_{m=1}^{M} Q(r, m)$ which comes out as $M^{|\mathbb{R}|}$. The computational cost associated with each path is $\beta_0(|\mathbb{R}| - 1)$ where β_0 is the average computational cost associated with any path segment in the trellis. The total complexity $Q(r, m)\beta_0(|\mathbb{R}| - 1)$. The process makes Exhaustive search method complex. As a moment-to-moment update is needed in the present work, a limited time span is available for optimizing γ_{ij}. The Exhaustive search method is likely to fail in finishing the optimization of γ_{ij} in the available time span.

As the complexity of Exhaustive search is higher, the Exhaustive search compromises with a lower magnitude of throughput within the designated time span for a moment-to-moment update. The magnitude of throughput reached with DQN learning requires much more time with Exhaustive search. Thus, the final throughput will be lower than that obtained with DQN learning.

E. DISCUSSIONS ON THE CONFLICTS OF INTERESTS AMONG THE ACCESS POINTS

Among the APs, there are conflicts of interest and hence the action-value function $Q(s, a)$ for different APs are related to each other. The action-value function $Q(s, a)$ mainly has two variables, s and a. The variable s signifies the state in which an AP-UE pair are present while the variable a signifies the action taken by the CU for each AP and the UEs associated to it to receive the highest reward. The action-value functions $Q(s, a)$s for different APs are related to each other through s and a, as s and a include the conflicts of interest between the APs. The conflict of interest between the APs occurs in two major ways: in interfering with the signals from other APs and in load balancing. As was mentioned earlier, the interference reduction is expressed in the expression (25) which is written as a constraint for the maximization of r_i. When the DQN algorithm runs for the maximization of r_i, this constraint on interference is included in it. Thus, all the output results are produced with due consideration of this constraint. The second conflict of interest is load balancing, which means that when the UEs get a high SINR from a particular AP, they try to associate with it. As a result, the load on this AP increases severely and it has to bifurcate it's bandwidth into more smaller parts for allotting spectrum to the UEs associated. Consequently, the effective achievable data rate decreases.

Generally in achievable sum rate maximization with conventional optimization methods like the maximum received SINR and maximum received power methods, the cost function is the final achievable sum rate of the system $r = \sum_{i=1}^{N} r_i$. When the problem max r is studied, it may happen that a particular AP from $i \in \mathcal{N}$ lies close to many

UEs and thus offers high SINR. Consequently, a large number of UEs will be associated with this AP causing the problem of load balancing. To address this issue, the most widely used mechanism is to instead maximize $\sum_{i=1}^{N} \log r_i$. This ensures that the final maximization is performed on $\log \prod_{i=1}^{N} r_i$, which ensures that no r_i remains lesser in magnitude as it will harm the final solution.

However, in DQN learning based maximization technique, maximizing the final sum rate r is difficult with learning based mechanism, as the data rate at each AP or each UE needs to be maximized separately. The maximization is performed on each r_i first and then all the r_is are summed up. Thus, this remains the limitation of our work.

VII. CONCLUSION

In this article, the joint optimization problem for bandwidth, power and association parameter allocation in a hybrid RF/VLC system in the downlink has been addressed. It is observed that the problem is neither concave nor convex. To overcome the limitations of conventional optimization algorithms in solving such a problem, a centralized DQN based learning algorithm has been designed, which is based on learning from the hybrid RF/VLC environment. The state vector for DQN is formulated with the constraint terms in the optimization problem, while the action vector for DQN is based on the choice of bandwidth, power, and association parameter. The optimal policy is obtained with the help of an action value function. For opting the appropriate action for optimal policy formulation, the CU picks the appropriate values of bandwidth, power and association parameter from the action vector set. A transfer learning based mechanism for a newly entering UE in the system has also been proposed, which uses the already gathered information in the network for the new entrant. Simulation results verify that the proposed learning based algorithm outperforms the exhaustive search algorithm, the received SINR based association and the received power based association algorithms by more than 10% in terms of achiveable sum-rate and 14.28% in terms of the number of iterations needed for convergence. It is also found that the CU is successful in applying the transfer learning algorithm for using the already gathered information for the new incoming UE in the system, with the maximum achievable sum-rate reached in 54% lesser number of iterations.

