Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 5937358, 9 pages
https://doi.org/10.1155/2020/5937358

Research Article

Decentralized and Dynamic Band Selection in Uplink Enhanced Licensed-Assisted Access: Deep Reinforcement Learning Approach

Fitsum Debebe Tilahun and Chung G. Kang

School of Electrical Engineering, Korea University, Seoul, Republic of Korea

Correspondence should be addressed to Chung G. Kang; ccgkang@korea.ac.kr

Received 15 November 2019; Revised 27 January 2020; Accepted 11 February 2020; Published 27 March 2020

Academic Editor: Mostafa Z. Chowdhury

Copyright © 2020 Fitsum Debebe Tilahun and Chung G. Kang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Enhanced licensed-assisted access (eLAA) is an operational mode that allows the use of unlicensed band to support long-term evolution (LTE) service via carrier aggregation technology. The extension of additional bandwidth is beneficial to meet the demands of the growing mobile traffic. In the uplink eLAA, which is prone to unexpected interference from WiFi access points, resource scheduling by the base station, and then performing a listen before talk (LBT) mechanism by the users can seriously affect the resource utilization. In this paper, we present a decentralized deep reinforcement learning (DRL)-based approach in which each user independently learns dynamic band selection strategy that maximizes its own rate. Through extensive simulations, we show that the proposed DRL-based band selection scheme improves resource utilization while supporting certain minimum quality of service (QoS).

1. Introduction

The rapid mobile traffic demand has resulted in the scarcity of the available radio spectrum. To meet this ever-increasing demand, extending systems like long-term evolution (LTE) to unlicensed spectrum is one of the promising approaches to boost users' quality of service by providing higher data rates [1]. In this regard, initiatives such as the licensed-assisted access (LAA) [2], LTE-unlicensed (LTE-U) [3], and Multe-Fire (MF) systems [4] can be mentioned. The focus of this article is, however, on the LAA system, which 3GPP has initially introduced and standardized in Rel.-13 for downlink operations only [2]. By using the carrier aggregation (CA) technology, carriers on licensed band are primarily used to carry control signals and critical data, while the additional secondary carriers from unlicensed band are used to opportunistically boost the data rates of the users [5]. To obey regional spectrum regulations such as restrictions on the maximum transmitting power and channel occupancy time [6] while fairly coexisting with the existing systems such as WiFi, it is mandatory for an LAA base station (BS) to perform listen before talk mechanism before transmitting over unlicensed band [7–9]. The enhanced version of LAA, named as enhanced licensed-assisted access (eLAA), that supports both uplink and downlink operations was later approved in Rel.14 [10]. The uplink eLAA mode over unlicensed band is designed to meet the channel access mechanisms of the two bands, meaning the BS performs LBT and allocates uplink resources for the scheduled users, and then the scheduled users perform the second round of LBT to check whether the channel is clear or not before uplink transmission [11]. The degradation of uplink channel access due to two rounds of LBT mechanism is investigated in [12–14]. If a scheduled user senses an active WiFi access point (AP) which is hidden to the BS, then the channel cannot be accessed, wasting the reserved uplink resource. Scheduling based approach in uplink eLAA, while there are unexpected interference sources, can significantly affect the utilization of uplink resources.

To improve the utilization of unlicensed band resources, several approaches have been suggested. In [15–17], multi-subframe scheduling (MSS), a simple modification of the conventional scheduling, is proposed. MSS enables a single uplink grant to indicate multiple resource allocation across multiple subframes. Providing diverse transmission opportunities may enhance the resource utilization; however, the resources can still be wasted if the user fails to access the channels. In [14, 18], schemes that switch between random access and scheduling are proposed, but their focus is limited to unlicensed spectrum. Joint licensed and unlicensed band resource allocation that takes a hidden node into account is proposed in [19] for the downlink eLAA system. Furthermore in [20], a scheme that does not require uplink grant along with the required enhancement to the existing LTE system is proposed.

In this paper, we attempt a new learning approach in which each user makes dynamic band selection (licensed or unlicensed) independently for uplink transmission, without waiting for scheduling from BS. To this end, we implemented each user as a DRL agent that learns the optimal band selection strategy relying only on its own local observation, i.e., without any prior knowledge of WiFi APs' activities and time-varying channel conditions. Through continuous interactions with the environment, the potential users to be affected by hidden nodes learn the activities of WiFi APs and make use of it in the band selection process. The learned policy not only guarantees channel access but also ensures a transmission rate above a certain threshold, despite the presence of unpredictable hidden nodes. Such a learning approach would be a useful means of handling the underlying resource utilization problems in uplink eLAA.

The rest of the paper is organized as follows. Section 2 describes the system model considered in the paper. Section 3 gives a brief overview on deep reinforcement learning (DRL), followed by DRL formulation of the band selection problem. The proposed deep neural network architecture and training algorithm are also discussed. Simulation results are presented in Section 4, and finally conclusion is drawn in Section 5.

