IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2022, accepted 19 September 2022, date of publication 22 September 2022,
date of current version 30 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208572

RESEARCH ARTICLE

A Deep Learning-Based Transmission Scheme Using Reduced Feedback for D2D Networks

TAE-WON BAN (Member, IEEE)
Department of Intelligent Communication Engineering, Gyeongsang National University, Tongyeong 53064, South Korea
e-mail: twban35@gnu.ac.kr

This work was supported by the National Research Foundation of Korea (NRF) through the South Korean Government, Ministry of Education, under Grant 2020R1I1A3061195 (Development of Wired and Wireless Integrated Multimedia-Streaming System Using Exclusive OR-Based Coding).

ABSTRACT In this study, we investigate frequency division duplex (FDD)-based overlay device-to-device (D2D) communication networks. In overlay D2D networks, D2D communication uses a dedicated radio resource to eliminate the cross-interference with cellular communication and multiple D2D devices share the dedicated radio resource to resolve the scarcity of radio spectrum, thereby causing co-channel interference, one of the challenging problems in D2D communication networks. Various radio resource management problems for D2D communication networks can't be solved by conventional optimization methods because they are modelled by non-convex optimization. Recently, various studies have relied on deep reinforcement learning (DRL) as an alternative method to maximize the performance of D2D communication networks overcoming co-channel interference. These studies showed that DRL-based radio resource management schemes can achieve almost optimal performance, and even outperform the state-of-art schemes based on non-convex optimization. Most of DRL-based transmission schemes inevitably require feedback information from D2D receivers to build input states, especially in FDD networks where the channel reciprocity between uplink and downlink is not valid. However, the effect of feedback overhead has not been well investigated in previous studies using DRL, and none of the studies reported on reducing the feedback overhead of DRL-based transmission schemes for FDD-based D2D networks. In this study, we propose a DRL-based transmission scheme for FDD-based D2D networks where input states are built by using reduced feedback information, thereby reducing feedback overhead. The proposed DRL-based transmission scheme using reduced feedback information achieves the same average sum-rates as that using full feedback, while reducing the feedback overhead significantly.

INDEX TERMS Autonomous transmission, device-to-device (D2D), deep reinforcement learning (DRL), transmission scheme, feedback.

I. INTRODUCTION

Recently, device-to-device (D2D) direct communication has been attracting a lot of attention because it can significantly improve spectral efficiency in mobile communication networks by spatially reusing the same radio spectra. D2D direct communication also plays crucial roles for public safety in disaster situations where base stations (BSs) are not available by allowing direct communication between two devices without traversing BSs or core network [1], [2]. D2D communication has already been included to the standards of 3rd generation partnership project (3GPP) [3] and will be able to provide a connectivity for both public safety and various commercial applications such as unmanned aerial vehicles (UAVs) [4], vehicle-to-vehicle (V2V) [5], and Internet-of-things (IoT) [6] without deploying infrastructure. Especially, 5G new radio (NR) radio access network (RAN) supports vehicle-to-everything (V2X) sidelink as a key application of D2D communication [7], [8].

Despite its various advantages, the performance of D2D communication can be seriously degraded by co-channel interference [8]. Performance degradation caused by

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

co-channel interference can be mitigated in the *in-coverage* scenarios where a cellular network can control D2D communication. However, such a control is impossible in the *out-of-coverage* scenarios where D2D devices are located beyond the coverage of cellular networks and the efficiency of D2D communication is thus severely degraded, which is one of the main reasons that various studies have focused on investigating interference in D2D networks [9], [10], [11]. Not only a centralized scheduling algorithm that yields almost optimal sum-rates with a reduced computational complexity, but a distributed algorithm was also proposed in [9]. The distributed scheduling algorithm can select one D2D transmitter with the greatest channel gain, but results in the the waste of radio resource because of a considerable amount of signaling. A game-theoretic framework for optimal mode selection and spectrum partitioning was proposed for D2D-enabled cellular networks, where D2D communication uses a dedicated resource without a cross-interference with cellular communication [10]. Based on an energy and spectral efficiency evaluation framework for large-scale D2D-enabled cellular networks, an optimal mobile traffic offloading scheme for D2D overlay cellular networks was derived by using tractable closed-form expressions for energy and spectral efficiency [11].

On the other hand, deep reinforcement learning (DRL) has been widely used as an alternative approach to solve many mathematically intractable problems in wireless networks [12], [13]. Especially, various challenging problems such as resource allocation and interference mitigation in D2D networks can't be solved by conventional optimization methods because most of them are non-convex [14], and many studies thus relied to DRL to solve non-convex problems in D2D networks [15], [16], [17]. A previous study investigated a joint problem of resource block allocation and power control and proposed a DRL-based scheme to solve the joint problem [15]. A joint problem of channel selection and power control was investigated to maximize weighted sum-rates of an overlay D2D network where multiple D2D pairs share a single channel simultaneously and network performance is degraded because of severe co-channel interference [16]. A distributed DRL-based transmission scheme that allows each D2D pair to make optimal decisions autonomously was proposed, achieving approximate performance of the state-of-art fractional programming-based algorithm. A deep learning based transmit power allocation scheme that can automatically determine the optimal transmit power levels of co-spectrum cellular users and D2D users based on a deep neural network was proposed [17].

Even though DRL-based novel approaches can achieve almost optimal performance or outperform the state-of-art conventional schemes without using non-convex optimization methods, as shown in [15], [16], and [17], they require agents to collect channel information to build input layers. Specifically, the input states in [15] consist of instantaneous signal channel gain and interference channel gains between D2D transmitter and receiver, and interference channel gains between cellular user and D2D receiver. Local channel state information and outdated non-local channel information are only used for input states in [16]. The input layer of the DRL used in [17] is also formed by the channel gain matrix. In time division duplex (TDD)-based D2D networks, each D2D transmitter can easily obtain channel information thanks to the channel reciprocity of uplink and downlink, i.e., each D2D transmitter can estimate downlink channel information based on uplink channel information obtained by measuring uplink sounding symbols transmitted by D2D receivers. In frequency division duplex (FDD)-based D2D networks, each D2D receiver is required to send downlink channel information to D2D transmitters to enable them to build input states because the channel reciprocity of uplink and downlink is not valid, which inevitably results in tremendous feedback overhead.

