Reconfigurable Intelligent Surface Enhanced Device-to-Device Communications

Preprint · June 2020

2 authors:

Zelin Ji
Queen Mary, University of London
6 PUBLICATIONS   **14** CITATIONS

SEE PROFILE

Zhijin Qin
Queen Mary, University of London
127 PUBLICATIONS   **3,442** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Machine Learning for Wireless Communications View project

Project   Semantic Communications View project

Reconfigurable Intelligent Surface Enhanced Device-to-Device Communications

Zelin Ji, Zhijin Qin

Queen Mary University of London, London, UK

Email: {z.ji, z.qin}@qmul.ac.uk

Abstract—**Reconfigurable intelligent surface (RIS) technology is a promising method to enhance the device-to-device (D2D) communications. To maximize the sum rate of the cellular and D2D networks, a joint optimization of the position and the phase shift of RIS in D2D communications is considered in this paper. To solve the non-convex sum rate maximum problem, we propose a novel convolutional neural network (CNN) based deep Q-network (DQN) that jointly optimizes the RIS position and its phase shift with lower complexity. Numerical results illustrate that the proposed algorithm can achieve higher sum rate compared to the benchmark algorithms, meanwhile meeting the quality of service (QoS) requirements at D2D receivers and the base station (BS).**

I. INTRODUCTION

As one of the key technologies of the fifth generation (5G) and beyond communication systems, device-to-device (D2D) communications can enhance the communications performance by reducing the latency and improving energy and spectrum efficiency. However, due to the underlay resource reusing mode, the interference management in D2D communications becomes challenging [1]. The existing works on D2D communications mainly focus on transmit power and channel assignment optimization [2], [3]. Unfortunately, due to the fast channel variations in D2D communications, few works contribute to the optimization of wireless communication environment.

Reconfigurable intelligent surfaces (RISs) have attracted extensive attension in wireless communications due to the ability of proactively modifying the wireless communication environment. Equipped with an array of low-cost passive reflecting elements, RISs can adjust the phase shift and amplitude of each element. Compared with conventional relays, the advantages of RIS include energy consumption reduction and achievable data rate improvement [4]. The performance improvements have been verified by an RIS-based wireless communication prototype [5]. The phase of each element is controlled by the RIS controller, which receives control signal from the base station (BS). Although the control signal of RIS can be analog using varactors to provide continuous phase shift [6], the large response time and low phase accuracy of varactors make it impractical for wireless communications. In [7], [8], theoretical analyses have been provided for multi-bit controlled elements to strike a tradeoff between the performance and complexity.

However, to benefit the overall system performance, the optimization of RIS becomes challenge due to the large number of reflecting elements [9]. Many approaches have been applied to optimize RIS to achieve higher throughput or energy efficiency. Particularly, RIS has been successfully adopted in D2D networks in [10]–[12] to maximize the sum rate in D2D communication systems. To solve the non-convex sum rate maximizing problems, they tend to find the sub-optimal solution by using the block coordinate descent [10] and Riemannian pursuit method [11]. To achieve performance-complexity tradeoff, the projected sub-gradient method is leveraged for the phase shift design [12]. However, the channels in D2D networks vary over time, resulting algorithms based on the long-term optimization and high complexity not applicable any more.

Recently, machine learning (ML) methods, especially deep learning (DL), have become promising tools to address explosive mass data, mathematically intractable nonlinear non-convex problems and high-computation issues. DL based approaches can significantly reduce the complexity, and have been adopted in wireless communication systems, e.g., physical layer communications [13] and resource allocation [14]. Motivated by the potential applications of DL in solving sophisticated optimization problems, the authors in [15] have adopted the DL method for designing the RIS reflection matrices with restricted channel state information (CSI).

As a novel branch of DL, deep Q-network (DQN) has been proposed [16]. Embraced with reinforcement learning, DQN enables agents to learn and build knowledge by interacting with the environment and maximizing the desired reward, thereby showing its great potential in circumventing challenges of conventional DL. Particularly, DQN is beneficial to discrete phase shift design and wireless communication systems where radio channels vary over time. Leveraging DQN, the authors in [17] proposed to jointly design the beamforming at the BS and phase shifts at the reflecting RIS to maximize the sum rate.