2. System Model

We consider a single cell uplink eLAA system that consists of an eLAA base station (BS) and N user equipment (UE) that can also operate in unlicensed band through carrier aggregation technology. Let $\mathcal{N} = \{1, 2, \cdots, N\}$ denote a set of user indices which are uniformly distributed within the cell and $\mathcal{M} = \{1, 2, \cdots, M\}$ designate a set of unlicensed band interference sources such as WiFi access points (APs) which are located outside the coverage area of the cell within a certain distance. The system model is shown in Figure 1.

In order to get uplink access, each UE $n \in \mathcal{N}$ makes a scheduling request to the eLAA BS, who is responsible for allocating resources. Before granting uplink resources, the eLAA BS is required to undergo a carrier-sensing procedure within its coverage limit. Once the channel is clear, it reserves resources for uplink transmission. Then, the scheduled user performs another round of listen before talk procedure before transmission. If the user detects transmission from hidden nodes, nearby WiFi APs that are outside the carrier-sensing range of the eLAA BS, then the reserved uplink resources over unlicensed band cannot be accessed.

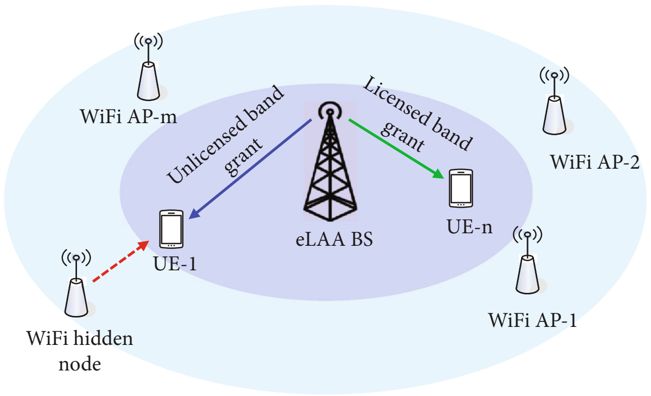

Figure 1: Uplink eLAA system model.

We assume the channel between the BS and the n-th UE, denoted as $h_n(t)$, evolves according to the Gaussian Markov block fading autoregressive model [21] as follows:

$$h_n(t) = \rho_n h_n(t-1) + \sqrt{1-\rho_n^2} e(t), \tag{1}$$

where ρ_n is the normalized channel correlation coefficient between slot t and $(t-1)$. From Jake's fading spectrum, $\rho_n = J_o(2\pi f_{d,n} \tau_o)$ where $f_{d,n}$, τ_o, and $J_o(\cdot)$ are the Doppler frequency, slot duration, and the zeroth-order Bessel function of the first kind, respectively. The error $e(t)$ is a circularly symmetric complex Gaussian variable, i.e., $e(t) \sim \mathcal{CN}(0, \Upsilon (d/d_o)^\alpha)$, where Υ is the path loss corresponding to the reference at a distance d_o and α is the path loss exponent. The channel is initialized as $h_n(0) \sim \mathcal{CN}(0, \Upsilon(d_n/d_o)^\alpha)$, where d_n is distance of the n-th user from the BS.

Let W_U and W_L be the total bandwidth in unlicensed and licensed bands, respectively. At time slot t, let the number of users associated with unlicensed and licensed band be $N_U(t)$ and $N_L(t)$, respectively. If all UEs on licensed band are uniformly allocated to orthogonal uplink resources, then the bandwidth of the UEs is constrained as

$$B_L(t) = \frac{W_L}{N_L(t)}. \tag{2}$$

Similarly, expecting that the total unlicensed bandwidth is equally shared among UEs in a virtual sense, then the bandwidth of UEs on unlicensed band can be constrained as

$$B_U(t) = \frac{W_U}{N_U(t)}. \tag{3}$$

Denoting P and N_0 as uplink transmit power and the noise spectral density, we may compute the signal-to-noise ratio (SNR) of the received signal at the BS for unlicensed band user n (assuming it occupies channel) as

$$\text{SNR}_{n,U}(t) = \frac{P|h_n(t)|^2}{B_U(t) \cdot N_0}. \tag{4}$$

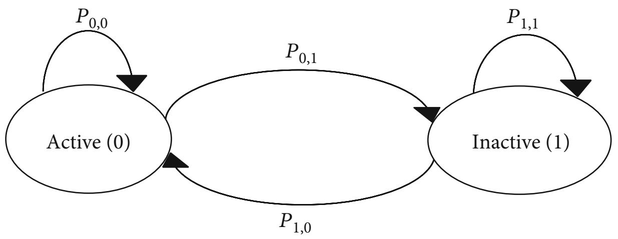

Figure 2: Activity model of WiFi AP as a two-state Markov chain.

Likewise, the SNR for licensed band user n is given as

$$\mathrm{SNR}_{n,L}(t) = \frac{P|h_n(t)|^2}{B_L(t) \cdot N_0}. \quad (5)$$

The dynamics of each WiFi APs activity are modeled as a discrete-time two-state Markov chain as shown in Figure 2. Each AP can be either in active (state = 0) or inactive (state = 1) state. The transition probability from state j to k is denoted as

$$P_{j,k} = \Pr\{s_{t+1} = k \mid s_t = j\}, \forall j, k \in \{0, 1\}. \quad (6)$$

Note that the users do not have the knowledge of the underlying dynamics of WiFi APs' activities, i.e., transition probabilities.