Despite various studies for D2D networks, to the best of our knowledge, there has been no investigation on feedback reduction for D2D networks because meaningful researches to reduce the feedback overhead have been focused on multi-antenna networks [18], [19]. A method that significantly reduces the required feedback load by utilizing a small number of receive antennas at each mobile was proposed in [18]. On the other way, a transmit antenna selection scheme for downlink transmission in massive antenna systems was proposed to reduce the feedback required by cellular base stations [19]. These schemes can not be applied to D2D networks because of the structural differences between D2D and cellular networks, despite their superiority. Thus, we investigate feedback reduction schemes for FDD-based D2D communication networks. Each agent builds its input states by only exploiting local channel information instead of global channel information to reduce feedback overhead. Furthermore, we also propose two feedback schemes, namely partial feedback scheme and binary feedback scheme, to reduce the feedback overhead further. In the partial feedback scheme, each D2D receiver feeds back its signal channel gain and interference channel gains that are greater than its signal channel gain. In the binary feedback scheme, each D2D receiver feeds back indicators of interference channel gains that are greater than its signal channel gain, instead of the real values of channel gains. Our numerical results show that the partial and binary feedback schemes can both achieve approximately optimal sum-rates in power or interference limited environments.

The main contributions of this paper are summarized as follows. The problem of feedback overhead for D2D networks is formulated and two feedback reduction schemes are proposed to reduce the feedback overhead for D2D networks based on the formulation. In addition, the proposed feedback reduction schemes are incorporated with a new transmission scheme using DRL to prevent performance degradation caused by the reduced feedback. The proposed feedback reduction schemes can significantly reduce the feedback overhead while achieving the same sum-rates, compared to the full feedback scheme. They also enable each D2D to

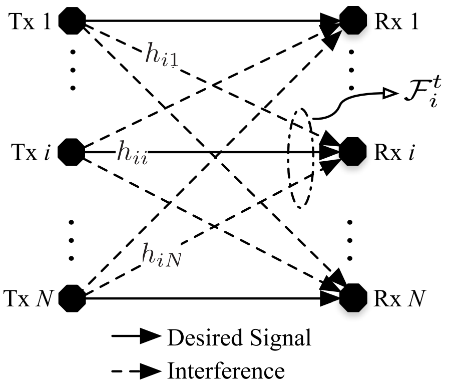


FIGURE 1. An illustrative D2D communication network with K D2D pairs.

determine whether to transmit data in a fully distributed manner.

The rest of this paper is organized as follows. The D2D communication network and channel model considered in this study are described in Sect. II. DRL-based distributed transmission schemes using reduced feedback are proposed in Sect. III. Numerical results are shown in Sect. IV and the conclusions of this paper are finally drawn in Sect. V.

II. A D2D COMMUNICATION NETWORK AND CHANNEL MODEL

Fig. 1 depicts an overlay D2D communication network that consists of K D2D transmitters and K D2D receivers. Each receiver is associated with a transmitter. We assume that the D2D communication uses a dedicated frequency spectrum reserved for D2D communication and all D2D devices share the frequency spectrum. We use frequency division duplex (FDD) as a duplex scheme, thereby allocating half of the total frequency to uplink and downlink, respectively. h_{ij}^t denotes the channel coefficient from transmitter j to receiver i at a certain transmission interval t, where $i, j \in \{1, 2, \cdots, K\}$. $h_{ij}^t \neq h_{ji}^t$ because channel reciprocity is not valid in FDD. h_{ij}^t's follow the complex Gaussian distribution$\sim \mathcal{CN}(0, 1)$ and $|h_{ij}^t|^2$'s are thus exponentially distributed with unit mean. We also assume that all channel coefficients are independent and identically distributed (*i.i.d.*). They are static for a transmission interval but vary randomly every transmission interval. All transmitters use a fixed transmit power P. A binary symbol $a_i^t \in \{0, 1\}$ determines whether each D2D transmitter i transmits data, and is the action of transmitter i. Then, the data rate of the D2D pair i is calculated as

$$\begin{aligned} c_i^t &= \log_2\left(1 + \frac{Pa_i^t|h_{ii}^t|^2}{\sum_{j=1,j\neq i}^K Pa_j^t|h_{ij}^t|^2 + N_0}\right) \\ &= \log_2\left(1 + \frac{\gamma a_i^t|h_{ii}^t|^2}{\sum_{j=1,j\neq i}^K \gamma a_j^t|h_{ij}^t|^2 + 1}\right), \end{aligned} \quad (1)$$

where N_0 denotes the thermal noise power and $\gamma = \frac{P}{N_0}$. γ is referred to as the signal-to-noise ratio (SNR) for notational simplicity, hereafter. Our goal is to enable each D2D transmitter i to autonomously determine its action a_i^t for maximizing the sum-rate $\sum_{i=1}^K c_i^t$ while reducing the feedback overhead.