APPENDIX
PROOF OF NON-CONCAVITY OF THE ACHIEVABLE DATA RATE OF AN AP

The objective function in the problem $\mathcal{P}$ for $i \in \mathcal{N}\backslash\{0\}$ is given in (60), as shown at the bottom of the page.

Though the sum of logarithmic functions is strictly concave. However, as α_{ij} is an indicator function, r_i will be neither concave nor convex in α_{ij}. It is proved that the function r_i will is neither concave nor convex in B_{ij} and P_i. Let us take a system as $i = 1, 2$ and $j = 1, 2$. Let us assume $\alpha_{11} = 1$, $\alpha_{12} = 0, \alpha_{21} = 0$ and $\alpha_{22} = 1$ and define vector $x = \{x_1, x_2, x_3, x_4\}$ where $x_1 = B_{11}, x_2 = B_{22}, x_3 = P_1$, and $x_4 = P_2$ and $\sqrt{w}G_{11} = a, \sqrt{\rho}G_{12} = b, \sqrt{\rho}G_{21} = c, \sqrt{w}G_2 2 = d$ and $N_0 = g$. Then,

$$r_i(x) = \frac{1}{2}x_1 \log_2\left(1 + \frac{a^2x_3^2}{gx_1 + b^2x_4^2}\right) + \frac{1}{2}x_2 \log_2\left(1 + \frac{d^2x_4^2}{gx_2 + c^2x_3^2}\right). \quad (61)$$

The hessian matrix of r_i wrt x, i.e., $\nabla_x^2\, r_i(x)$ is obtained to check for it's concavity. The elements of $\nabla_x^2\, r_i$ are obtained as

$$\frac{d^2r_i}{dx_1^2} = -\frac{a^2gx_3^2((a^2gx_3^2+2c^2gx_4^2)x_1+2a^2c^2x_4^2x_3^2+2c^4x_4^4}{ln(2)(gx_1 + c^2x_4^2)^2(gx_1 + a^2x_3^2 + c^2x_4^2)^2}, \quad (62)$$

$$\frac{d^2r_i}{dx_2^2} = -\frac{d^2gx_4^2((d^2gx_4^2+2b^2gx_3^2)x_2+2d^2b^2x_3^2x_4^2+2b^4x_3^4}{ln(2)(gx_2 + c^2x_3^2)^2(gx_2 + d^2x_4^2 + b^2x_3^2)^2}, \quad (63)$$

$$\begin{aligned}\frac{d^2r_i}{dx_3^2} = &-\frac{-2d^2b^2x_4^2x_2}{(b^2x_3^2 + gx_2)^2(\frac{d^2x_4^2}{b^2x_3^2+gx_2} + 1)} \\ &+ \frac{8d^2b^4x_4^2x_2x_3^2}{(b^2x_3^2 + gx_2)^3(\frac{d^2x_4^2}{b^2x_3^2+gx_2} + 1)} \\ &- \frac{4d^4b^4x_4^4x_2x_3^2}{(b^2x_3^2 + gx_2)^4(\frac{d^2x_4^2}{b^2x_3^2+gx_2} + 1)} \\ &+ \frac{2a^2x_1}{(gx_1 + c^2x_4^2)(\frac{a^2x_3^2}{gx_1+c^2x_4^2} + 1)} \\ &- \frac{4a^2x_1x_3^2}{(gx_1 + c^2x_4^2)^2(\frac{a^2x_3^2}{gx_1+c^2x_4^2} + 1)},\end{aligned} \quad (64)$$

and

$$\begin{aligned}\frac{d^2r_i}{dx_4^2} = &-\frac{-2a^2c^2x_3^2x_1}{(c^2x_4^2 + gx_1)^2(\frac{a^2x_3^2}{c^2x_4^2+gx_1} + 1)} \\ &+ \frac{8a^2c^4x_3^2x_1x_4^2}{(c^2x_4^2 + gx_1)^3(\frac{a^2x_3^2}{c^2x_4^2+gx_1} + 1)}\end{aligned}$$

