2021

Deep Reinforcement Learning based Handover Management for Millimeter Wave Communication

Mollel, Michael S.

International Journal of Advanced Computer Science and Applications,

https://dx.doi.org/10.14569/IJACSA.2021.0120298

Deep Reinforcement Learning based Handover Management for Millimeter Wave Communication

Michael S.Mollel[1], Shubi Kaijage[2], Michael Kisangiri[3]
The Nelson Mandela African Institution of Science and Technology (NM-AIST)

***Abstract*—The Millimeter Wave (mm-wave) band has a broad-spectrum capable of transmitting multi-gigabit per-second date-rate. However, the band suffers seriously from obstruction and high path loss, resulting in line-of-sight (LOS) and non-line-of-sight (NLOS) transmissions. All these lead to significant fluctuation in the signal received at the user end. Signal fluctuations present an unprecedented challenge in implementing the fifth generation (5G) use-cases of the mm-wave spectrum. It also increases the user's chances of changing the serving Base Station (BS) in the process, commonly known as Handover (HO). HO events become frequent for an ultra-dense dense network scenario, and HO management becomes increasingly challenging as the number of BS increases. HOs reduce network throughput, and hence the significance of mm-wave to 5G wireless system is diminished without adequate HO control. In this study, we propose a model for HO control based on the offline reinforcement learning (RL) algorithm that autonomously and smartly optimizes HO decisions taking into account prolonged user connectivity and throughput. We conclude by presenting the proposed model's performance and comparing it with the state-of-art model, rate based HO scheme. The results reveal that the proposed model decreases excess HO by 70%, thus achieving a higher throughput relative to the rates based HO scheme.**

Keywords*—*Handover management; 5G; machine learning; reinforcement learning; mm-wave communication

I. Introduction

Unlike its predecessors, the fifth-generation (5G) of mobile communication networks has been considered a paradigm shift due to its attractive service in terms of latency, data rates, device inter-connectivity, and network flexibility. These enhancements in Key Performance Indicators (KPIs) make 5G a game-changer by allowing new applications such as remote surgery, smart cities, device-to-device communication (D2D), industrial Internet, smart agriculture, etc. [1].

To meet these service requirements and demands, 3GPP has launched the New Radio (NR) standardization with the following use cases: enhanced mobile broadband (eMBB), massive Machine Type Communication (mMTC), and Ultra-Reliable Low-Latency Communication (URLLC) [2], [3]. eMBB aims at enhancing the system capacity and supporting the ever-increasing end-user data rate. eMBB introduces two significant technological enhancements: mm-wave use to achieve higher data rate and antenna array that supports massive multiple-input and multiple-output (MIMO) beamforming. URLLC introduces entirely new use-cases requirements to support vertical industries such as self-driving cars, remotely surgery for eHealth and other mission-critical use cases. The unique features introduced by URLLC include improved latency, reliability while guaranteeing high service availability and security. mMTC intends to provide cost-efficient and robust connection of billions of devices that transmit small packets of data (with 10s latency) but without overloading the network. Some factor to consider in mMTC are low power consumption, longtime availability of service, and coverage. mMTC can also be seen as a particular case of URLLC with more emphasis placed on reliability while less emphasis is placed on the latency [3]. The new use cases pave the way for increasing interconnected devices to the Internet, resulting in the Intenet of Things (IoT) development. IoT is a technology that targets to connect everyday devices (e.g., home appliances, wearable devices) to the Internet, making the scenario even severer. The considerable projection increase in the number of cellular IoT devices in the near future [4] entails 5G networks dealing with stringent requirements and an increasing number of connected devices.

Heterogeneous network (HetNet), Ultra-Dense Network (UDN) and the use of mm-wave are candidate solutions to overcome the possible challenges of 5G networks [5]. Together, they can significantly increase network throughput, available bandwidth and spectral efficiency [6]. HetNet is the deployment of various base station (BS) topology based on coverage footprint and type of Radio Access technology used [7]. Moreover, densification of the network is a phenomenon of deploying more small cells (SCs) in the network to increase cell density, coverage, and network throughput. The main challenge of deploying UDN is increased interference sources and signal fluctuation. For example, there are many access points (AP) and cells in crowded substations or stadiums; thus, signals can have more reflecting and scattering paths, contributing to high signal interference and fluctuation. On the other hand, the concept of utilizing a broader bandwidth refers to opening up a new frequency spectrum for mobile communication to increase the available bandwidth. mm-Wave frequencies offer great potentials in terms of data rate due to their larger bandwidth, and mm-wave bands have been designated as Frequency Range-2 (FR2) in 5G New Radio (NR) [8]. Nevertheless, the mm-wave spectrum comes with its limitation as it is more likely to suffer from extreme penetration losses due to higher frequencies. Thus, mm-wave use as carrier frequency decreases the BS footprint area, thereby resulting in multiple SCs in the network.

Network densification is an inevitable destination for network operators to provide a more sustainable and enhanced Quality of Service (QoS) for mobile users. However, network densification with SCs is not a solution without any side effects; it increases the number of HOs, which is characterized by changing from one BS to another BS for the user equipment (UE) when there is an ongoing communication (voice or data). Given the limited coverage area of SCs, the UE would need more HOs since there will be more BSs within an area of interest after the densification. Moreover, the different types of BSs from HetNet deployment will result in complicated HO signalling processes [6]. Furthermore, considering that the HO interval is inversely proportional to the UE speed [9], the case becomes even more severe in the case of high mobility user.