Note that the applications of DQN in communication systems usually eoploit the fully connected layer instead of convolutional layers, which results in significant increasing on the number of training parameters [18]. This motivate us to combine DQN with convolutional neural network (CNN) to jointly optimize the position and phase shift of the RIS, thereby reducing the number of parameters and computation complexity. The major contributions of this paper are summarized as follows.

1) The position and phase shift of RIS are jointly optimized

for the D2D networks, which is valuable to be considered before the installation of the RIS.

2) A novel CNN based DQN structure is proposed to reduce the number of training parameters significantly, therefore speed up the convergence time especially when the number of D2D pairs, cellular users and RIS elements are large.

3) In order to meet the quality of service (QoS) requirements of the D2D receivers and the BS, a dynamic reward is defined.

II. System Model and Problem Formulation

We consider the uplink transmission of a D2D network, which includes K cellular users, communicate with the BS in the conventional cellular mode, and I D2D pairs communicate with each other. To enhance transmission performance of the network, an RIS composed of N passive elements is installed. Assuming that the ith D2D transmitter $D_{t,i}$ communicates with the corresonding receiver $D_{r,i}$ by reusing the resource block for the uplink of the kth cellular user CU_k, then CU_k becomes the source of interference for ith D2D pairs.

A. System description

Practically, the location of a BS is fixed as shown in 1. We define the channel gain $\boldsymbol{h_t^T}$ from the ith D2D transmitter $D_{t,i}$ and the RIS as

$$\boldsymbol{h_{t,i}^T} = [h_{t,i}^1, \ldots, h_{t,i}^n, \ldots, h_{t,i}^N]^T, \tag{1}$$

where $\boldsymbol{h_{t,i}^T}$ represent transpose matrix of $\boldsymbol{h_{t,i}}$, $h_{t,i}^n$ represents the Nakagami fading channel gain between D2D transmitter $D_{t,i}$ and the nth RIS element, which can be denoted as

$$h_{t,i}^n = \beta m \sqrt{d_{t,i}^{-\alpha}}, \forall n \in N, \tag{2}$$

where β is a constant value, m is a variable that represents the Nakagami distribution, $d_{t,i}$ is the distance between the ith D2D transmitter $D_{t,i}$ and the RIS, and α represents the path loss exponent. Similarly, the channel gain between RIS and the ith D2D receiver $D_{r,i}$, between the kth cellular user CU_k and RIS, between RIS and BS are defined as $\boldsymbol{h_{r,i}} \in \mathbb{C}^{1 \times N}$, $\boldsymbol{h_k^T} \in \mathbb{C}^{N \times 1}$ and $\boldsymbol{h_{BS}} \in \mathbb{C}^{1 \times N}$, respectively. Note that the channel gains towards the RIS have the dimension $\mathbb{C}^{N \times 1}$, while the gains from the RIS have the dimension $\mathbb{C}^{1 \times N}$. The phase shift and amplitude attenuation A for all the RIS elements can be expressed by $\boldsymbol{\Theta} \triangleq diag[Ae^{j\theta_1}, Ae^{j\theta_2}, \ldots, Ae^{j\theta_N}]$, where $A \in [0,1]$ and $\theta \in [0, 2\pi)$.

For the line-of-sight (LoS) links, we denote $h_{tr,i} = \beta m \sqrt{(d_{tr,i})^{-\alpha}}$ as the channel gain for the link between D2D pairs ($D_{t,i}$ and $D_{r,i}$), where $d_{tr,i}$ represents the distance between $D_{t,i}$ and $D_{r,i}$. The channel gain from the kth cellular user CU_k to the ith D2D receiver $D_{r,i}$, from the ith D2D transmitter $D_{t,i}$ to BS, and from the kth cellular user CU_k to BS can be denoted by $h_{k,i}$, $h_{i,BS}$, and h_u in the similar way.