$$r_i = \sum_{j\in\mathcal{M}} \frac{1}{2}\alpha_{ij}B_{ij}^{VLC} \times \log_2\left(1 + w\frac{G_{ij}^2P_i^2}{N_0^vB_{ij}^{\text{VLC}} + \sum_{k\in\mathcal{N}\backslash\{i\}} \rho_jG_{kj}^2P_k^2\left(1 - \prod_{j'\in\mathcal{M}\backslash\{j\}}(1-\alpha_{kj'})\right)}\right). \quad (60)$$

$$-\frac{4a^4c^4x_3^4x_1x_4^2}{(c^2x_4^2+gx_1)^4(\frac{a^2x_3^2}{c^2x_4^2+gx_1}+1)}+\frac{2d^2x_2}{(gx_2+b^2x_3^2)(\frac{d^2x_4^2}{gx_2+b^2x_3^2}+1)}-\frac{4d^2x_2x_4^2}{(gx_2+b^2x_3^2)^2(\frac{d^2x_4^2}{gx_2+b^2x_3^2}+1)}. \quad (65)$$

It can be seen that $\frac{d^2\ r_i}{dx_1^2}$ and $\frac{d^2\ r_i}{dx_2^2}$ are always negative but the nature of $\frac{d^2\ r_i}{dx_3^2}$ and $\frac{d^2\ r_i}{dx_4^2}$ is not fixed. It can become positive or negative for varying values of B_{11}, B_{22}, P_1 and P_2. Thus, for the VLC network, r_i will neither be concave nor convex in B_{11}, B_{22}, P_1, and P_2. Next, the behavior of r_i in the RF network has been investigated. The system model consists of only one RF AP indexed as $i = 0$. The achievable data rate for the RF network is given as

$$r_i = \sum_{j\in\mathcal{M}} \alpha_{0j}B_{0j}\log_2\left(1+\frac{P_0G_{0j}}{N_0^rB_{0j}^{\mathrm{RF}}}\right). \quad (66)$$

As mentioned earlier, as α_{0j} is an indicator function, r_i will be neither concave nor convex in α_{0j}. Further, the concavity of r_i with P_0 and B_{0j} has been investigated. For simplicity in calculations, let us consider $\alpha_{0j} = 1$, $|\mathcal{M}| = 1$, $N_0^r = a$ and a vector $x = \{x_1\ x_2\}$ where $x_1 = B_{01}$ and $x_2 = P_0\ G_{0j}$. The achievable rate can be written as

$$r_i(x) = x_1\log_2\left(1+\frac{x_2}{ax_1}\right). \quad (67)$$

The elements of the Hessian matrix $\nabla_x^2\ r_i(x)$ are obtained as

$$\frac{d^2r_i(x)}{dx_1^2} = \frac{-x_2^2}{ln(2)x_1(ax_1+x_2)^2}, \quad (68)$$

and

$$\frac{d^2r_i(x)}{dx_2^2} = \frac{-a^2x_1}{ln(2)(ax_1+x_2)^2}. \quad (69)$$

It can be seen that both the elements of $\nabla_x^2\ r_i(x)$ are negative. The higher order of $|\mathcal{M}|$ will result to sum of such similar functions. Sum of concave functions is a concave function. This means that for the RF network, r_i will be jointly concave in B_{0j} and P_0. However, it is neither concave nor convex in the indicator function α_{0j}. Thus, jointly it will be neither concave nor convex in α_{0j}, B_{0j} and P_0.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable reviews on the work. The reviews were instrumental in improving it.