III. PROPOSED DRL-BASED SCHEME USING REDUCED FEEDBACK

A. DRL-BASED TRANSMISSION SCHEME

We investigate a DRL-based transmission scheme where each D2D transmitter can determine autonomously whether to transmit data based on the feedback from its associated D2D receiver. Fig. 2 illustrates the architecture of the DRL-based transmission scheme. s_i^t denotes the input state of D2D transmitter i at time t, which consists of signal channel gain and interference channel gains that the receiver i feeds back. a_i^t is the output that is the binary action of D2D transmitter i at time t denoting whether to transmit data. In various environments, input states and actions are bilaterally co-related. Thus, actions are chosen based on input states and the chosen actions affect the next input states. However, actions and input states are unilaterally co-related in our case because the chosen actions don't change the next input state consisting of channel gains, even though actions are chosen based on input states. We, thus, use a dueling deep Q-network (DQN) as a learning model because it is particularly useful when actions and input states are unilaterally co-related [20], [21]. As shown in Fig. 2, the dueling DQN comprises of three main layers; feature layer, state-value layer, and advantage layer, denoted by θ, α, and β, respectively. Each main layer has three fully connected sub-layers and each fully connected sub-layer is followed by the rectified linear unit (ReLU) activator. The top stream, which consists of the feature layer and the state-value layer, yields a scalar output denoted by $V(s_i^t; \theta, \alpha)$ that can estimate the value of input state as a scalar value. The bottom stream, which consists of the feature layer and the advantage layer, yields an array output denoted by $A(s_i^t, a; \theta, \beta)$ that can estimate the advantage for each action a. For this study, $a \in \{0, 1\}$ and the output size of the bottom layer is 2. The top stream and bottom stream are aggregated, thereby yielding the final state-action values Q, calculated as

$$\begin{aligned} Q_i^t(s_i^t, a; \theta, \alpha, \beta) = V(s_i^t; \theta, \alpha) \\ + A(s_i^t, a; \theta, \beta) - \mathbb{E}_a[A(s_i^t, a; \theta, \beta)], \end{aligned} \quad (2)$$

where $\mathbb{E}_a[A(s_i^t, a; \theta, \beta)]$ is the average value of $A(s_i^t, a; \theta, \beta)$ over a and is included to resolve the lack of identifiability of Q value [20]. Without $\mathbb{E}_a[A(s_i^t, a; \theta, \beta)]$, we can't recover V and A values uniquely for a given Q value, which leads to poor performance in practical cases. Then, the D2D transmitter i chooses its action at time t, denoted by a_i^t, which has the greatest state-action value as

$$a_i^t = \underset{a\in\{0,1\}}{\arg\max}\, Q_i^t(s_i^t, a; \theta, \alpha, \beta). \quad (3)$$

If $a_i^t = 1$, the transmitter i transmits data to the receiver i, which broadcasts the data-rate c_i^t defined in (1). Otherwise,

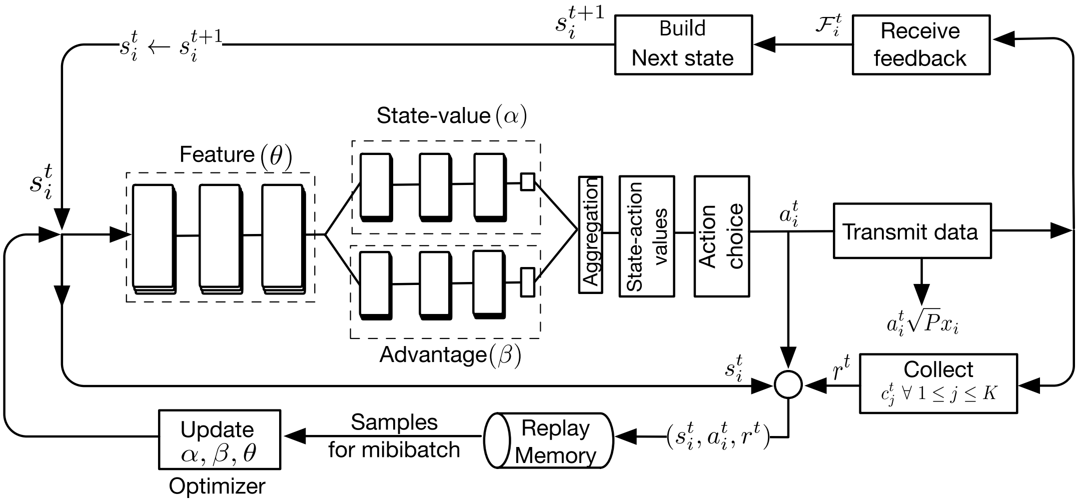


FIGURE 2. **The architecture of DRL-based distributed transmission scheme using dueling DQN.**

the transmit i doesn't transmit data. The reward that the transmitter i can achieve by taking the action a_i^t is defined as the sum-rate of K D2D pairs and can be calculated as

$$r_i^t \triangleq \sum_{j=1}^{K} c_j^t \tag{4}$$

by using the broadcast information from receivers. r_i^t's are all the same for all i's and the subscript i can be omitted. Then, the tuple (s_i^t, a_i^t, r^t) is saved in the replay memory for experience replay. Using randomly chosen samples from the replay memory, all the parameters α, β, and θ are gradually updated by Adam optimizer to minimize the difference between the action-state value for the selected action a_i^t, $Q_i^t(s_i^t, a_i^t; \theta_i^t, \alpha_i^t, \beta_i^t)$, and the target Q value, $\hat{Q}_i^t$. The target Q value is calculated by

$$\hat{Q}_i^t = r^t + \eta \max_{a'} Q_i^t(s_i^{t+1}, a'; \theta_i^t, \alpha_i^t, \beta_i^t), \tag{5}$$

where η denotes a discount rate for future rewards. Contrary to typical environments, the environment in this study is unilateral and the actions don't affect the subsequent input states and η is thus set to 0 as in [22] and [21]. The optimizer gradually trains α, β, and θ by using the gradient descent method to minimize the mean square error $||Q_i^t(s_i^t, a_i^t; \theta_i^t, \alpha_i^t, \beta_i^t) - \hat{Q}_i^t||^2$ as follows:

$$\theta \leftarrow \theta - \nu\nabla\theta,\ \alpha \leftarrow \alpha - \nu\nabla\alpha,\ \beta \leftarrow \beta - \nu\nabla\beta, \tag{6}$$

where ν denotes a learning rate.