HO process involves exchanging information between serving BS, target BS and Core Network (CN). Exchanged messages, commonly known as signalling overheads, is necessary during the three (3) steps involved in HO, which are HO preparation, execution and completion. If excessive and undesirable HO increases, then both signalling overhead and average HO interruption time increases [10], [11]. The high signalling overhead and HO interruption time result in a significant increase in latency, thereby undermining the attempt to meet with 5G network specifications, particularly the URLLC use cases. Besides, the average throughput also decays with the increasing number of HOs, resulting in degraded quality of experience (QoE) for the users [12]. Therefore, it is apparent that special consideration should be given to HO management to ultimately achieve and unleash the potential of the 5G networks by meeting all its requirements.

To meet the 5G expectation, novel and advanced HO control that minimizes the effects of HO are required. The focus is on reducing unnecessary, and unwanted HO events such as ping-pong and frequent HOs, and the main parameters to be considered are the total number of HOs per UE trajectory and the time spent during HO. These parameters together define HO cost, which is the multiplication of both parameters [12]. In other words, the total number of HOs and the time spent during a single HO should be reduced to get away with one of the negative implications of using mm-wave spectrum in the UDNs. The former can be achieved through an intelligent method by avoiding 'unnecessary' HOs, whereas the latter is a characteristic of the RAT [13]. Therefore, in this paper, we present an intelligent method based on DRL for HO reduction in mm-wave BSs in a UDN environment.

The rest of this article is organized as follows. First, in Section II, we describe HO management in 5G networks, and a review of the state-of-the-art HO management approaches, then in Section III, the Deep Reinforcement Learning (DRL) framework was introduced as well as how it is linked to HO problem. Next, in Section IV, the use case is presented as well as a description of the simulation environment. In Section V, we evaluate the performance of the proposed model and compare it with the rate based HO scheme. Finally, Section VI concludes the paper.

II. HO Management in 5G Networks

HO is described as the process of transferring an ongoing UE's resource from one channel to another in wireless mobile communication. The process mainly involves a change of connection from either serving BS, carrier frequency channel or prioritizing a new technology found within the UEs' vicinity. One of the key design strategies for the successful implementation of 5G networks is the efficient handling of HO to make UEs seamlessly change BS association, thereby limiting unnecessary HO. HO process in mobile communication involves three states. The first stage is the measurement or information gathering phase, where the UE measures the signal strength (other parameter measurements are also possible) of every potential neighbour BS and the current serving BS. The second phase is about the HO decision, where the current serving BS decides to initialize the HO based on the measured data from the first stage. The third phase is the cell exchange phase, when the UE releases the serving BS and connects to the new BS [14].

Traditionally, HO is of two types, hard and soft HOs. In the case of hard HO, the connection must be released from the serving BS before the connection with the target BS can be established. In soft HO, the serving BS connection is maintained and used for a while in parallel with the target BS connection [14]. 5G mm-wave communication supports the hard HO method in most cases [8]. Besides, it supports dual connectivity, which means that the UE can be connected to more than one BS. However, when it comes to HO in dual connectivity, the individual connections perform hard HO, and new HO scenarios emerge, which lead to more HO complications in mm-wave communication [15].

Mm-wave communication is already severely affected by blockages and high path loss; thus, deploying multiple mm-wave BSs would result in additional challenges, particularly from HO management's perspective. Hence, by adopting hard HO in mm-wave communication, the UE will often experience intermittent connections, leading to poor QoE regardless of QoS. One of the causes of UE dissatisfaction from mm-wave BS might be either blockage or interference, leading to a reduction in the SNR of the serving BS; these situations present a ping-pong problem. Another cause of UE dissatisfaction from mm-wave BS is when UE moves out of signal range since it is known that the UE experiences excellent coverage of mm-wave communication when it is within 200m from the serving BS [16]. The challenge is selecting BS intelligently during HO in such a way that leads to a few ping pong, reducing unnecessary HO, and maintaining UE-BS connectivity for a long duration. Generally, optimal BS selection to re-associate with UE is needed to reduce the problem mentioned above. In legacy technology, fourth generation and all technology which use sub 6 GHz, the issue of HO is less severe considering the sparse nature of BS deployment compared to 5G, which uses mm-wave frequencies. Furthermore, sub 6 GHz has a broad coverage compared to mm-wave, making unnecessary HO less frequent. It is worth noting that the HO process involves several procedures, but we present the general conditions required for HO to occur for the sake of simplicity.

A. HO Process in 5G

In 5G, 3GPP [8] defines six HO events for entering and leaving. These events are A1, A2, A3, A4, A5, and A6 and are used to trigger HO. They are described as follows [17]: Event A2 and A1 are activated when the UE's channel condition drops below and exceeds the configured threshold, respectively. They are also used to start and stop inter-frequency neighbour search. Intra-frequency HO is initiated by event A3 when the neighbouring channel's condition is higher than the service channel's condition based on the configured threshold. Event A4 and A5 are typically used for inter-frequency HO, where the target cell's signal strength has to be higher than the absolute threshold for the A4 event to be triggered. In addition to

Event A4, however, event A5 requires that the serving BS radio frequency (RF) condition be below a certain threshold. Event A6 is similar to event A3 but is used for intra-frequency HO to the secondary frequency on which the UE is encamped. Event A4 and A5 can also be used for conditional HO management, e.g. load balancing. Event B1 and B2 specifies the entering and leaving condition for inter-RAT HO [8]. The threshold values are all configured value, and if they are correctly configured, they can significantly reduce the number of unnecessary HOs. In this paper, we assume for UE to HO, one of the trigger conditions for HO must be met.