Let τ represent the transmission probability of an active WiFi AP. In slot t, let $N_{n,\mathrm{cont}}(t)$ be the number of contending active APs within the sensing range of n-th UE. Assuming that all activities of WiFi AP's are independent, the probability of UE n having at least one hidden node is

$$P_{n,\mathrm{hid}}(t) = 1 - (1 - \tau)^{N_{n,\mathrm{cont}}(t)}. \quad (7)$$

In order to calculate the uplink rate (throughput) of the users, we refer to the lookup table, given in Table 1, which maps the received SNR to spectral efficiency (SE) [22]. Then, the uplink rate of UE n using unlicensed band is given as

$$R_{n,U}(t) = B_U(t)\,\mathrm{SE}(t)(1 - P_{n,\mathrm{hid}}(t)). \quad (8)$$

Similarly, the uplink rate of UE n using licensed band is given as

$$R_{n,L}(t) = B_L(t)\,\mathrm{SE}(t). \quad (9)$$

In each time slot t, the goal of each UE is to select the band that maximizes the uplink rate. Note that if a certain band, e.g., licensed band, is overloaded by a large number of UEs, the individual rate of the users in the band will be significantly reduced. This will constraint each UE to take advantage of the unlicensed band whenever the APs are inactive. Hence, learning the WiFi APs' activities and channel conditions is critical to effectively use the uplink resources while boosting individual data rate.

Table 1: Lookup table for SNR-to-spectral efficiency mapping.

Index	Minimum SNR (dB)	Spectral efficiency (bps/Hz)
1	−6.7	0.1523
2	−4.7	0.2344
3	−2.3	0.3770
4	0.2	0.6016
5	2.4	0.8770
6	4.3	1.1758
7	5.9	1.4766
8	8.1	1.9141
9	10.3	2.4063
10	11.7	2.7305
11	14.1	3.3223
12	16.3	3.9023
13	18.7	4.5234
14	21.0	5.1152
15	22.7	5.5547

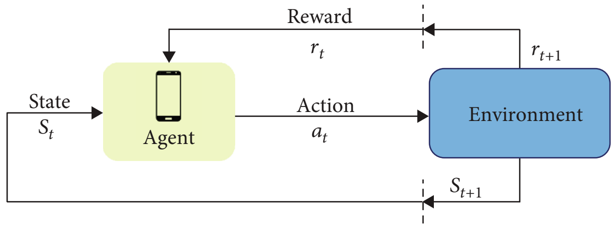

Figure 3: Reinforcement learning framework.

3. DRL-Based Decentralized Dynamic Band Selection

3.1. Deep Reinforcement Learning (DRL): Overview. In reinforcement learning (RL), an agent learns how to behave by sequentially interacting with the environment. As shown in Figure 3, at each time t, the agent observes the state $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the state space, and executes action $a_t \in \mathcal{A}$ from the action space $\mathcal{A}$. The interaction with the environment produces the next state s_{t+1} and scalar reward r_{t+1}.

The goal of the agent is to learn an optimal policy that maximizes the discounted long-term cumulative reward, expressed as

$$R_t = \sum_{t-1}^{T} \gamma^{t-1} r_{t+1}, \quad (10)$$

where $\gamma \in [0, 1]$is the discounting factor and T is the total number of time steps (horizon) [23].

One of the most widely used model-free RL methods is Q-learning in which the agent learns policy by iteratively evaluating the state-action value function $Q(s, a)$, defined

```
Initialize replay buffer $\mathscr{D}$
Initialize action value function $Q$ with parameter $\boldsymbol{\theta}$
Initialize target action value function $\widehat{Q}$ with parameter $\boldsymbol{\theta}' = \boldsymbol{\theta}$
Input the initial state to the DQN
for $t = 1, 2, .. \cdots$ do
    Execute action $a_t$ from $Q$ using $\varepsilon$-greedy policy
    Observe $r_{t+1}$ and $s_{t+1}$ from the environment.
    Store the transition $(s_t, a_t, r_{t+1}, s_{t+1})$ into the replay buffer $\mathscr{D}$
    Sample random minibatch of transitions from $\mathscr{D}$
    Evaluate the target $y_j = r_j + \gamma \max_{\mathbf{a}'} \widehat{Q}(s_{j+1}, a' ; \theta')$
    Perform a gradient descent step on $(y_j - Q(s_j, a_j ; \theta))^2$ with respect to $\boldsymbol{\theta}$
    Every $C$ steps, update the target network $\widehat{Q}$ according to $\boldsymbol{\theta}' \longleftarrow \boldsymbol{\theta}$
end for
```

Algorithm 1. DQN algorithm.

as the expected return starting from the state s, taking the action a, and then, following the policy π. In order to derive the optimal policy, at a given state s, the action that maximizes the state-action value function should be selected, i.e.,

$$a * (s) = \arg \max_{a} Q(s, a) \tag{11}$$

and then similarly follow optimal actions in the successor states.