Overall, the channel gain h_D between D2D link and $h_{C,D}$ between cellular user CU_k to D2D receiver $D_{r,i}$ can be given by
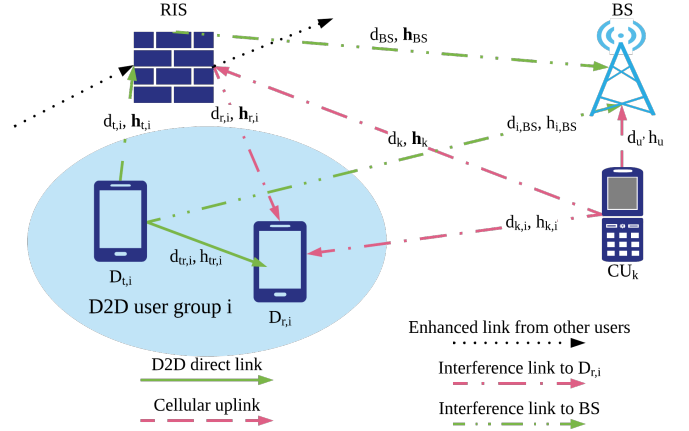

Fig. 1. System model of the of RIS enhanced D2D network.

$$h_D = \underbrace{\boldsymbol{h_{t,i}^T \Theta h_{r,i}}}_{\text{Reflection link}} + \underbrace{h_{tr,i}}_{\text{LoS link}}, \tag{3}$$

and

$$h_{C,D} = \boldsymbol{h_{r,i} \Theta h_k^T} + h_{k,i}. \tag{4}$$

In the similar way, the overall channel gain from CU_k and $D_{t,i}$ to the BS can be represented as h_C and $h_{D,BS}$, respectively. The signal y_i received by $D_{r,i}$ is

$$y_i = h_D x_i + \underbrace{h_{C,D} x_k}_{\text{Interference signal}} + \underbrace{z}_{\text{Noise}}, \tag{5}$$

where $x_{t,i} \triangleq \sqrt{p_i} u_i$ and $x_k \triangleq \sqrt{p_k} u_k$ denotes the signal from $D_{t,i}$ and CU_k, p_i and u_i denotes transmission power of the D2D transmitter and unit variance entries with zero mean, $z \sim N(0, \sigma^2)$ denotes the AWGN noise signal with σ^2 variance. Then, the signal-to-interference-plus-noise ratio (SINR) γ_i^D at ith D2D receiver and γ_k^C at the BS for the kth cellular user can be denoted as

$$\gamma_i^D = \frac{p_i |h_D|^2}{\sum_{k=1}^K \rho_{k,i} p_k |h_{C,D}|^2 + \sigma^2}, \tag{6}$$

and

$$\gamma_k^C = \frac{p_k |h_C|^2}{\sum_{i=1}^I \rho_{k,i} p_i |h_{D,BS}|^2 + \sigma^2}, \tag{7}$$

where $\rho_{k,i}$ is the resource reuse coefficient of kth cellular user and ith D2D pair, $\rho_{k,i} = 1$ when ith D2D pair reuses the resource of CU_k, and $\rho_{k,i} = 0$ otherwise.

B. Problem formulation

Our objective is to maximize the sum rate of D2D and cellular networks, which could given by

$$
\begin{aligned}
R &= \sum_{i=1} R_i + \sum_{k=1} R_k \\
&= \sum_{i=1} B_i log_2(1 + \gamma_i^D) + \sum_{i=k} B_k log_2(1 + \gamma_k^C),
\end{aligned}
\tag{8}
$$

where B_i and B_k represents the bandwidth of ith D2D transmission link and the uplink of kth cellular user CU_k.

This paper aims to maximize the sum rate in (8) by jointly optimize the phase shift and position of RIS. The joint RIS positioning and phase shift problem can be formulated as

$$\text{P1:} \quad \underset{\{\boldsymbol{\Theta}, \boldsymbol{s_{RIS}}, \boldsymbol{\rho}\}}{\text{maximize}} \quad R \tag{9a}$$