However, these HO events only show the minimum requirements for the UE to undergo HO. The HO trigger events do not include any intelligence in deciding which BS to associate UEs with, especially when choosing among multiple BSs. Hence, it always chooses the BS that provides the highest empirical rewards, for instance, BS with the highest signal to interference plus noise ratio (SINR) or highest reference signal received power (RSRP). Furthermore, the selection of the optimal BS to HO does not only depend on the BS which provides the maximum instantaneous reward SINR but other factors such as throughput, which depends on bandwidth and number of UEs, also need to be considered, especially for mm-wave BS, thus, making the matter of optimal BS selection an open issue.

The conventional event-based HO trigger depends only on the UE's measurement report (MR) rather than the general network perception, which often results in sub-optimal HO decisions. Moreover, in 5G, HO decisions would be taken at the network level, where both the distribution and load of users alongside BSs status would be considered. Intelligence is therefore required to make optimum decisions regarding selecting the target BS by incorporating or considering other appropriate features during the BS selection process.

B. State-of-the-Art HO Management Approaches

In [18], the authors addressed the HO prediction method in 5G and used RL to find the optimal beam that the UE should select to maximize throughput. Their method assumes that the state fed to RL is the combination of all RSRP values seen from all surrounding BSs. However, considering the states as discrete values in such a complex environment, the proposed solution does not generalize the HO solution. The states created by combining RSRP are continuous intrinsic values and not discrete values as assumed. The actual network generates continuous RSRP values.

More recently, there have been several studies that solve HO using multi-armed bandit. The armed bandit is the classic probability-based RL problem. In [19] the authors assume the UE as an agent and set the BS as an arm which the UE chooses to maximize its return, which is the average throughput for their case. The dynamics of the environment was well-considered and captured in the learning process. However, they only considered UE dynamics in their work without considering the dynamics of the environment, such as moving and stationary obstacles, which can make the solution more complex. They also did not consider user trajectory. Despite the success of [20] in optimizing HO from an energy point of view, the proposed model is still insufficient as it ignores some vital factors such as UEs trajectory and distribution as well as the available bandwidth in the target BS.

In addition, different heuristic approaches have also been proposed as an alternative solution to the HO problem. Several researchers have focused their attention on different HO management techniques using these approaches. For example, [14] demonstrates how inter-cell interference coordination (ICIC) can be used to enhance HO decision performance. There is also a more advanced version of ICIC known as enhanced Inter-Cell Interference Coordination (eICIC), which can reduce the HO failure ratio (HoF) and the radio link failure (RLF) compared to the case without eICIC. However, despite the advantages of this method, it involves extensive overhead signalling during coordination between the BS and finding the global solution regarding when and which BS to HO, thereby increasing delay and degrading UE's QoE. A BS skipping technique for mobile UEs that demonstrates a significant increase in the overall UE throughput was proposed in [12]. The authors take advantage of a coordinating BS in deciding which BS to select to reduce the number of HOs. They also added a HO cost function, which penalizes the action of HO and maintains the minimum SINR as much as possible to avoid taking HO. Their method has been proven to work based on stochastic analysis, but the fundamental question remains how to skip BSs smartly. Hence, there is a need to develop intelligent BS skipping techniques which incorporates all the necessary factors during decision making.

In order to overcome the stated challenges while achieving high throughput in mm-wave communication, we propose a DRL algorithm that intelligently selects the BS that will prolong UE-BS association while guaranteeing maximum throughput. We develop an efficient method that alleviates the effect of HO and help realize the potential of mm-wave frequency in 5G systems. We leverage the availability of extensive data that the network generates during the training phase. The advantage of the proposed method is that it learns offline before its deployment to the BS controller to assist in HO decision. The model aims to maximize the system's average throughput by considering the signal to noise ratio (SNR), UE velocity, number of HOs per UE trajectory, and network load balancing.

III. Reinforcement Learning Assisted HO Management

Our objective is to achieve the maximum throughput, which is achieved if the whole network environment is considered. The network environment includes, but is not limited to, UE trajectories, velocity and distribution, blockages, and BS distribution and UE velocity. Some of these factors vary with time, while others do not. Therefore, it is difficult for the heuristic approaches to solve the HO problem while including changing factors over time. Hence, the solution is to explore the environment and exploit the actions that achieve the intended objective. Artificial Intelligence (AI) has a class of algorithms known as RL that solves this problem; these algorithms learn through trial-and-error. When combined with Deep Neural Network (DNN), RL forms DRL, which performs exhaustive search and learns by themselves through experience from interacting with the environment to achieve the objective of maximizing or minimizing the objective function.

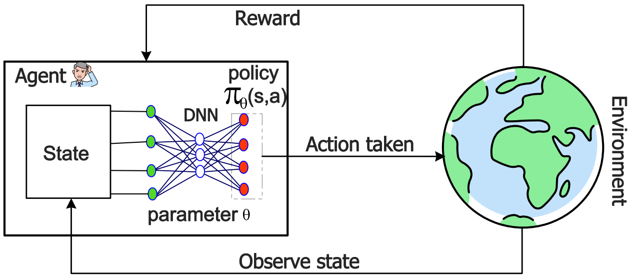


Fig. 1. General Framework of RL.

A. Reinforcement Learning

This section gives a brief overview of the RL and DRL framework and further discusses how the HO optimization problem is formulated and solved using the DRL algorithm.