In Q-learning, a lookup table is constructed that stores the action value $Q(s, a)$ for every state-action pair (s, a). The entries of the table are updated by iteratively evaluating the Bellman optimality equation as:

$$\begin{aligned} Q(s_t, a_t) \longleftarrow & \; Q(s_t, a_t) \\ & + \beta\left[r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right] \end{aligned} \tag{12}$$

where $\beta \in [0, 1]$ is the learning rate. However, the look up table approach in Q-learning is not scalable for problems with the large state and action spaces. DRL approximates the value functions with deep neural network (DNN) instead. In deep Q-network (DQN), the action-value function $Q(s, a ; \theta)$ is estimated by DNN, parametrized by θ, which takes the state as input. Then, action is selected according to the following ε-greedy policy:

$$a_t = \begin{cases} \text{random action}, & \text{with a probability of } \varepsilon, \\ \arg \max_{a} Q(s, a ; \theta), & \text{with a probability of } (1-\varepsilon). \end{cases} \tag{13}$$

To stabilize the learning process, it is common to use a replay buffer $\mathscr{D}$ that stores transitions $e = (s_t, a_t, r_{t+1}, s_{t+1})$ and mini batch of samples are randomly drawn from the buffer to train the network. Moreover, a separate quasi-static target network, parametrized by θ', is used to estimate the target value of the next state. The loss function is computed as

$$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{(s,a,r,s')\sim\mathscr{D}}\left[\left(r + \gamma \max_{\mathbf{a}'} Q\left(s', a' ; \theta'\right) - Q(s, a ; \theta)\right)^2\right] \tag{14}$$

θ is updated by following stochastic gradient of the loss as $\theta \longleftarrow \theta - \beta\nabla\theta L(\theta)$, while the target parameter θ' is updated according to $\theta' \longleftarrow \theta$ every C steps [24]. The details of DQN algorithm is summarized in Algorithm 1.

3.2. DRL Formulation for Dynamic Band Selection. Each user is implemented as DRL, specifically by deep Q-network (DQN) agent that relies on the output of their deep neural network to make dynamic band selection decisions between licensed and unlicensed bands. The DRL formulation is presented below.

(i) Action

In each time slot t, the n-th agent samples an action $a_n(t)$ from the action set

$$\mathscr{A} = \{\text{Licensed, Unlicensed}\}. \tag{15}$$

(ii) State

After executing the action $a_n(t)$, the agent receives binary observation and reward from the environment. The observation is either $o_n(t) = 1$ if the uplink rate in the selected band exceeds the minimum threshold rate or $o_n(t) = 0$ otherwise. The state of the agent is defined as history of an action-observation pairs with length H:

$$s_n(t) \triangleq \{(a_n(i), o_n(i))\}_{i=t-H}^{t} \tag{16}$$

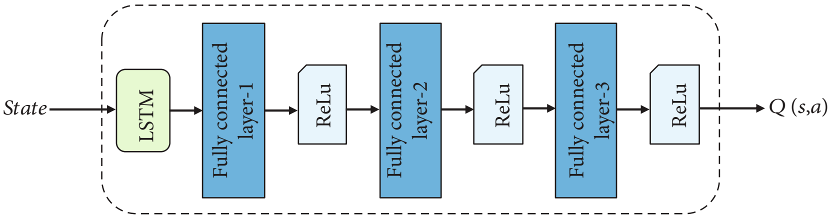

Figure 4: Structure of the proposed deep neural network.

(iii) Reward

Depending on the selected action, the agent receives the following scalar reward:

$$\text{If } a_n(t) = \begin{cases} \text{Unlicensed, } r_{t+1}^n = \begin{cases} R_{n,U}(t), & \text{if } R_{n,U}(t) \geq R_{U,\min} \\ 0, & \text{otherwise} \end{cases} \\ \text{Licensed, } r_{t+1}^n = \begin{cases} R_{n,L}(t), & \text{if } R_{n,L}(t) \geq R_{L,\min} \\ 0, & \text{otherwise} \end{cases} \end{cases} \quad (17)$$

where $R_{n,U}(t)$ and $R_{n,L}(t)$ are given according to Equations (8) and (9), while $R_{U,\min}$ and $R_{L,\min}$ are the uplink minimum threshold rates on unlicensed and licensed band, respectively.

3.3. Deep Neural Network Description. For dynamic band selection, each UE trains independent DQN. The structure of the deep neural network is shown in Figure 4.

The deep neural network consists of long short-term memory (LSTM) layer, fully connected layers, and rectified linear unit (ReLu) activation function.

Long short-term memory (LSTM) is one class of recurrent neural networks (RNNs) which are designed to learn a specific pattern in a sequence of data by taking time correlation into account. They were initially introduced to overcome the vanishing (exploding) gradient problem of RNNs in the course of back propagation. Regulated by gate functions, the cell (internal memory) state of an LSTM learns how to aggregate inputs separated by time, i.e., which experiences to keep or throw away [25]. In our formulation, note that the states of the agents, which are histories of action-observation pairs, have long-term dependency (correlation) emanating from the dynamics of WiFi APs' activities that follow a two-state Markov property, and the time-varying channel conditions according to Gaussian Markov block-fading autoregressive model. LSTM is crucial for the learning process since it can capture the actual state by exploiting the underlying correlation in the history of action-observation pairs. Therefore, the state must pass through this preprocessing step before it is directly fed to the neural network.