$$\text{subject to} \quad \gamma_i^D \geq \gamma_{min}^D, \forall i, \tag{9b}$$

$$\gamma_k^C \geq \gamma_{min}^C, \forall k, \tag{9c}$$

$$0 < \theta_n \leq \pi, \forall n \in N, \tag{9d}$$

$$\boldsymbol{s_{RIS}} \in \mathbb{R}^2, \tag{9e}$$

where γ_{min}^D and γ_{min}^C are the minimum SINR requirements at the D2D receiver and the BS, respectively. Coordinate $\boldsymbol{s_{RIS}}$ restricts a 2-dimension space for the installation of the RIS. The constraint (9d) makes $P1$ non-convex. To solve the non-convex problem by mathematical tools, we have to utilize exhaustive search, which is impractical when the number of D2D pairs, cellular users, and RIS elements become large. Generally, classical mathematical tools can be leveraged to acquire suboptimal solutions [10]–[12]. Alternatively, instead of solving challenging non-convex problem by mathematical tools, we adopt CNN based DQN algorithm, which is more applicable to solve problems with high dimension inputs and large state and action space.

III. PROBLEM SOLUTIONS

In this section, we propose a CNN based DQN learning model is proposed to solve the joint RIS positioning and phase shift problem. Particularly, the RL components are first defined and the DQN components are then introduced. Then the proposed algorithm is explained in detail.

A. RL components definition

In our proposed algorithm, the state contains the position information of D2D users, cellular users, BS and RIS, as well as the phase shift $\boldsymbol{\Theta}$. The position vectors $\boldsymbol{S_t} = [s_{t,1}, \ldots, s_{t,I}]$, $\boldsymbol{S_r} = [s_{r,1}, \ldots, s_{r,I}]$, $\boldsymbol{S_u} = [s_{u,1}, \ldots, s_{u,K}]$, $\boldsymbol{s_{RIS}}$ and $\boldsymbol{s_{BS}}$ represents the position information of D2D transmitters, D2D receivers, cellular users, RIS and BS, respectively. The input state $\mathcal{S} = [\boldsymbol{S_t}, \boldsymbol{S_r}, \boldsymbol{S_u}, \boldsymbol{s_{RIS}}, \boldsymbol{s_{BS}}; \boldsymbol{\Theta}]$, which has a cardinality $|\mathcal{S}|$ of $(2I + K + N + 2)$.

Action set $\mathcal{A}$ represents the possible action choice for the RIS controller. Generally, the position of RIS are fixed after installation, while the phase shift can be adjusted, so the action space contains the phase shift adjustment and position choice of RIS. At iteration t, action a_t consists of two parts: i) the variable quantity of phase shift matrix, $\Delta\boldsymbol{\Theta} = \{\Delta\theta_1, \ldots, \Delta\theta_N\}$, where $\Delta\theta_n \in \{-\delta, 0, +\delta\}, \forall n \in N$; ii) the position choice of RIS, $\boldsymbol{s} \in \{v_1, \ldots, v_O\}$, where O represents the number of grids in the communications system. Formally, the action $a_t = [\Delta\boldsymbol{\Theta}; \boldsymbol{s_{RIS}}]$, which has a cardinality $|a|$ of $(N + 1)$. Action set $\mathcal{A}$ includes all possible actions with the cardinality $|\mathcal{A}| = 3^N \times O$.
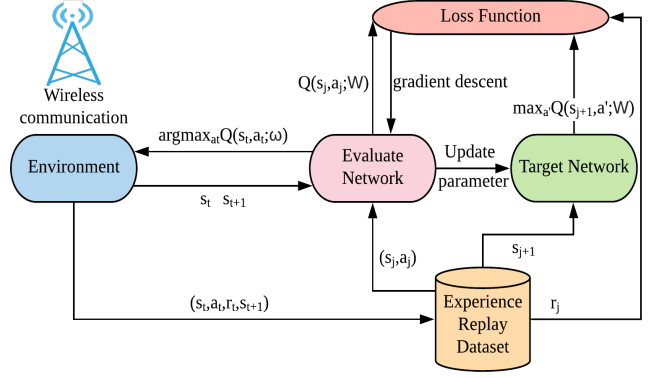

Fig. 2. The proposed CNN based DQN algorithm for the RIS position and phase shift optimization.