1) RL Framework: RL is a subfield of AI that enables machines to create artificially intelligent agents that learn to optimize their accumulated reward by interacting with the environment. In RL, the agent receives feedback after each action. The feedback includes the reward and the next state of the environment. The relationship between agent, action and environment is shown in Fig. 1 [21]. The agent learns the best policy through multiple interactions with the environment, and the learning procedure is detailed in the following paragraphs.

Here, we first define the main elements of RL. At time t, the agent observes the state of the environment, $s_t \in S$, where S is the set of possible states. After observing state s_t agent takes an action, $a_t \in A(s_t)$ where $A(s_t)$ is the set of possible actions at state s_t. After selecting and taking the action a_t from state s_t, agent receives the immediate reward r_{t+1} from state-action pair (s_t, a_t). The selected action in state s_t moves the agent to state s_{t+1} at time $t + 1$. It is essential for the environment to have state dynamics such that P$(s_{t+1}|s_t, a_t)$ exists. There are two approaches to solving RL problems: The first approach is based on policy search, and the second approach is based on the value function approximation. Their names reflect their behaviour. The former searches directly for the optimal policy based on a parameterizing policy such as NN. The later keeps improving the value function estimate by selecting actions greedily according to the previously updated value function and indirectly learning optimal policy.

RL methods have a dilemma, which is the trade-off between exploitation and exploration. This has to do with how the agent learns the environment through trial and error. Should the agent be encouraged to perform exploitation or exploration during learning? Exploitation implies that the agent acts more greedily by taking the best actions that maximize the reward. Exploration means the agent act less greedy, so it can learn about the environment more to find optimal actions. The most common solution to this dilemma is the e-greedy policy where the agent explores with probability less than $\epsilon \in [0, 1]$ and exploits the best action otherwise is applicable to value function. For policy search methods, the problem is less severe.

2) DRL Framework: All RL methods based on tabular solution suffer from the so-called ”the curse of dimensionality”, which means that computational requirements increase exponentially with an increase in the number of states. Moreover, for the task involving continuous states, the problem becomes severe. To overcome this problem, DRL is introduced by exploiting the advantage of neural networks (NN) in the traditional RL. The idea behind DRL is to train neural networks to approximate optimal policy [21].

In [22], the authors combine deep convolutional neural networks (CNN) with RL to develop a novel artificial agent capable of learning successful policies directly from high-dimensional sensory input data. The CNN is used to represent the action-value function, denoted as $Q(s, a; \theta)$, where $Q(s, a)$ represents the action-value function and the parameter θ is the weight of the neural networks. θ is updated every time Q - network performs an iteration with the mean square error as the loss function. The loss function is the mean square error between the action-value $Q(s, a; \theta)$ and target values $r + \gamma \cdot \max_{a'} Q^*(s', a'; \theta^-)$.

It is imperative to train the neural network using training samples from both the previous and current episodes. This is necessary because approximating the optimal policy direct using only current samples results in slower learning and undesirable temporal correlations. To solve this problem, the concept of experience replay, in which previous experiences by the agent at each time-step $(s_\mathrm{t}, a_\mathrm{t}, r_\mathrm{t}, s_{\mathrm{t+1}})$ as well as recent experience are stored for subsequent use in the training phase. The experience replay buffers previous experiences and randomly selects the training set over the data. This results in the gradual smoothing of the data distribution to avoid the bias of the sample data.

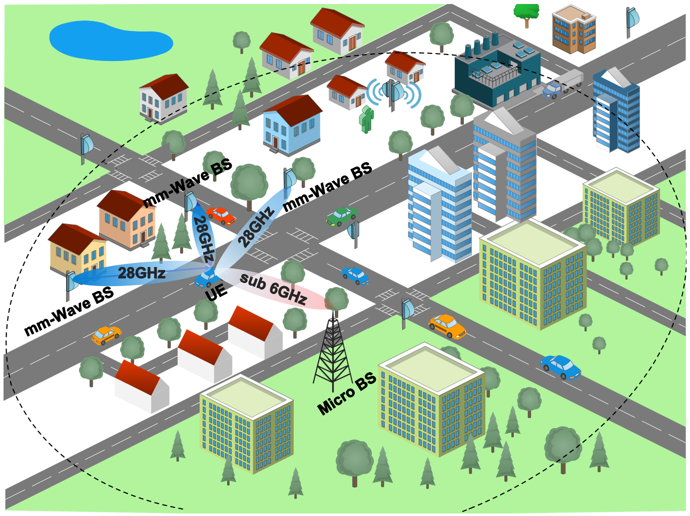


Fig. 2. Overview of Heterogeneous Network (HetNets) with Dense mm-wave BS, UE’s and sub 6 GHz BS in the Urban Area.

IV. DRL-Aided Intelligent BS Selection

In this section, we explain our considered system model. Then, we describe the proposed DRL optimal BS selection framework. It is worth noting that the DRL framework is based on Deep Q Network (DQN) and that both terms would be used interchangeably for the rest of this paper.

A. System Model

We consider Fig. 2 as our use-case system model, which demonstrate a simplified 5G HetNet where the mm-wave SCs are placed close to each other as part of the HetNet. For simplicity, we assume that every BS and UE has a single antenna and 28 GHz, 2.1 GHz are used for mm-wave BS and sub-6 GHz BSs respectively. The environment consists of a sub-6 GHz macro BS, UE's, and the mm-wave BSs in Fig. 2. Wireless Insite (WI) software is used to develop the environment, and it uses ray tracing, which provides accurate results that mimic the actual network environment. SINR is a popular metric for measuring channel quality. In the system model, however, we consider SNR, and the reason is that mm-wave antennas are capable of forming directional beams; therefore, Inter-cell interference contribution is assumed to be negligible.