A deep neural network consists of multiple fully connected layers, in which each of the layers abstracts certain feature of the input. Let $\mathbf{x}$ be the input to the layer, while $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias vector, respectively. The output vector of a layer, denoted as $\mathbf{y}$, in a fully connected layer can be described by the following operation:

$$\mathbf{y} = f(\mathbf{Wx} + \mathbf{b}), \quad (18)$$

where f is the element-wise excitation (activation) that adds nonlinearity. In our simulations, we input the states to an LSTM layer with hidden units of 64, whose output is fed to two fully connected hidden layers with 128 and 64 neurons. The output layer produces action values $Q(s, a)$ for both actions. ReLu activation function is used on all the layers to avoid the vanishing gradient problem [26]. The target network also adopts the same neural network structure.

3.4. Training Algorithm Description. The DQNs of the agents are individually trained according to Algorithm 2. The loss function given by Equation (14) is used to train the DQN. The hyperparameters are summarized in Table 2.

Note that the agents do not have a complete knowledge of the environment, such as the action of other agents, the underlying dynamics of the WiFi APs' activities, and varying channel conditions. Instead, through sequential interaction with the environment, each agent makes decisions on band selection solely based on local feedbacks (reward and observation) from the base station. This significantly reduces the training complexity (cost) at each user. Moreover, since the training can be conducted in an offline manner, the trained weights can be used in deployment phase. Retraining the weights is done infrequently; for example, if the environment significantly changes.

4. Simulation Results

4.1. Simulation Setup. For each realization, we first distribute 10 users uniformly in a square area of 100 m × 100 m. Within 30 m distance from the coverage area of the cell, WiFi APs are distributed in homogeneous Poisson point process (PPP) with rate λ. Figure 5 illustrates the network model of one realization for node deployment of BS, users, and APs.

We set the dynamics of each WiFi AP activity according to the following transition matrix:

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}, \quad (19)$$

```
for each agent $n \in \mathcal{N}$ do
    Initialize replay buffer $\mathcal{D}_n$
    Initialize action value function $Q_n$ with parameter $\boldsymbol{\theta}_n$
    Initialize target action value function $\widehat{Q}_n$ with parameter $\boldsymbol{\theta}'_n = \boldsymbol{\theta}_n$
    Generate initial state $s_{n,1}$ from the environment simulator
end for
for $t = 1, 2, .. \cdots$ do
    for each agent $n \in \mathcal{N}$ do
        Execute action $a_{n,t}$ from $Q_n$ using $\varepsilon$-greedy policy
        Collect reward $r_{n,t+1}$ and observation $o_{n,t+1}$
        Observe the next state $s_{n,t+1}$ from the environment simulator
        Store the transition $(s_{n,t}, a_{n,t}, r_{n,t+1}, s_{n,t+1})$ into $\mathcal{D}_n$
        Sample random minibatch of transitions from $\mathcal{D}_n$
        Evaluate the target $y_{n,j} = r_{n,j} + \gamma \max_{a_{n,j+1}} \widehat{Q}_n(s_{n,j+1}, a_{n,j+1}; \boldsymbol{\theta}'_n)$
        Perform a gradient descent step on $(y_{n,j} - Q_n(s_{n,j}, a_{n,j}; \theta_n))^2$with respect to $\boldsymbol{\theta}_n$
        Every $C$ steps, update the target network $\widehat{Q}_n$ according to $\boldsymbol{\theta}'_n \longleftarrow \boldsymbol{\theta}_n$
    end for
end for
```

Algorithm 2. DQN training algorithm for dynamic band selection.

Table 2: Hyperparameters.

Parameter	Value
Discount factor γ	0.9
Learning rate β	0.01
Exploration ε in ε-greedy policy	0.05 to 0.01
Target network update frequency C	300
Mini batch size	32
Replay buffer $\mathcal{D}$ size	1000

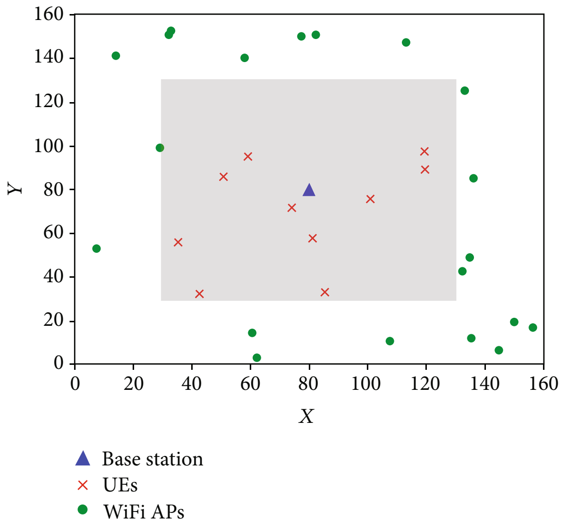

Figure 5: Network model for node deployment (layout).

Table 3: Simulation parameters.