REFERENCES

[1] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2047–2077, 4th Quart., 2015.
[2] J. Luo, L. Fan, and H. Li, "Indoor positioning systems based on visible light communication: State of the art," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2871–2893, 4th Quart., 2017.
[3] A.-M. Cailean and M. Dimian, "Current challenges for visible light communications usage in vehicle applications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2681–2703, 4th Quart., 2017.
[4] O. Babatundi, L. Qian, and J. Cheng, "Downlink scheduling in visible light communications," in *Proc. 6th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2014, pp. 1–6.
[5] D. A. Basnayaka and H. Haas, "Design and analysis of a hybrid radio frequency and visible light communication system," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4334–4347, Oct. 2017.
[6] M. Obeed, A. M. Salhab, M.-S. Alouini, and S. A. Zummo, "On optimizing VLC networks for downlink multi-user transmission: A survey," 2018, *arXiv:1808.05089*. [Online]. Available: http://arxiv.org/abs/1808.05089
[7] S. Bayat, R. H. Y. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Aug. 2014.
[8] A. R. Elsherif, W.-P. Chen, A. Ito, and Z. Ding, "Resource allocation and inter-cell interference management for dual-access small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1082–1096, Jun. 2015.
[9] M. Chen, S. Chang Liew, Z. Shao, and C. Kai, "Markov approximation for combinatorial network optimization," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6301–6327, Oct. 2013.
[10] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[11] S. C. Chen, N. Bambos, and G. J. Pottie, "Admission control schemes for wireless communication networks with adjustable transmitter powers," in *Proc. INFOCOM Conf. Comput. Commun.*, Jun. 1994, pp. 21–28.
[12] S. Pietrzyk and G. J. M. Janssen, "Radio resource allocation for cellular networks based on OFDMA with qos guarantees," in *Proc. IEEE Global Telecommun. Conf. GLOBECOM*, Nov. 2004, pp. 2694–2699.
[13] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 884–895, Jun. 1999.
[14] I. Stefan and H. Haas, "Hybrid visible light and radio frequency communication systems," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.
[15] M. Obeed, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "Joint optimization of power allocation and load balancing for hybrid VLC/RF networks," *J. Opt. Commun. Netw.*, vol. 10, no. 5, pp. 553–562, May 2018.
[16] Y. S. M. Pratama and K. W. Choi, "Bandwidth aggregation protocol and throughput-optimal scheduler for hybrid RF and visible light communication systems," *IEEE Access*, vol. 6, pp. 32173–32187, 2018.
[17] M. Obeed, H. Dahrouj, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "DC-bias and power allocation in cooperative VLC networks for joint information and energy transfer," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5486–5499, Dec. 2019.
[18] M. Kashef, M. Ismail, M. Abdallah, K. A. Qaraqe, and E. Serpedin, "Energy efficient resource allocation for mixed RF/VLC heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 883–893, Apr. 2016.
[19] H. Li, Z. Huang, Y. Xiao, S. Zhan, and Y. Ji, "A power and spectrum efficient NOMA scheme for VLC network based on hierarchical predistorted LACO-OFDM," *IEEE Access*, vol. 7, pp. 48565–48571, 2019.
[20] A. Khreishah, S. Shao, A. Gharaibeh, M. Ayyash, H. Elgala, and N. Ansari, "A hybrid RF-VLC system for energy efficient wireless accessm," Jun. 2018, *arXiv:1806.05265*. [Online]. Available: https://arxiv.org/abs/1806.05265
[21] M. Kafafy, Y. Fahmy, M. Abdallah, and M. Khairy, "Power efficient downlink resource allocation for hybrid RF/VLC wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.
[22] T. Rakia, H.-C. Yang, F. Gebali, and M.-S. Alouini, "Dual-hop VLC/RF transmission system with energy harvesting relay under delay constraint," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.
[23] R. Jiang, Q. Wang, H. Haas, and Z. Wang, "Joint user association and power allocation for cell-free visible light communication networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 1, pp. 136–148, Jan. 2018.
[24] L. Li, Y. Zhang, B. Fan, and H. Tian, "Mobility-aware load balancing scheme in hybrid VLC-LTE networks," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2276–2279, Nov. 2016.

[25] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33275–33284, 2018.