B. Proposed Optimal Base Station Selection based on DRL

In this section, we present our design and the proposed DRL-based architecture. Fig. 3 shows the main components of the proposed DRL framework, and the description of each component is presented in the following session.

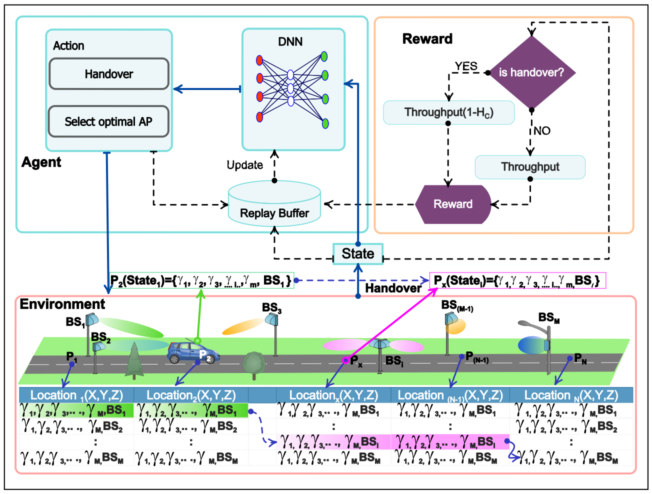

Action
Handover
Select optimal AP
DNN
Agent
Update
Replay Buffer
Reward
YES
is handover?
Throughput(1-H$_C$)
NO
Throughput
Reward
State
$P_2(State_1)=\{\gamma_1, \gamma_2, \gamma_3, \gamma_{...i...}, \gamma_m, BS_1\}$
Handover
$P_x(State_i)=\{\gamma_1, \gamma_2, \gamma_3, \gamma_{...i...}, \gamma_m, BS_i\}$
Environment
BS_1
BS_2
BS_3
$BS_{(M-1)}$
BS_M
P_1
P_2
P_x
BS_i
$P_{(N-1)}$
P_N
Location $_1$(X,Y,Z)
Location$_2$(X,Y,Z)
Location$_x$(X,Y,Z)
Location $_{(N-1)}$(X,Y,Z)
Location $_N$(X,Y,Z)
$\gamma_1, \gamma_2, \gamma_3, ..., \gamma_M, BS_1$
$\gamma_1, \gamma_2, \gamma_3, ..., \gamma_M, BS_2$
$\gamma_1, \gamma_2, \gamma_3, ..., \gamma_M, BS_i$
$\gamma_1, \gamma_2, \gamma_3, ..., \gamma_M, BS_M$

Fig. 3. DRL-based Framework Comprising Environment, States, Actions, and Rewards.

a) Agent: An agent is an entity that can interact with the environment. It observes the state of the environment, takes action and receives the consequence of the action taken. For this problem, we model the agent as a BS controller, and the reason for doing this is because the DRL model requires training resources. The BS controller is chosen because it possesses resource in terms of time, computation power, data set, and, more crucially, the entire network's global information consisting of mm-wave BSs. It should also be noted that the UE collects the input state features in the measurement report (MR) and shares them with the agent.

b) Action: In the HetNet, the association strategy between UE and BS mainly depend on the HO events A1-A6 [23]. However, always choosing the target BS with the highest SNR or RSRP lead to the sub-optimal decision. The wireless environment's dynamic nature is correlated with mobile and stationary obstacles, the presence of several nearby mm-wave BS, and signal fluctuation due to path loss. These factors increase the number of HOs for mobile UE unless appropriately handled. Fig. 3 shows M mm-wave BS, and arbitrary UEs, moving from point P$_1$ to P$_N$, and in each point, P$_x$(X, Y, Z) is in cartesian coordinates. Intuitively, there are more than one BSs that if the UE connects to it, it can prolong UE connectivity with fewer HOs and guarantee maximum user throughput. Hence, we define the action $a \in A(s)$ as the scalar representation of the serving BS at state s. The action space $A(s)$ includes all BSs along the UE route.

c) State space: The state explains the current condition of the network environment and determines what happens next. For our problem, the state is the UE Cartesian coordinate point P$_x$. However, due to the difficulties involved in localizing mobility location, SNR is chosen instead to represent Point P$_x(X, Y, Z)$. We consider SNR received from all BSs at Point P$_x$ to represent location P$_x$ instead of actual P$_x$ in Cartesian coordinates. Logically, the combination of SNRs from BSs is unique continuous values that are the same as point P$_x$ in the Cartesian coordinates throughout the UE route. Therefore, we can relate UE's current position to a combination of BSs SNR values. The advantage of SNR is that UE always receives MR containing accurate SNR from the serving and neighbouring BSs, and we can use this potential information.