Parameter	Value
Total bandwidth in unlicensed band (W_U)	10 MHz
Total bandwidth in licensed band (W_L)	10 MHz
Uplink transmission power (P)	20 dBm
Receiver noise power (N_0)	−147 dBm
Path loss exponent (α)	3.76
Reference distance (d_o)	1 m
Channel gain at reference distance (Υ)	−35.3 dB
Channel correlation coefficient (ρ_n)	0.95
Doppler frequency ($f_{d,n}$)	70 Hz
Transmission probability of active WiFi AP (τ)	0.7

We further assume that the uplink transmission of a user over unlicensed band can be interfered from any active WiFi AP $m \in \mathcal{M}$ within 30 m range. Table 3 summarizes the values of all simulation parameters used for evaluating the proposed algorithm.

4.2. Performance Evaluation. We compared the policy learned by the DRL agents to two benchmark schemes: random policy and fixed distance policy. In random policy, each user randomly decides which band to select, while in fixed policy, decision is made based on the location of the user. Assuming the BS knows the location of the users at each slot t; hence, the distance from BS, only users within D meters from the base station transmit using unlicensed band resources, since they are less susceptible to interfering WiFi APs. The others transmit using licensed band resources. Since we assumed transmission from a WiFi AP can affect

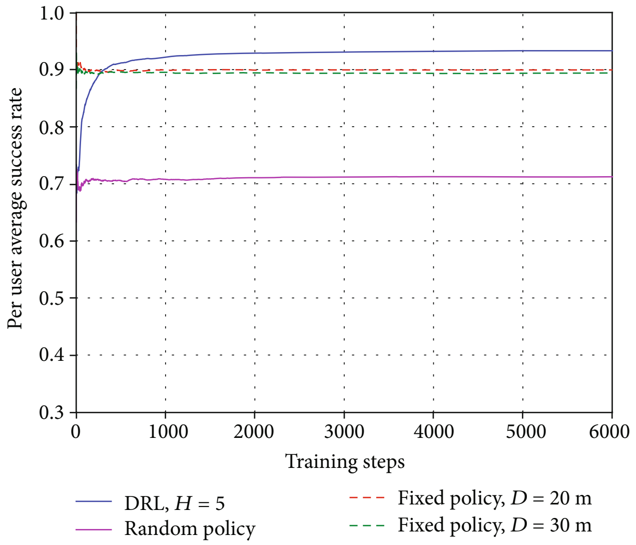

Figure 6: Per user average success rate with the training process.

unlicensed band uplink transmission of any user within 30 m distance, according to the node deployment in Figure 5, in fixed policy users with $D = 20$m from the BS are assigned to unlicensed band resources. The trained DRL policy of each agents should learn this distance without any prior assumption while selecting band. Furthermore, by learning the activities of the APs, the agents should make a dynamic selection.

Figure 6 compares the per user average success rate of the users for different thresholds at history length $H = 5$, $\lambda = 0.5 \times 10^{-2}$, and $R_{L,\min} = R_{U,\min} = 4$Mbps. The dynamic DRL agents entertain around 90% of success rate, outperforming the users of the fixed distance-based policy with all of the thresholds we set. The gain from the fixed distance-based policy is attributed to two factors. The first one is that DRL agents, without any prior assumption, learn the optimal distance d^* from the BS to make a decision on band selection. In other words, if user $n \in \mathcal{N}$ is located outside the optimal distance range $(d_n > d^*)$, then it transmits over licensed band to avoid interference from nearby WiFi APs. The second factor is that the agents capture the dynamics of both time-varying channel and WiFi APs' activities while making use of it in dynamically selecting band. It implies that during the absence of transmission from nearby WiFi APs, even if $(d_n > d^*)$, user $n \in \mathcal{N}$ exploits the opportunity of transmitting over unlicensed band; hence, avoids overloading other users on licensed band.

To further investigate the gain coming from dynamic decision on band selection, we evaluate the per user average success rate of the users for different throughput thresholds in Figure 7. As the threshold values (over both bands) increase from 3 to 5, the gap on performance (per user average success rate) also increases. This indicates that capability of the DRL agents is crucial to maintaining appreciable success rate under a stringent requirement on quality of service (QoS).

In Figure 8, per user average throughput obtained by the three policies for history length $H = 5$, $\lambda = 0.5 \times 10^{-2}$, and $R_{L,\min} = R_{U,\min} = 4$Mbps is compared. As depicted, the per user average throughput achieved by DRL agents outperforms the other two schemes. The ability of DRL to adapt to changing environment and learn robust policy enabled the agents to outperform a fixed distance-based policy which falls short when either of the bands is overloaded. In other words, even if there is an opportunity to transmit on unlicensed band, due to inactivity of nearby WiFi APs, cell edge users in fixed

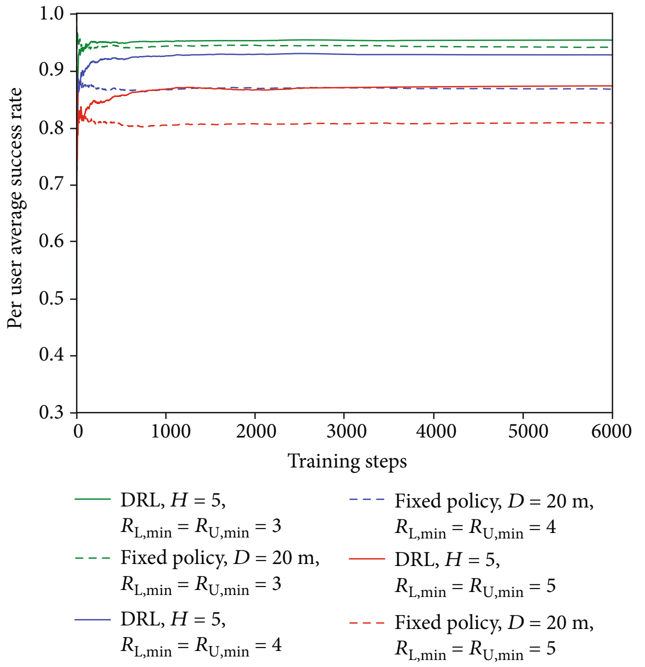

Figure 7: Per user average success rate at varying thresholds with the training process.