[26] M. Hammouda, S. Akın, A. Maria Vegni, H. Haas, and J. Peissig, "Hybrid RF/VLC systems under QoS constraints," 2018, *arXiv:1804.05211*. [Online]. Available: http://arxiv.org/abs/1804.05211

[27] C. Shen, S. Lou, C. Gong, and Z. Xu, "User association with lighting constraints in visible light communication systems," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2016, pp. 222–227.

[28] R. Zhang, Y. Cui, H. Claussen, H. Haas, and L. Hanzo, "Anticipatory association for indoor visible light communications: Light, follow me!," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2499–2510, Apr. 2018.

[29] *DARPA Spectrum Collaboration Challenge (SC2) at Mobile World Congress Americas*. Accessed: Oct. 23, 2019. [Online]. Available: https://spectrumcollaborationchallenge.com/

[30] *Kick-off Meeting*, The Johns Hopkins Univ., Dapra, Arlington County, VA, USA, Jan. 2017.

[31] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.

[32] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 300–309, Jan. 2012.

[33] K. Cohen, Q. Zhao, and A. Scaglione, "Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 1575–1578.

[34] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," 2017, *arXiv:1702.07031*. [Online]. Available: http://arxiv.org/abs/1702.07031

[35] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[36] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT Press, 1998.

[37] K. Arulkumaran, M. Peter Deisenroth, M. Brundage, and A. Anthony Bharath, "A brief survey of deep reinforcement learning," 2017, *arXiv:1708.05866*. [Online]. Available: http://arxiv.org/abs/1708.05866

[38] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Dept. Comput. Sci., King's College, Univ. Cambridge, Cambridge, U.K., May 1989. [Online]. Available: http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf

[39] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[40] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3996–4012, Sep. 2016.

[41] Z. Li, C. Wang, and C.-J. Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2217–2228, Aug. 2017.

[42] E. Ghadimi, F. Davide Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[43] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[44] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018, doi: 10.1109/TCCN.2018.2809722.

[45] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.

[46] X. Wan, G. Sheng, Y. Li, L. Xiao, and X. Du, "Reinforcement learning based mobile offloading for cloud-based malware detection," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[47] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[48] U. Challita, W. Saad, and C. Bettstetter, "Cellular-connected UAVs over 5G: Deep reinforcement learning for interference management," 2018, *arXiv:1801.05500*. [Online]. Available: http://arxiv.org/abs/1801.05500

[49] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[50] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan. 2016.

[51] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[52] (2017). *IEEE Network Special Issue: Exploring Deep Learning for Efficient and Reliable Mobile Sensing*. Accessed: Jul. 14, 2017. [Online]. Available: http://www.comsoc.org/netmag/cfp/ exploring-deep-learning-efficient-and-reliable-mobile-sensing

[53] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, "Machine learning for networking: Workflow, advances and opportunities," *IEEE Netw.*, vol. 32, no. 2, pp. 92–99, Mar. 2018.

[54] M. A. Alsheikh, D. Niyato, S. Lin, H.-P. Tan, and Z. Han, "Mobile big data analytics using deep learning and apache spark," *IEEE Netw.*, vol. 30, no. 3, pp. 22–29, May 2016.

[55] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[56] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.

[57] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-Art deep learning: Evolving machine intelligence toward Tomorrow's intelligent network traffic control systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2432–2455, 4th Quart., 2017.

[58] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, 4th Quart., 2018.

[59] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," 2017, *arXiv:1710.02913*. [Online]. Available: http://arxiv.org/abs/1710.02913

[60] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*. [Online]. Available: http://arxiv.org/abs/1312.5602

[61] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[62] J. Ye and Y.-J. A. Zhang, "DRAG: Deep reinforcement learning based base station activation in heterogeneous networks," 2018, *arXiv:1809.02159*. [Online]. Available: http://arxiv.org/abs/1809.02159

[63] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, "Learning-based energy-efficient resource management by heterogeneous RF/VLC for ultra-reliable low-latency industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5565–5576, Aug. 2020, doi: 10.1109/TII.2019.2933867.