Hence, at point P$_x$, the state space for an arbitrary UE is given as, $s = \{\gamma_1, \gamma_2, \gamma_3 ......\gamma_m, BS_{i\in m}\}$ where γ_i is the SNR of BS i, i is the index variable in m BS, and BS$_{i\in m}$ is a serving BS index in one-hot encoded vector. One-hot encoding [24] is the vector transformation of an integer variable into the binary value of zeros except for the index of the integer. For instance, if the serving Bs index at point P$_x$ is BS$_{i=3}$ and there are a total of five BSs $m = 5$, hence, it's equivalent one-hot encoding vector become $BS_{(i=3)} = [0, 0, 1, 0, 0]$.

d) Reward Design: The reward is an abstract term reflecting environmental feedback. The importance of reward is to motivate the agent to learn to reach the target through reward maximization, and our goal is to maximize UE throughput while minimizing HOs. It is also essential to design the reward in such a way that it avoids giving delayed rewards since it may cause the so-called credit assignment problem [20], [21]. We introduce an immediate reward function estimating the immediate impact of the action taken to achieve the agent's target. We design the immediate reward so that the number of HOs and instantaneous received SNR value are combined. We derive the reward from the throughput equation as follows: The instantaneous throughput can be expressed as:

$$\mathbb{T} = \frac{B}{N} \times \log_2(1 + \mathcal{SNR}_i) \tag{1}$$

where B is the maximum bandwidth allocation per serving BS, N is the total number of UEs connected to the BS, and $\mathcal{SNR}_i$ is received SNR from serving BS$_i$. The reward is obtained by incorporating the impact of HO cost to eqn. 1. Hence, the reward can be expressed as:

$$r\left(s_{t+1}, a, s_t\right) = \begin{cases} \mathbb{T}\left(1 - \mathcal{H}_c\right), & \text{if HO occurs} \\ \mathbb{T}, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{H}_c$ is the HO cost [25] which is a unit-less quantity that is used to measure the fraction of time without useful transmission of data along the user's trajectory due to the transfer of HO signalling and the switching of radio links between serving and target BSs.

For model to work, we assume that the average SNR represents the long term experienced SNR at a particular point and that the agent uses these accurately collected SNR values to calculate the reward. We also assume the time delay values of 2 sec per HO for UE's HO from mm-wave BSs to mm-Wave BSs and 0.7 sec per HO for UE's HO mm-wave BSs to sub-6GHz BSs and vice versa [12].

1) Learning algorithm: Fig. 3 shows the proposed model framework built DQN algorithm, summarized in Algorithms 1. In this Algorithm 1, the first thing the agent does is to observe the type of service and if the SNR received from the serving BS is greater than the threshold then it maintains the serving BS else agent decides by taking action a following the ϵ-greedy policy. For a moving UE in particular , at position p, the UE takes action a according to the stated policy $\pi_\theta(s, a)$. Then, after one step of UE $p+1$, the environment generates the next state s_{P+1}. The experienced transition (s,a,r) is stored in the replay memory $\mathbb{D}$, after which the UE receives the next state (s_{p+1}) and perform action a_{p+1} determine by π_θ, and process continue until it reaches terminal state.

V. Performance Evaluation

This section evaluates the proposed DRL-based algorithm's performance, but first, we describe the simulation set-up and parameters and then presenting the simulation results and discussions. We also compare the performance of the proposed DRL model and with the benchmark HO policy [23], which is rate based HO (RBH) strategy.

A. Simulation Setups

The environment, agent and reward are constructed as follows: The environment is constructed using ray tracing simulator WI, and states that are obtained from the environment consist of different number of BSs ranging from 10 - 70 BSs, random obstacle, the random walking model for UE with speed 1 - 10 ms^{-1} and UE's trajectories is of length 500 m length. Python with Keras library and TensorFlow framework was used to implement the agent, and reward is generated based on throughput as expressed in Eqn 2. The summary of the simulation parameters is presented in Table I. In addition, the hyper-parameters used in the implementation of the DQN are shown in the Table. II.

B. Results

The user's velocity was set to 8 ms^{-1}, and 10 mm-wave BSs were considered in the first experiment. Also, the SNR threshold values considered is within the range of 1 dB and 7 dB. We analyse the relationship between the number of HOs and the threshold SNR, which is the UE triggering condition to HO. Fig. 4 shows the different values of the minimum SNR against the number of HO. From the figure, it can be clearly observed that the proposed model outperforms the RBH. The minimum HO reduction gain is seen when the threshold SNR is 7. The trend shows that for any SNR, the proposed DQN based model outperforms RBH. Overall, the proposed DQN model resulted in a 70% HO reduction compared to the benchmark RBH method.

Algorithm 1: Deep Q-Learning

```
Initialize replay memory $\mathbb{D}$ to capacity $\mathbb{N}$;
Initialize action-value function $Q$ with random weight $\theta$;
Initialize the target action-value function $\hat{Q}$ with weight $\theta^- = \theta$
Initialize the target $Q$-network replacement frequency $f_u$;
Repeat:
Get Initial state
Assign terminal state $\leftarrow$ False
Repeat The agent observes the state:
if SNR of Serving BS$_s \geq$ minimum SNR for service $\mathcal{C}_i$ then
    Action: $\leftarrow$ Index of serving BS$_s$;
else
    Action: $\leftarrow$ agent takes an action following $\epsilon$-greedy policy;
end
The agent observe new state $s_{p+1}$ after UE move from point $p$ to another point $p+1$
From action $a(p)$ taken above, calculates the immediate reward $r(s(p), action(p))$ in position $p$
The agent stores all new experiences $(s(p), a(p), r(p), s(p+1), terminalstate)$ into the replay memory $\mathbb{D}$
Agent run experience replay once every $f_u$ steps;
Sample random mini-batch of $\mathbb{Z}$ experience $(s(p), a(p), r(p), s(p+1),$ terminal state) from the reply memory $\mathbb{D}$;
set $y_s = \begin{cases} r_{s(p)}, & \text{for terminal } s(p+1) \\ r_{s(p)} + \gamma\ max_{a'}\ Q(s, a'; \theta), & \text{otherwise} \end{cases}$
Agent performs a gradient descent step on $(y_j - Q(s(p), a(p); \theta))^2$
The agent updates the DQN wight $\theta$ once every $\mathbb{C}$ ;
Every $\mathbb{C}$ step reset $\hat{Q} = Q$, i.e $\theta^- = \theta$;
```