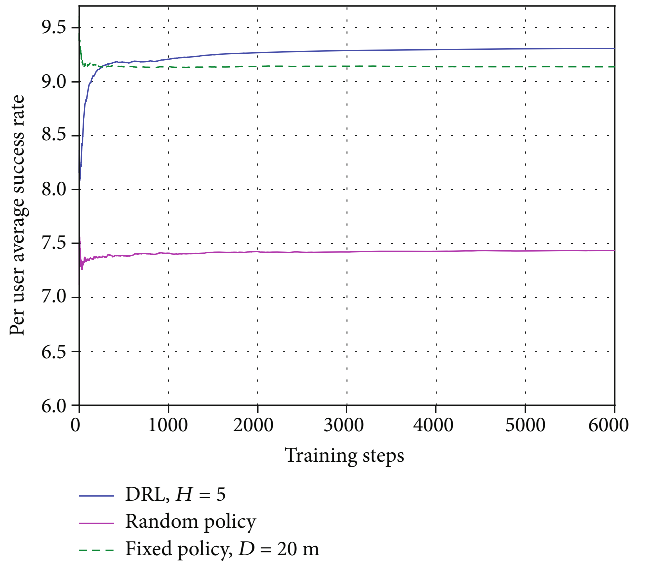

Figure 8: Per user average throughput of the policies with the training process.

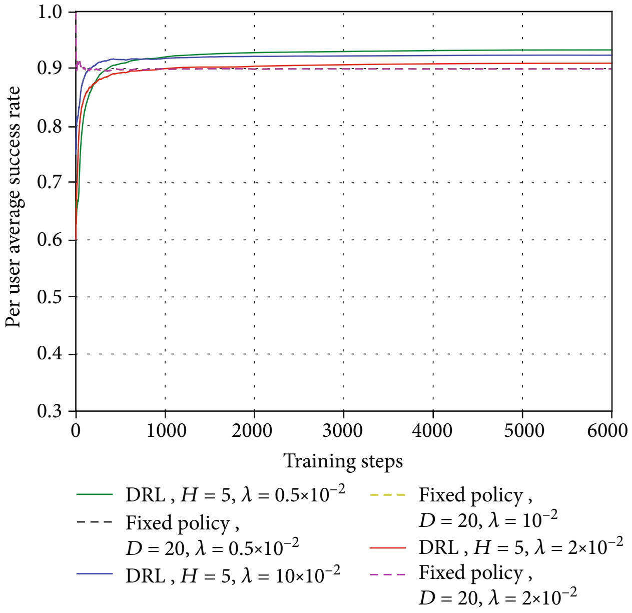

Figure 9: Effect of number of WiFi APs on per user average success rate.

distance-based policy fail to take advantage of it. Further gain can be obtained by tuning the hyperparameters.

The effect of the number of interfering WiFi APs on the performance of the DRL agents is investigated for history length $H = 5$, and $R_{L,\min} = R_{U,\min} = 4$Mbps in Figure 9. As the number of WiFi APs increases (when λ increases), the gain due to dynamic decision on band selection reduces since the number of contenders for unlicensed band resources increases. However, the agents still retain the gain coming from learning the optimal distance for band selection. The performance of the fixed distance-based policy is unaffected by the number of WiFi APs.

Next, in Figure 10, we compare the effect of history size on the performance of the DRL agents. We observe that shorter history sizes tend to converge relatively faster. The variation of convergence time of the learned policy is however marginal. This implies the convergence time of the learned policy is generally less sensitive to history size. Note that all the results are averaged from three numerical simulations.

5. Conclusion and Future Works

To improve the underlying resource utilization problem in uplink eLAA, we presented a learning-based fully decentralized dynamic band selection scheme. In particular, employing the deep reinforcement learning algorithm, we have implemented each user as an agent that makes a decision based on the output of the DQN, without waiting for scheduling from BS. It is shown that despite the lack of the knowledge of the underlying dynamics of WiFi APs' activities, the DRL agents successfully learn a robust policy to make a dynamic decision on band selection. Such dynamic and decentralized learning approach can significantly improve the resource utilization problem associated with unlicensed band, due to hidden nodes, in the uplink eLAA system. In a future study, we want to extend this work to more complicated scenarios that involve joint resource allocation over the two bands. Moreover, to improve the gain presented in this paper, different architectures and hyperparameters should be investigated.