B. NEW FEEDBACK SCHEMES

All transmitters broadcasts reference symbols, by which all D2D receivers can estimate channel gains. Each D2D receiver i can estimate N channel gains consisting of one signal channel gain from D2D transmitter i and $(N-1)$ interference channel gains from other transmitters by measuring the reference symbols transmitted by N D2D transmitters with no exchange of information with other devices. This assumption can be supported by the channel state information reference signal (CSI-RS) and sidelink communication procedure specified by 3rd generation partnership project (3GPP) [23]. In a conventional full feedback scheme, each D2D receiver i feeds back N channel gains as follows:

$$\mathcal{F}_{\text{full},i}^t = \left\{[j, |h_{ij}^t|^2], 1 \le j \le K\right\}, \tag{7}$$

where j denotes the identifier of D2D transmitter j. To reduce the amount of feedback information given in (7), we propose two feedback schemes: partial feedback scheme and binary feedback scheme. In the partial feedback scheme, each D2D receiver i feeds back the channel gains that are equal to or greater than its signal channel gain $|h_{ii}|^2$. Thus, the feedback information of receiver i can be described by

$$\mathcal{F}_{\text{partial},i}^t = \left\{[j, |h_{ij}^t|^2] \middle| j \in \left\{k \middle| |h_{ik}^t|^2 \ge |h_{ii}^t|^2, 1 \le k \le K\right\}\right\}. \tag{8}$$

In the binary feedback scheme, each D2D receiver i only feeds back identifiers of D2D transmitters whose channel gains are equal to or greater than the signal channel gain $|h_{ii}|^2$ as follows:

$$\mathcal{F}_{\text{binary},i}^t = \left\{[j] \middle| j \in \left\{k \middle| |h_{ik}^t|^2 \ge |h_{ii}^t|^2, 1 \le k \le K\right\}\right\}. \tag{9}$$

The feedback information in the binary feedback scheme consists of the same number of elements as the partial feedback scheme, however, the binary feedback scheme can further reduce the feedback overhead compared to the partial feedback scheme because the identifiers of channel gains are only fed back without real-valued channel gains. After transmitting data, D2D transmitter i receives the feedback information $\mathcal{F}_i^t$ transmitted by the receiver i, and builds the next input state s_i^{t+1} using $\mathcal{F}_i^t$, where $\mathcal{F}_i^t \in \{\mathcal{F}_{\text{full},i}^t, \mathcal{F}_{\text{partial},i}^t, \mathcal{F}_{\text{binary},i}^t\}$.

If $\mathcal{F}_i^t = \mathcal{F}_{\text{full},i}^t$, s_i^t is formulated as

$$s_{\text{full},i}^t = \left\{|h_{ii}^t|^2, |h_{i1}^t|^2, \cdots, |h_{i(i-1)}^t|^2, |h_{i(i+1)}^t|^2, \cdots, |h_{iK}^t|^2\right\}. \tag{10}$$

If $\mathcal{F}_i^t = \mathcal{F}_{\text{partial},i}^t$, the D2D receiver i feeds back only $|h_{ij}^t|^2$ $\forall j$ that are greater than or equal to $|h_{ii}^t|^2$. Thus, the input state is described as

$$s_{\text{partial},i}^t[k] = \begin{cases} s_{\text{full},i}^t[k] & \text{if } s_{\text{full},i}^t[k] \geq |h_{ii}^t|^2 \\ 0 & \text{otherwise,} \end{cases} \quad 1 \leq k \leq K, \tag{11}$$

where $s_{\text{partial},i}^t[k]$ and $s_{\text{full},i}^t[k]$ denote the k-th elements of $s_{\text{partial},i}^t$ and $s_{\text{full},i}^t$, respectively. If $\mathcal{F}_i^t = \mathcal{F}_{\text{binary},i}^t$, the D2D receiver i feeds back only identifiers of transmitters j's satisfying $|h_{ij}^t|^2 \geq |h_{ii}^t|^2|$. Thus, the input state can be described as

$$s_{\text{binary},i}^t[k] = \begin{cases} 1 & \text{if } s_{\text{full},i}^t[k] \geq |h_{ii}^t|^2 \\ 0 & \text{otherwise,} \end{cases} \quad 1 \leq k \leq K. \tag{12}$$

We mathematically analyze the feedback overhead of three different feedback schemes to evaluate how much the proposed feedback schemes can reduce the feedback overhead, compared the full feedback scheme. We have K D2D pairs in our system model, as depicted in Fi.g 1. Thus, $\lceil \log_2 K \rceil$ bits and F bits are required to identify one D2D transmitter and digitize one real-valued channel gain, respectively. The number of channel gains that each D2D receiver feeds back to its transmitter is not fixed in the partial and binary feedback schemes, as opposed to the full feedback scheme where it is always K. Let $\mathbb{E}[T]$ be the average number of channel gains that each D2D receiver feeds back to its transmitter in the proposed feedback schemes. Then, the total bits required for three feedback schemes can be calculated as

$$\begin{aligned} \mathcal{O}_{\text{full}} &= \left(\lceil \log_2 K \rceil + F\right) \times K, \\ \mathcal{O}_{\text{partial}} &= \left(\lceil \log_2 K \rceil + F\right) \times \mathbb{E}[T], \\ \mathcal{O}_{\text{binary}} &= \lceil \log_2 K \rceil \times \mathbb{E}[T]. \end{aligned} \tag{13}$$

In the partial and binary feedback schemes, $|h_{ii}^t|^2$ is used as a reference value when each D2D receiver i determines which channel gains to feed back. Thus, $1 \leq \mathbb{E}[T] \leq K$. If we let p be the probability that the receiver i feeds back $|h_{ij}^t|^2$, then the p is always constant regardless of i and $j (j \neq i)$ because we consider *i.i.d.* channels in this paper. $\mathbb{E}[T]$ can be thus calculated by

$$\begin{aligned} \mathbb{E}[T] &= \sum_{k=0}^{K-1} (k+1) \binom{K-1}{k} p^k (1-p)^{K-1-k} \\ &= (1-p)^{K-1} \sum_{k=0}^{K-1} (k+1) \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k \\ &= (1-p)^{K-1} \left[\sum_{k=0}^{K-1} k \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k \right. \\ &\qquad \left. + \sum_{k=0}^{K-1} \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k \right]. \end{aligned} \tag{14}$$

The overhead ratios of the partial and binary feedback schemes compared to the full feedback scheme can be calculated as

$$\rho_{\text{partial}} \triangleq \frac{\mathcal{O}_{\text{partial}}}{\mathcal{O}_{\text{full}}} = \frac{\mathbb{E}[T]}{K}, \tag{15}$$

$$\rho_{\text{binary}} \triangleq \frac{\mathcal{O}_{\text{binary}}}{\mathcal{O}_{\text{full}}} = \frac{\lceil \log_2 K \rceil}{\lceil \log_2 K + F \rceil} \rho_{\text{partial}}, \tag{16}$$

respectively.