[64] M. Kashef, M. Abdallah, N. Al-Dhahir, and K. Qaraqe, "On the impact of PLC backhauling in multi-user hybrid VLC/RF communication systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[65] J. R. Barry, *Wireless Infrared Communications*. Norwell, MA, USA: Kluwer, 1994.

[66] P. Kyosti and J. Meinila, *Winner II Channel Models*, document D1.1.2 V1.2. IST-4-027756 Winner II, Sep. 2007.

[67] S. Hranilovic and F. R. Kschischang, "Capacity bounds for power- and band-limited optical intensity channels corrupted by Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 784–795, May 2004.

[68] A. Lapidoth, S. M. Moser, and M. A. Wigger, "On the capacity of free-space optical intensity channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4449–4461, Oct. 2009.

[69] A. A. Farid and S. Hranilovic, "Capacity bounds for wireless optical intensity channels with Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6066–6077, Dec. 2010.

[70] A. Chaaban, J.-M. Morvan, and M.-S. Alouini, "Free-space optical communications: Capacity bounds, approximations, and a new sphere-packing perspective," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 1176–1191, Mar. 2016.

[71] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid RF/VLC networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 443–447, Apr. 2020, doi: 10.1109/LWC.2019.2958121.
[72] Z. Mlika, E. Driouch, and W. Ajib, "User association under SINR constraints in HetNets: Upper bound and NP-hardness," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1672–1675, Aug. 2018, doi: 10.1109/LCOMM.2018.2840714.
[73] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1025–1039, Jun. 2015, doi: 10.1109/JSAC.2015.2417011.
[74] C. R. N. Athaudage, "On computational complexity of optimized temporal decomposition algorithm for speech signal modelling," in *Proc. 6th Int. Conf. Signal Process.*, Aug. 2002, pp. 437–440, doi: 10.1109/ICOSP.2002.1181084.
[75] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2019, *arXiv:1901.00137*. [Online]. Available: https://arxiv.org/abs/1901.00137
[76] C. H. Liu, Q. Lin, and S. Wen, "Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3516–3526, Jun. 2019.
[77] A. L. Blum and R. L. Rivest, "Training a 3-node neural network is NP-complete," *Neural Netw.*, vol. 5, no. 1, pp. 117–127, Jan. 1992.

SHIVANSHU SHRIVASTAVA received the B.E. degree in electronics and telecommunications from the Shri Shankaracharya College of Engineering and Technology, Bhilai, India, in 2010, and the Ph.D. degree in communication engineering from the Department of Electronics and Electrical Engineering, IIT Guwahati, in 2017. From 2017 to 2018, he was with the Department of Electrical Engineering, IIT Kanpur, India, as the Project Investigator of the Science and Engineering Research Board Project entitled Designing Energy Efficient Hybrid RF/VLC Based CRANs for 5G. Since 2019, he has been with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, as a Postdoctoral Researcher. His research interests include visible light communications, cognitive radios, and wireless communications.

BIN CHEN (Senior Member, IEEE) received the B.E. and M.S. degrees in electronic engineering from Lanzhou University, in 1997 and 2002, respectively, and the Ph.D. degree in communication engineering from Nanyang Technological University, in 2007. He is currently an Associate Professor with Shenzhen University. His research interests include blockchain, optical networking, and neural networks.

CHEN CHEN (Member, IEEE) received the B.S. and M.Eng. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2010 and 2013, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2017. He is currently a Tenured-Track Assistant Professor with the School of Microelectronics and Communication Engineering, Chongqing University, China. His research interests include visible light communications, Li-Fi, visible light positioning, optical access networks, and digital signal processing.

HUI WANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, in 1990, 1993, and 1996, respectively, all in telecommunication. He is currently a Professor with the College of Electronics and Information Engineering, Shenzhen University. His research interests include wireless communication, signal processing, and distributed computing systems, in which, he is an author or a coauthor of more than 50 international leading journals, conferences, and book chapters.

MINGJUN DAI (Member, IEEE) received the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, in 2012. He is currently an Associate Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include distributed systems, network coding design, and optimization of wireless networks.