The reward represents whether we encourage or punish an action, so it is defined based on the objective function given in (8). For a successful transmission at iteration t, i.e., the constraints (9b) and (9c) are satisfied, the reward r_s can be defined as $r_s = R(t)$, where $R(t)$ represents the achievable rate R at iteration t. However, if any of the constraints are not satisfied, the expected QoS cannot be achieved. This kind of action results in penalty due to energy waste, and we defined the new reward for the transmission failure as

$$r_\text{f} = \begin{cases} \sum\limits_{i=1} R_i, & \text{if (9b) is not satisfied;} \\ \sum\limits_{k=1} R_k, & \text{if (9c) is not satisfied;} \\ 0, & \text{otherwise;} \end{cases} \tag{10}$$

The fail reward is to encourage the communication system to improve the SINR which is not satisfied the requirement. The overall reward can be expressed as

$$r_t = \begin{cases} r_s, & \text{if (9b) and (9c) are satisfied;} \\ r_\text{f}, & \text{else.} \end{cases} \tag{11}$$

The optimal action-value function obeys an important identity known as the Bellman equation. The optimal strategy is to select the action that maximizes [16]:

$$Q^*(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r_t + \Gamma \max_{a' \in A} Q^*(s_{t+1}, a')|s_t, a_t], \tag{12}$$

where $Q^*(s, a)$ is the desired value function such that $Q(s_t, a_t) \to Q^*(s_t, a_t)$ as $t \to \infty$. However, it is impractical since the iteration is discrete. Instead, the neural networks (NN) are applied to be function approximator to estimate the action-value function, i.e., $Q(s_t, a_t; \boldsymbol{W}) \approx Q^*(s, a)$. When the state and action space become large, this method does not need to maintain the large Q-table as conventional RL approaches do, thereby expanding the applications of RL in wireless communications greatly.

B. DQN components

Leveraging the NN, the DQN model can find the relationship between the input location of D2D and cellular users

TABLE I
POSITION AND PHASE SHIFT INPUT FORMAT TO CNN BASED DQN

Position	$s_{t,1}$	$s_{r,1}$	...	$s_{t,I}$	$s_{r,I}$	$s_{u,1}$	...	$s_{u,K}$	s_{RIS}	s_{BS}
Phase	θ_1	θ_2	...	θ_n	...	θ_N	0	0	0	...

and the corresponding position and phase shift combination of RIS. The components in DQN is defined as

- Agent: The agent in our DQN model is RIS controller. RIS controller will proceed the inputs and execute the outputs of DQN to control RIS.
- Input: The DQN model takes the states $\mathcal{S}$ as the input, which includes the position and phase shift information.
- Output: The output of the DQN model is the evaluate Q-value for state-action pairs. The output layer contains $|\mathcal{A}|$ units, which represents the number of possible actions. As shown in Fig. 2, two identical networks are set: evaluation network and target network. In the evaluation network, current state s_j is the input information, and the output is the evaluate Q-value for each action. In the target network, next expected state s_{j+1} is the input, while the output is the Q-value for the each action in the next state.

C. DQN learning process

At each iteration, the BS observes the states and send them to RIS controller as the input. DQN in RIS controller will process the input and output the approximately Q-value for each state-action pair. The policy for selecting actions is to make a trade-off between exploitation and exploration, so we apply decaying ϵ-greedy algorithm [19]. The agent will choose actions uniformly from $\mathcal{A}$ with probability of (1-ϵ), while choosing the action which maximize the Q-value with the probability of ϵ.

In the adopted DQN, the training data set, also named replay memory $\mathcal{D} = [e_1, \ldots, e_t, \ldots]$ for NN is stored according to agent's experience at each iteration t, where the experience $e_t = (s_t, a_t, r_t, s_{t+1})$ is called transition, including the state, action and reward information. The training minibatch (s_j, a_j, r_j, s_{j+1}) is sampled from the training data set. During the training process, parameters are updated to the Q estimation network at each step to generate the estimated Q-value. Q target network is updated after every g steps according to the parameters in the Q estimation network. The training process for DQN is to minimize the error function which represents the estimated Q-value and the realistic Q-value. For the DQN in this work, the error function can be expressed by:

$$\text{Loss}(\boldsymbol{W}) = \mathbb{E}[(q_{target} - Q(s_j, a_j; \boldsymbol{W}))^2], \quad (13)$$

where $q_{target} = r_j + \Gamma \max_{a'} Q(s_{j+1}, a'; \boldsymbol{W}^*)$ is the target Q-value for minibatch j, which is the output of Q target network. $\boldsymbol{W}$ and $\boldsymbol{W}^*$ denotes the weights of the evaluation network and the target network, respectively. The weights are optimized by the gradient descent method [16].