Theorem 1: $\rho_{partial}$ converges to $\frac{1}{2}$ as K asymptotically increases.

Proof: If we use the following binomial expansion [24]

$$(1+x)^n = \sum_{i=0}^{n} \binom{n}{i} x^i, \tag{17}$$

$\sum_{k=0}^{K-1} \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k$ in (14) can be simplified to

$$\sum_{k=0}^{K-1} \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k = \left(1 + \frac{p}{1-p}\right)^{K-1}. \tag{18}$$

By taking the derivative of (17) over x, we can obtain

$$n(1+x)^{n-1} = \sum_{i=0}^{n} \binom{n}{i} i x^{i-1} = \frac{1}{x} \sum_{i=0}^{n} \binom{n}{i} i x^i, \tag{19}$$

which can be rewritten as

$$\sum_{i=0}^{n} \binom{n}{i} i x^i = nx(1+x)^{n-1}. \tag{20}$$

By using (20), $\sum_{k=0}^{K-1} k \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k$ in (14) can be simplified to

$$\begin{aligned} \sum_{k=0}^{K-1} k \binom{K-1}{k} \left(\frac{p}{1-p}\right)^k &= (K-1) \left(\frac{p}{1-p}\right) \\ &\quad \times \left(1 + \frac{p}{1-p}\right)^{K-2}. \end{aligned} \tag{21}$$

Then, (14) can be simplified as

$$\begin{aligned} \mathbb{E}[T] &= (1-p)^{K-1} \left[(K-1) \left(\frac{p}{1-p}\right) \left(1 + \frac{p}{1-p}\right)^{K-2} \right. \\ &\quad \left. + \left(1 + \frac{p}{1-p}\right)^{K-1} \right] \\ &= (1-p)^{K-1} \left(\frac{1}{1-p}\right)^{K-1} \left[p(K-1) + 1\right] \\ &= p(K-1) + 1 \end{aligned} \tag{22}$$

by using (18) and (21).

If D2D nodes are uniformly distributed, they have different path losses but the same distribution [25]. For two channel gains with the same distributions, $p = \frac{1}{2}$ by using Lemma 1. Then, (22) can be more simplified as

$$\mathbb{E}[T] = \frac{K+1}{2}. \tag{23}$$

Then, ρ_{partial} can be calculated as

$$\rho_{\text{partial}} = \frac{\mathbb{E}[T]}{K} = \frac{K+1}{2K} = \frac{1}{2}\left(1 + \frac{1}{K}\right), \tag{24}$$

and the proof of Theorem 1 is completed by

$$\lim_{K\to\infty} \rho_{\text{partial}} = \frac{1}{2}. \tag{25}$$

□

Lemma 1: For two independent random variables X and Y with the same pdf $f_X(x) = f_Y(y)$, $\Pr\{X > Y\} = \frac{1}{2}$.

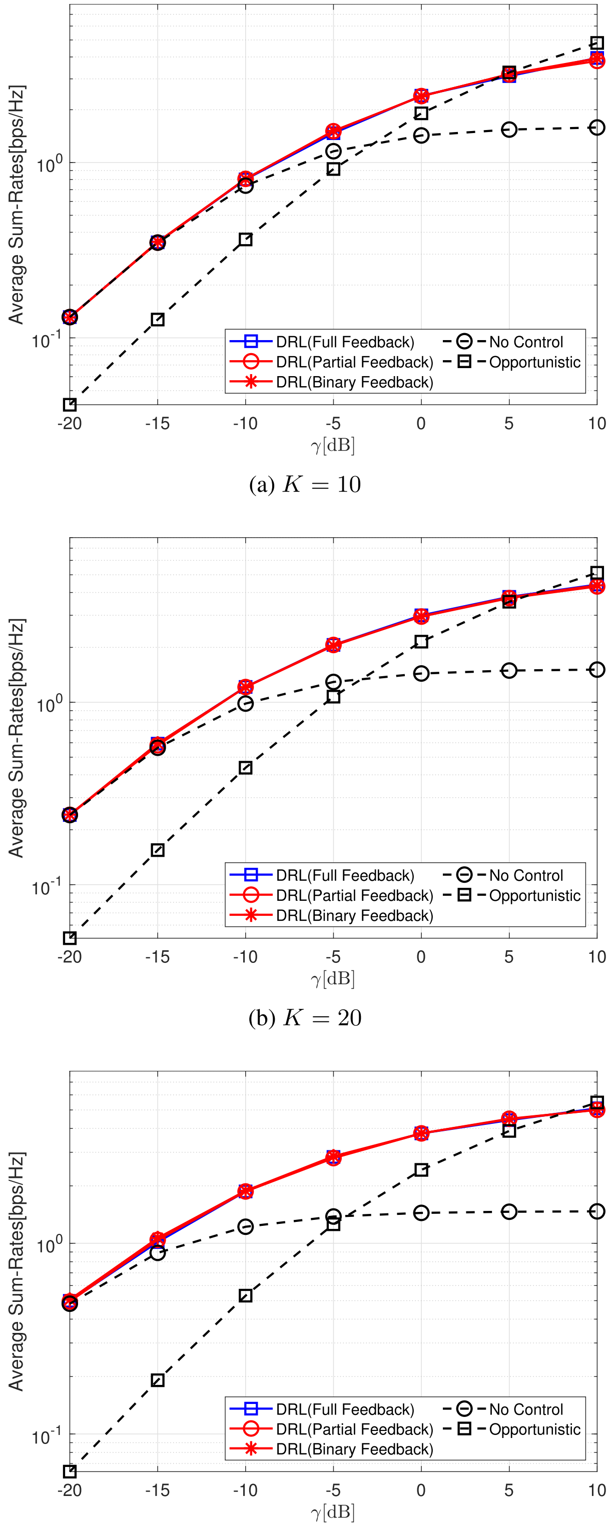
Proof:

$$\begin{aligned}
\Pr\{X > Y\} &= \iint\limits_{\{(x,y):x>y\}} f_{X,Y}(x,y)dxdy \\
&= \int_0^{+\infty}\int_y^{+\infty} f_X(x)f_Y(y)dxdy \qquad (26)\\
&= \int_0^{+\infty} f_Y(y)\left(1 - F_X(y)\right)dy \\
&= \int_0^{+\infty} \left(f_Y(y) - f_Y(y)F_X(y)\right)dy \\
&= 1 - \int_0^{+\infty} f_Y(y)F_Y(y)dy \qquad (27)\\
&= 1 - \int_0^{1} F_Y(y)\mathrm{d}F_Y(y) \qquad (28)\\
&= 1 - \left[\tfrac{1}{2}F_Y(y)^2\right]_{F_Y(y)=0}^{F_Y(y)=1} = \frac{1}{2}, \qquad (29)
\end{aligned}$$

where (26) is valid because the locations of D2D nodes are uncorrelated to each other and (27) is valid because D2D nodes are distributed based on the same distribution. (28) is valid because $f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$ [26]. □

IV. NUMERICAL RESULTS

We analyze the performance of the DRL-based distributed transmission scheme for D2D networks with three different feedback schemes in terms of average sum-rates and feedback overhead. All samples for training the dueling DQN are generated by simulations according to the channel model described in Sect. II. The network parameters of the dueling DQN for the three feedback schemes are first trained by Adam optimizer, with 500,000 samples, and the average sum-rates are then derived from the dueling DQN with the trained parameters using 100,000 samples different from the samples used for training. The mini-batch size, the learning rate ν, and the size of hidden layers for the dueling DQN are set to 10, 10^{-4}, and 1024, respectively. The parameters required for the dueling DQN are summarized in Table 1.



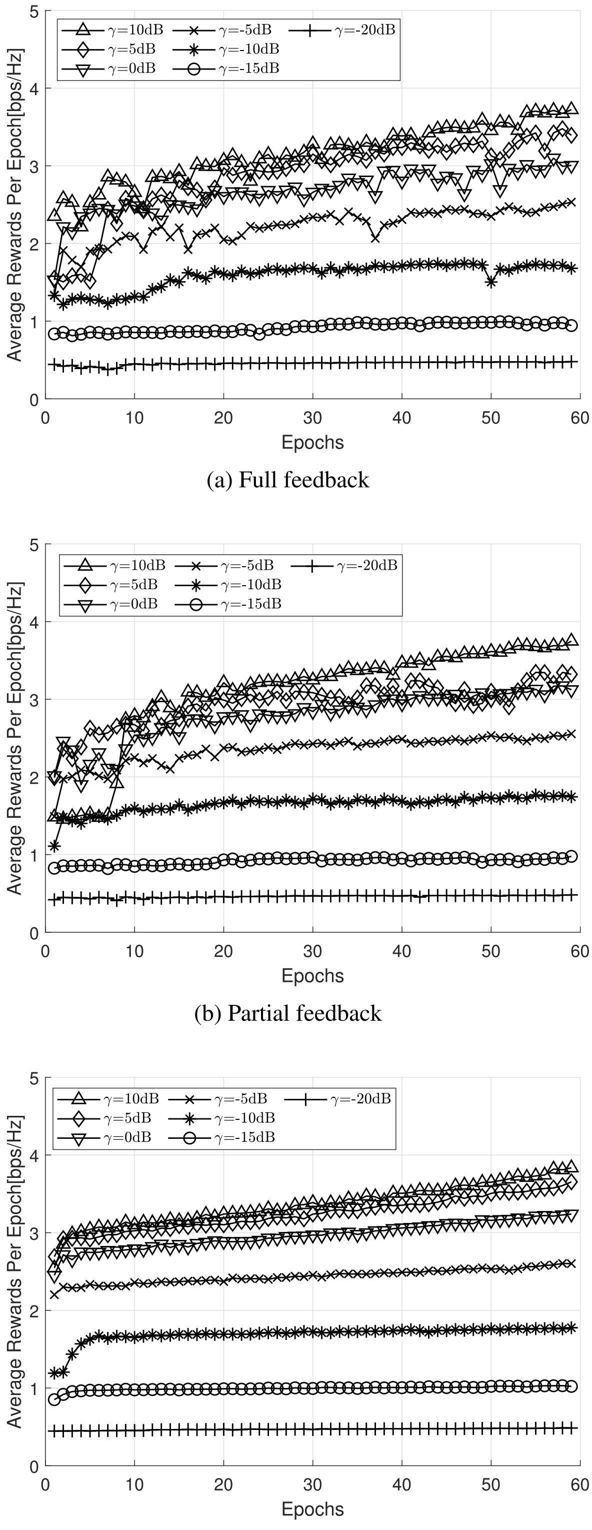
(a) $K = 10$

(b) $K = 20$

(c) $K = 50$

FIGURE 3. Average sum rates.