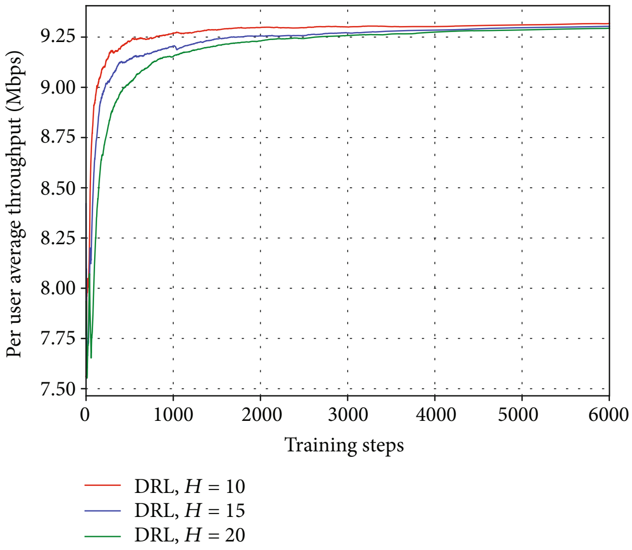

Figure 10: Effect of history size on convergence of the training.

Data Availability

We have not used specific data from other sources for the simulations of the results. The proposed algorithm is implemented in python with TensorFlow library.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Brain Korea 21 Plus Project in 2019.

References

[1] Qualcomm, *Qualcomm research LTE in unlicensed spectrum: harmonious coexistence with WiFi*, Qualcomm, 2014.

[2] 3GPP, *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Licensed Assisted Access to Unlicensed Spectrum;(Release 13)*, 3GPP, 2015, TR 36.889 V13.0.0.

[3] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, X. Shen, and L.-L. Xie, "LTE-unlicensed: the future of spectrum aggregation for cellular networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 150–159, 2015.

[4] "MulteFire Alliance Formed to Bring Enhanced Wireless Performance to Unlicensed Spectrum," https://www.multefire.org/2015/12/16/multefire-alliance-formed-to-bring-enhanced-wireless-performance-to-unlicensed-spectrum/.

[5] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)*, 3GPP, 2010, TR 36.814 v9.0.0.

[6] ETSI, *Broadband Radio Access Networks (BRAN); 5 GHz high performance RLAN*, ETSI, 2014, EN 301 893.

[7] B. Li, T. Zhang, and Z. Zeng, "LBT with adaptive threshold for coexistence of cellular and WLAN in unlicensed spectrum," in *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pp. 1–6, Yangzhou, China, October 2016.

[8] C. K. Kim, C. S. Yang, and C. G. Kang, "Adaptive listen-before-talk (LBT) scheme for LTE and Wi-Fi systems coexisting in unlicensed band," in *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 589–594, Las Vegas, NV, January 2016.

[9] C. S. Yang, C. K. Kim, J. Moon, S. Park, and C. G. Kang, "Channel access scheme with alignment reference interval adaptation (ARIA) for frequency reuse in unlicensed band LTE: Fuzzy Q-learning approach," *IEEE Access*, vol. 6, pp. 26438–26451, 2018.

[10] 3GPP, Huawei, HiSilcon, *Design for frame structure 3 with DL and UL subframes for eLAA*, 3GPP, 2016, R1-162604.

[11] 3GPP, *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 14)*, 3GPP, 2017, TR 36.213, V14.2.0.

[12] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[13] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, Glasgow, UK, May 2015.

[14] Y. Gao, X. Chu, and J. Zhang, "Performance analysis of LAA and WiFi coexistence in unlicensed spectrum based on Markov chain," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, December 2016.

[15] R. Karaki, J.-F. Cheng, E. Obregon et al., "Uplink performance of enhanced licensed assisted access (eLAA) in unlicensed spectrum," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, San Francisco, CA, March 2017.

[16] 3GPP, *LG Electronics, Qualcomm, and ZTE, "Way forward on multi-subframe scheduling in LAA"*, 3GPP, 2016, R1-161409.

[17] 3GPP, Ericsson, *On UL channel access procedures for PUSCH*, 3GPP, 2016, R1-163150.

[18] S.-Y. Lien, J. Lee, and Y.-C. Liang, "Random access or scheduling: optimum LTE licensed-assisted access to unlicensed spectrum," *IEEE Communications Letters*, vol. 20, no. 3, pp. 590–593, 2016.

[19] T. Zhang, J. Zhao, and Y. Chen, "Hidden node aware resource allocation in licensed-assisted access systems," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, December 2017.

[20] J. Zhang, W. Chang, H. Niu, S. Talarico, and H. Yang, "Grantless uplink transmission for LTE operated in unlicensed spectrum," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, Montreal, QC, February 2017.

[21] H. A. Suraweera, T. A. Tsiftsis, G. K. Karagiannidis, and A. Nallanathan, "Effect of feedback delay on amplify-and-forward relay networks with beamforming," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 1265–1271, 2011.

[22] H. Zarrinkoub, *Understanding LTE with MATLAB: from mathematical modeling to simulation and prototyping*, John Wiley & Sons, 2014.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an introduction*, MIT press, 1998.

[24] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT press, 2016.

[26] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, Madison, WI, USA, 2010.