For the second experiment, we evaluate the running time for the two methods, as shown in Fig. 5. The parameters in this experiment are as follows: UE velocity = 8 ms^{-1}, and γ_{th} = 20 dB. Fig. 5 shows that all the policies follow a similar trend. It can be observed that our proposed model takes a longer time than RBH to decide the BS to HO the UE. This is because the proposed model considers more parameters when making a HO decision than the RBH method. Moreover, there is a linear relationship between increasing the number of mm-wave BS and running time for both policies.

Finally, we evaluate the proposed model's performance in terms of the number of HOs and throughput at different UE velocities in the last experiment. The experimental parameters are set as follows: γ_{th} = 20 dB, λ = 50 BSkm^{-2}, and UE velocity = 8 ms^{-1}. The average system throughput and the number of HOs for both HO management policies against the

TABLE I. Simulation Parameters

Parameter	**Value**
BS intensity	10 - 70 (BS/km^2)
mm-wave frequency	28 GHz
mm-wave bandwidth	1 GHz
BS transmit power	30 dBm
Thermal noise density	−174 dBm/Hz
Delay without data transmission	0.75, 2 sec

TABLE II. Design Parameters for the Developed DQN Model

Parameter	**Value**
Hidden layers, Neuron size	6, {32, 64, 128, 256, 64}
Activation function hidden layers	relu
Activation function output layer	linear
Initial exploration training	1
Final exploration training	0.2
Learning rate, α and Discount Factor, γ	0.01 , 0.9
Mini-batch size $\mathbb{C}$, Optimizer	32, Adam
Replay memory size, $\mathbb{D}$	10000

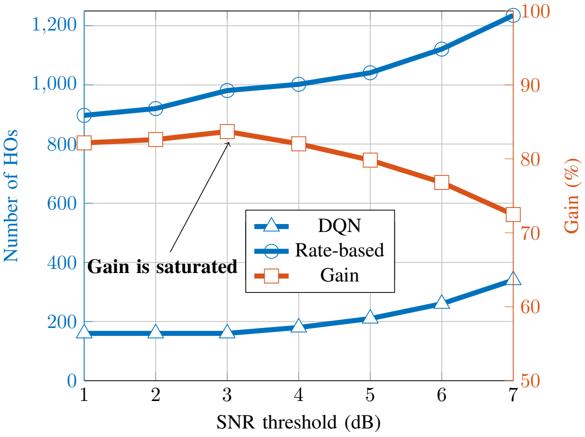


Fig. 4. Number of HO Against Different SNR Threshold.

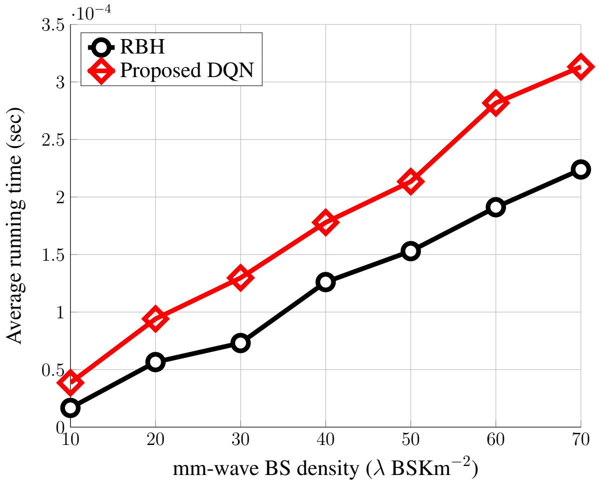


Fig. 5. Average Running Time as a Function of Number of mm-wave BS.

UE velocity are shown in Fig. 6. Fig. 6(a) shows a slight and gradual increase in the number of HOs for both models; however, the proposed DQN model outperforms the RBH policy. Compared to low-speed UE, the effect of HO on the average throughput is more significant for high-speed UE, as seen in Fig. 6(b). Nevertheless, in comparison to RBH, our model proposed performs better.

VI. Conclusion

Mm-wave BS deployment will become ever denser with the emergence of new 5G use cases that demand high data rate. Using mm-wave for communication between UE and BS leads to more HOs for arbitrary UE, and deploying dense mm-wave BSs increases the problem. This paper presents a DQN based model that smartly learn how to maximum UE throughput while minimizing HO's effect. The proposed DQN model and the benchmark rate based HO mechanisms are simulated, and their comparative performance analysis has been performed based on throughput and the number of HOs. According to the simulation results, it can be clearly seen that the proposed approach gives more successful results than the traditional approach in terms of throughput and number of HO occurrences. A new HO strategy that can learn by feeding various state features such as images will be presented in the future. Moreover, the idea of sharing the learnt strategy with the UEs in the learning phase in order to fasten the training process will be considered in the ultra-dense 5G network environment.

Acknowledgment

The authors would like to thank the African development Bank (AfDB) for funding this work.