Fig. 3 shows average sum-rates of the DRL-based distributed transmission scheme with three feedback schemes for various values of γ when $K = 10, 20$, or 50. The average sum-rates of Opportunistic and No Control schemes are also presented as references for evaluating the performance of the DRL-based transmission schemes. In the No Control



(a) Full feedback

(b) Partial feedback

(c) Binary feedback

FIGURE 4. **Average rewards of DRL-based transmission schemes obtained during training. Each epoch consists of 5,000 episodes. $K = 50$.**

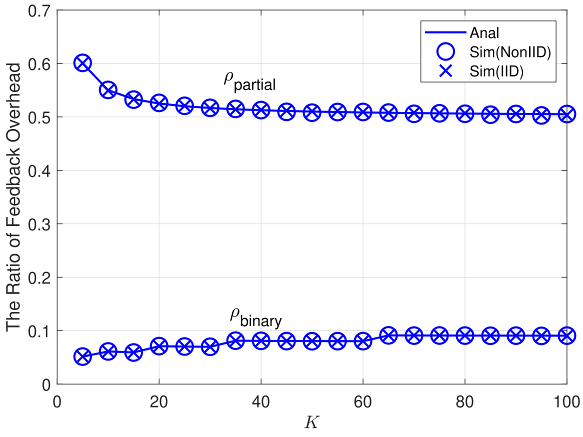


FIGURE 5. **Overhead ratios of the partial and binary feedback schemes to the full feedback scheme. F = 32bits/channel gain.**

scheme, all D2D transmitters always transmit data using their peak transmit power. On the other hand, in the Opportunistic scheme, a centralized node gathers all the signal channel gains and allows only one D2D transmitter with the highest signal channel gain to transmit data. The No Control scheme is optimal in power-limited environments while the Opportunistic scheme is optimal in interference-limited environments [27]. As shown in Fig. 3-(a), with $K = 10$, the No Control scheme outperforms the Opportunistic scheme for $\gamma \leq -3$dB because the interference level is low. The DRL-based transmission schemes achieve the same average sum-rates as the No Control scheme for $\gamma \leq -14$ dB and outperforms the No Control scheme for -14dB$\leq \gamma \leq -3$dB, regardless of the feedback schemes. For $\gamma \geq -3$dB, the Opportunistic scheme outperforms the No Control scheme because the co-channel interference among D2D transmitters becomes significant. For -3dB$\leq \gamma \leq 5$dB, the DRL-based transmission schemes outperform the Opportunistic scheme. For $\gamma \geq 5$dB, the Opportunistic scheme outperforms the DRL-based transmission schemes. However, if we consider that the Opportunistic scheme requires an extra centralized node to gather all signal channels and to select one D2D transmitter to transmit data, the difference in the average sum-rates is marginal. In addition, the partial feedback scheme and the binary feedback scheme can achieve the same average sum-rates as the full feedback scheme for all γ values, despite the reduced amount of feedback overhead. Figs. 3-(a) and (b) show that as K increases from 10 to 50, the range of γ where the No Control scheme outperforms the Opportunistic scheme is reduced and the difference of two schemes' average sum-rates increases because the co-channel interference among D2D transmitters becomes more significant as K increases. In addition, the DRL-based transmission schemes are superior to the No Control and Opportunistic schemes in a wider range of γ values. Specifically, the DRL-based transmission schemes can achieve average sum-rates equal to or greater than what the No Control scheme or the Opportunistic scheme can achieve when $\gamma \leq 5$dB for $K = 10$, while they outperform the No Control and Opportunistic schemes when $\gamma \leq 4$dB for $K = 50$. In addition, the partial and binary feedback schemes can achieve the same average sum-rates as the full feedback scheme even though K increases.

TABLE 1. Simulation parameters.

Parameters	Values
Batch size	10
Learning rate	10^{-4}
Feature layer(θ)	(Linear, ReLU, Linear, ReLU, Linear, ReLU)
Neurons of feature layer	(1024, 1024, 1024)
State-value and advantage layers	(Linear, ReLU, Linear, ReLU, Linear)
Neurons of state-value and advantage layers	(1024, 1024, 1024)
Size of replay memory	500,000
Training samples	500,000
Testing samples	100,000

Fig. 4 shows the average rewards that the dueling DQN can achieve for three feedback schemes while learning is in progress when $K = 50$. When γ is low, it is optimal for all transmitters to transmit data and there is no difference among three feedback schemes in terms of learning speed and stability. As γ increases, the proposed feedback schemes can train the dueling DQN more quickly and more stably, compared to the full feedback scheme. This implies that the proposed feedback schemes can accelerate the training of the dueling DQN by reducing the amount of feedback information included input states, while not corrupting the core context in the feedback information, thereby achieving the same performance as the full feedback scheme.

Fig. 5 shows the overhead ratios of the partial feedback scheme and binary feedback scheme to the full feedback scheme, defined in (15) and (16), respectively. F is set to 32 bits/channel gain according to the single-precision binary floating-point format in IEEE 754 [28]. As K increases, ρ_{partial} decreases and converges to 0.5 as in Theorem 1. ρ_{binary} slightly increases but is still lower than ρ_{partial}. More specifically, $\rho_{\text{partial}} \approx 0.5$ and $\rho_{\text{binary}} \approx 0.1$ when $K = 100$, indicating that the partial and binary feedback schemes can reduce the amount of feedback overhead by 50% and 90%, compared to the full feedback scheme, respectively, achieving the same average sum-rates. Fig. 5 also shows that the proposed scheme can achieve the same feedback reduction effect for non-*i.i.d.* channels as for *i.i.d.* channels. Non-*i.i.d.* channels are generated by assuming that all D2D nodes are uniformly distributed in a square with a side length of 10 m and the path loss for a separation distance d is calculated by d^{-3}.