D. Proposed CNN based DQN algorithm for the control of RIS

Generally, in each iteration, the fully connected network need to train $|\mathcal{S}| \times |\mathcal{A}|$ parameters, and even more if there are hidden layers. In order to reduce the number of training parameters and improve the training complexity, CNN is an effective approach. The proposed CNN algorithm consists of a convolutional layer, followed by a flatten layer and a fully connected layer which connected to the output layer.

- Convolutional layer: In the proposed CNN based DQN algorithm, the convolutional layer is employed to extract the information of input, i.e., the position and the phase shift information in this paper. The inputs information are stored in a matrix shown in Table I. The number of columns $U = \max(2I + K + 2, N)$. To make the format complete, the rest elements are set to 0. Mathematically, the input of the convolutional layer is $\mathcal{S} \in \mathbb{R}^{2 \times U}$, then the output of convolutional layer is given by:

$$\boldsymbol{v}_1^\lambda = f(\mathcal{S} \otimes \boldsymbol{\omega}_1^\lambda + \boldsymbol{b}_1^\lambda), \quad (14)$$

where $f(\cdot) = \max(0, \cdot)$ is rectified linear unit (ReLU) activation function, $\otimes$ represents the convolution operation, $\boldsymbol{\omega}_1^\lambda \in \mathbb{R}^{2 \times \mu}$ and $\boldsymbol{b}_1^\lambda$ represents the convolution kernel and bias in first convolutional layer of the feature map λ. Note that output feature map is $\boldsymbol{v}_1^\lambda \in \mathbb{R}^{U - \mu + 1}$.

- Flatten layer: A flatten layer is applied after convolutional layer, whose inputs are Λ feature maps extracted by the convolutional layer, and it generates the feature vector $\boldsymbol{v}_2 \in \mathbb{R}^\nu$, where $\nu = \Lambda \times (U - \mu + 1)$.
- Hidden layer: The hidden layer is fully-connected with 256 rectifier units.
- Output layer: The output layer is a fully-connected linear layer with single output of each valid action, represents the evaluated Q-value or the target Q-value for a specific state action pair in evaluation network or target network, respectively.
- Complexity comparison: Compared with the conventional DQN need to train $(|\mathcal{S}| \times |\mathcal{A}|)$, the number of training parameter in the first layer of CNN based DQN is only $(2 \times \mu \times \Lambda)$. Clearly, CNN can reduce the complexity as state and action space becomes large.

Overall, by receiving the input information of position information of D2D pairs, cellular users and the BS from the wireless environment, the RIS controller can train the weights and update NNs to estimate the action value function. The proposed algorithm learns the policy to jointly optimize the installation position and phase shift of the RIS with much lower complexity. The details of the proposed algorithm are shown in Algorithm 1.

Algorithm 1 CNN based DQN algorithm for the RIS

 1: **Input**: Environment simulator, Q network, replay memory $\mathcal{D}$, minibatch size;
 2: Initialize: action-value function Q with random weights $\boldsymbol{W}$, replay memory $\mathcal{D}$, RIS position and phase;
 3: **repeat**
 4:     for each iteration step:
 5:     choose action a_t from aciton space $\mathcal{A}$ according to ϵ-greedy algorithm;
 6:     Execute a_t, calculate dynamic reward r_t by (11) and observe s_{t+1};
 7:     Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$;
 8:     Replay memory:
 9:     Sample random minibatch of transitions (s_j, a_j, r_j, s_{j+1}) in $\mathcal{D}$;
10:     Calculate q_{target} by (13)
11:     Perform a gradient descent step on $(q_{target} - Q(s_j, a_j; \boldsymbol{W}))^2$;
12: **until** s_t is the goal state
13: **Return**: Action value function and optimized action a.