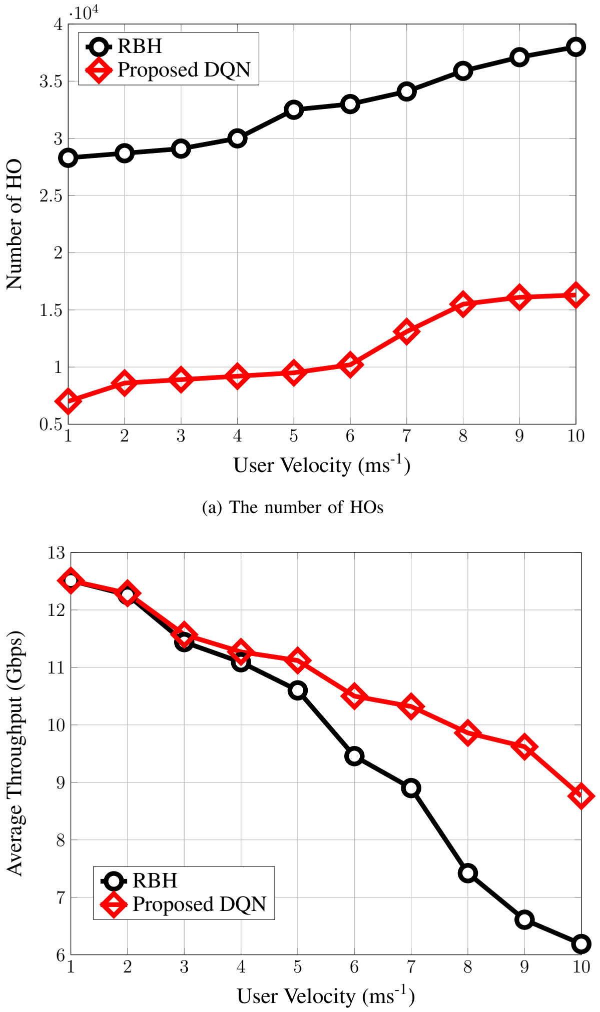
References

[1] A. Al-Dulaimi, X. Wang, and C. L. I, *5G Communication System: A Network Operator Perspective*. Wiley, 2018, pp. 625–652.

[2] 3GPP, "5G; Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), TS 38.913, Sept. 2018.

[3] S. Lien, S. Hung, D. Deng, and Y. J. Wang, "Efficient ultra-reliable and low latency communications and massive machine-type communications in 5g new radio," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–7.

[4] Ericsson, "Ericsson mobility report," Ericsson, Tech. Rep., Nov. 2018. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-november-2018.pdf



(a) The number of HOs

(b) Average system throughput

Fig. 6. Relationship between HO Performance and UE Velocity.

[5] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[6] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5g: challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78–85, 2016.

[7] S. Dastoor, U. Dalal, and J. Sarvaiya, "Issues, solutions and radio network optimization for the next generation heterogeneous cellular network—a review," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017, pp. 1388–1393.

[8] 3GPP, "5G; NR; Base Station (BS) radio transmission and reception," 3rd Generation Partnership Project (3GPP), TS 38.104, July 2018.

[9] A. Talukdar, M. Cudak, and A. Ghosh, "Handoff rates for millimeter-wave 5G systems," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–5.

[10] E. Ndashimye, N. Sarkar, and S. Ray, "A network selection method for handover in vehicle-to-infrastructure communications in multi-tier networks," *Wireless Networks*, vol. 26, 08 2018.

[11] M. Tayyab, X. Gelabert, and R. Jäntti, "A survey on handover management: From lte to nr," *IEEE Access*, vol. 7, pp. 118 907–118 930, 2019.

[12] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in dense cellular networks: A stochastic geometry approach," in *2016 ieee international conference on communications (icc)*. IEEE, 2016, pp. 1–7.

[13] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From lte to 5g for connected mobility," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 156–162, 2017.

[14] G. Gódor, Z. Jakó, Ádám Knapp, and S. Imre, "A survey of handover management in lte-based multi-tier femtocell networks: Requirements, challenges and solutions," *Computer Networks*, vol. 76, pp. 17 – 41, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614003715

[15] I. Shayea, M. Ergen, M. Hadri Azmi, S. Aldirmaz Çolak, R. Nordin, and Y. I. Daradkeh, "Key challenges, drivers and solutions for mobility management in 5g networks: A survey," *IEEE Access*, vol. 8, pp. 172 534–172 552, 2020.

[16] M. Attiah, M. Isa, Z. Zakaria, M. Abdulhameed, M. Mohsen, and I. Ali, "A survey of mmwave user association mechanisms and spectrum sharing approaches: an overview, open issues and challenges, future research trends," *Wireless Networks*, vol. 26, 05 2020.

[17] S. M. A. Zaidi, M. Manalastas, H. Farooq, and A. Imran, "Mobility management in emerging ultra-dense cellular networks: A survey, outlook, and future research directions," *IEEE Access*, vol. 8, p. 183505–183533, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3027258

[18] M. Bonneau, "Reinforcement learning for 5G handover," 2017.

[19] Y. Sun, G. Feng, S. Qin, Y. Liang, and T. P. Yum, "The smart handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1456–1468, June 2018.

[20] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover control in wireless systems via asynchronous multiuser deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4296–4307, 2018.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[23] 3GPP, "5G;NR;Radio Resource Control (RRC);Protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.331, 10 2018, version 15.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[25] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M. Alouini, "Handover management in 5g and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016.