V. CONCLUSION

We investigated DRL-based transmission schemes where each D2D transmitter can autonomously train its neural network to determine whether to transmit data to maximize the sum-rates of D2D communication networks using FDD as a duplex scheme. DRL-based transmission schemes for D2D communication networks are preferred to conventional centralized schemes using non-convex optimization because they can be easily implemented in distributed ways and can yield better performance in various environments. In various DRL-based transmission schemes, each D2D transmitter generates input states based on the channel gains perceived at its receiver. In TDD systems, each transmitter can build input states without the feedback from a receiver by using channel reciprocity. However, each transmitter needs the feedback from its receiver to obtain the channel gains in FDD systems where channel reciprocity is not valid. Thus, we proposed the partial and binary feedback schemes to reduce the amount of feedback overhead for FDD-based D2D communication networks, while enjoying the advantages of DRL-based transmission schemes. In the conventional full feedback scheme, each D2D receiver measures all channel gains from surrounding D2D transmitters including signal channel gain and interference channel gains, and feed them back to a transmitter. The channel gains that are equal to or greater than its signal channel gain are only fed back to a transmitter in the partial feedback scheme, while the identifiers of the channel gains are only fed back without real-valued channel gains in the binary feedback scheme. Our numerical results showed that the DRL-based transmission schemes can achieve optimal average sum-rates in power-limited environments with low γ values, and outperforms both No Control and Opportunistic schemes in moderate γ values. For high γ values, the average sum-rates of the DRL-based transmission schemes are lower than those of Opportunistic scheme. However, the gap in average sum-rates between two schemes becomes marginal as K increases. Contrary to the Opportunistic scheme that is a centralized scheme and requires a centralized control node, the DRL-based schemes enable each D2D transmitter to determine its action autonomously. In addition, the partial and binary feedback schemes for DRL can both achieve the same average sum-rates as the full feedback scheme in all the environments considered in our simulations, while significantly reducing feedback overhead. The feedback overhead of the partial feedback scheme converges to 50%, compared to that of the full feedback scheme, as K asymptotically increases. More specifically, when $K = 50$, the partial and binary feedback schemes reduce the feedback overhead by approximately 49% and 92%, respectively, compared to the full feedback scheme.

REFERENCES

[1] Y. Qiao, Y. Li, and J. Li, "An economic incentive for D2D assisted offloading using Stackelberg game," *IEEE Access*, vol. 8, pp. 136684–136696, 2020.

[2] S. Yasukawa, H. Harada, S. Nagata, and Q. Zhao, "D2D communications in LTE advanced release 12," *NTT Docomo Tech. J.*, vol. 17, no. 2, pp. 56–64, 2015.

[3] *NR and NG-RAN Overall Desciprtion;Stage 2*, document TS 38.300, Version 16.5.0, 3rd Generation Partnership Project, Mar. 2021.

[4] L. Shi and S. Xu, "UAV path planning with QoS constraint in device-to-device 5G networks using particle swarm optimization," *IEEE Access*, vol. 8, pp. 137884–137896, 2020.

[5] W. Sun, E. G. Ström, F. Brännström, K. C. Sou, and Y. Sui, "Radio resource management for D2D-based V2V communication," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6636–6650, Aug. 2016.

[6] K. M. Malarski, F. Moradi, K. D. Ballal, L. Dittmann, and S. Ruepp, "Internet of reliable things: Toward D2D-enabled NB-IoT," in *Proc. 5th Int. Conf. Fog Mobile Edge Comput. (FMEC)*, Apr. 2020, pp. 196–201.

[7] *Architecture Enhancements for 5G System to Support Vehicle-to-Everything (V2X) Services*, document TS 23.287, Version 16.5.0, 3rd Generation Partnership Project, Dec. 2020.

[8] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 3rd Quart., 2021.

[9] T.-W. Ban and B. C. Jung, "On the link scheduling for cellular-aided device-to-device networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9404–9409, Nov. 2016.

[10] Y. Zhang, C.-Y. Wang, and H.-Y. Wei, "Incentive compatible overlay D2D system: A group-based framework without CQI feedback," *IEEE Trans. Mobile Comput.*, vol. 17, no. 9, pp. 2069–2086, Sep. 2018.

[11] G. Zhao, S. Chen, L. Qi, L. Zhao, and L. Hanzo, "Mobile-traffic-aware offloading for energy- and spectral-efficient large-scale D2D-enabled cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3251–3264, Jun. 2019.

[12] M. Chen, A. Liu, W. Liu, K. Ota, M. Dong, and N. N. Xiong, "RDRL: A recurrent deep reinforcement learning scheme for dynamic spectrum access in reconfigurable wireless networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 364–376, Mar. 2022.

[13] M. Chen, W. Liu, T. Wang, S. Zhang, and A. Liu, "A game-based deep reinforcement learning approach for energy-efficient computation in MEC systems," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107660. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121009229

[14] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[15] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3143–3156, Mar. 2021.

[16] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.

[17] C. Du, Z. Zhang, X. Wang, and J. An, "Deep learning based power allocation for workload driven full-duplex D2D-aided underlaying networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15880–15892, Dec. 2020.

[18] N. Jindal, "A feedback reduction technique for MIMO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 2699–2703.

[19] M. Benmimoune, E. Driouch, W. Ajib, and D. Massicotte, "Feedback reduction and efficient antenna selection for massive MIMO system," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–6.

[20] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 48, New York, NY, USA, Jun. 2016, pp. 1995–2003.

[21] T.-W. Ban, "An autonomous transmission scheme using dueling DQN for D2D communication networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16348–16352, Dec. 2020.

[22] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep *Q* learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[23] *NR;Physical Layer Procedures for Data (Release 16)*, document 3GPP TS 38.214, Version 16.10.0, Jun. 2022.

[24] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Amsterdam, The Netherlands: Elsevier, 2007.

[25] Z. Bharucha and H. Haas, "The distribution of path losses for uniformly distributed nodes in a circle," *Res. Lett. Commun.*, vol. 2008, pp. 1–4, Jan. 2008.

[26] A. Papoulis, S. Pillai, and S. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.

[27] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[28] *IEEE Standard for Floating-Point Arithmetic*, IEEE Standard 754-2019 (Revision of IEEE 754-2008), 2019, pp. 1–84.

TAE-WON BAN (Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Kyungpook National University, South Korea, in 1998 and 2000, respectively, and the Ph.D. degree from the Department of Electrical and Electronic Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2010. He was a Researcher and a Network Engineer with Korea Telecom (KT), from 2000 to 2012. In KT, he researched 3G WCDMA, LTE, and Femto Systems. He was also responsible for a traffic engineering and spectrum strategy. He is currently a Professor with the Department of Intelligent Communication Engineering, Gyeongsang National University, South Korea. His current research interests include OFDM, MIMO, and radio resource management based on deep learning for mobile communication systems.