Fig. 3. Reward comparison of proposed CNN based DQN algorithm, conventional DQN, Q-learning and Random. $N = 16$, $p_i = 15$dBm, $p_k = 30$dBm, average reward $r = \sum r_t/t$.

IV. NUMERICAL RESULTS

In this section, performance of the proposed CNN based DQN algorithm is evaluated by comparing it with the benchmark algorithms. We take 1 D2D pair and 1 cellular user in the considered network as an example, i.e., $I = K = 1$. We place them in a square with size of $100m \times 100m$. Particularly, D2D transmitter, D2D receiver, cellular user, and the BS are fixed at position (40,20), (60,20), (25,55) and (75,55) respectively. The whole area is divided in $O = 25$ identical squares, where RIS can be installed in any of them. The variable quantity δ of the phase shift is $\frac{\pi}{4}$. To guarantee the QoS, the minimum SINR requirements for D2D receiver γ_{min}^D and for cellular uplink γ_{min}^C are -10dB and -13dB, respectively. Gaussian noise variance σ^2 is -116dBm. The path loss parameter d is set to 3. The learning rate α is set $\frac{1}{\text{Iterations}}$ to guarantee the convergence performance of the algorithm. Discount factor Γ is set to 0.9. The exploration rate ϵ is set to 0 at the beginning and increases by 0.01 at each iteration until it reaches 0.9 so that RL can explore the new actions with reasonable probability while guaranteeing the system performance.

A. Convergence Performance of Proposed Algorithms

Fig. 3 demonstrates the convergence of the proposed CNN based DQN algorithm, while other algorithms including conventional DQN algorithm, and Q-learning algorithm are also demonstrated as benchmarks. In the random benchmark scheme, the actions are chosen randomly from the action space $\mathcal{A}$ and the agents do not learn anything from the environment. Note that the reward r_t is calculated by (11), so it is not equal to the sum rate R. From the figure, the proposed algorithm achieves the highest reward with fast convergence speed, while conventional DQN algorithm can achieve the same reward with more iterations. Q-learning can achieve fastest convergence, however, the reward is lower than DQN and
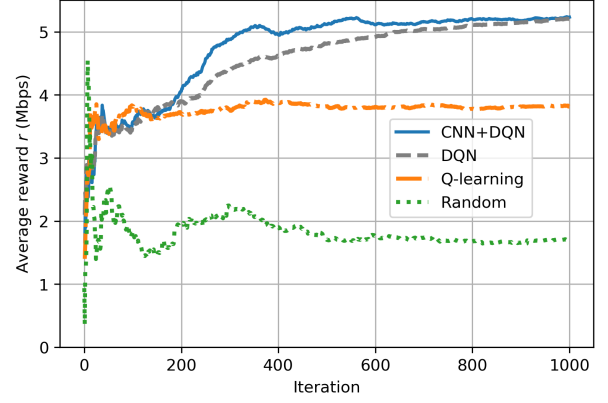
the proposed algorithm. Compared with Q-learning, DQN and the proposed CNN based DQN algorithm discard the Q-table, which contains a large number of state-action pairs. Compared with conventional DQN, our proposed algorithm invokes CNN to reduce the number of training parameters. Additionally, the proposed algorithm outperforms the conventional DQN by invoking decaying ϵ-greedy policy.
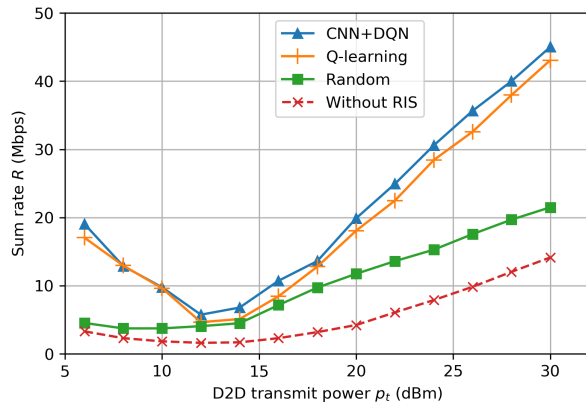
B. Sum Rate versus Transmit Power

Fig. 4(a) demonstrates the sum rate with different D2D transmit power. We can note that the schemes with RIS always outperform that without RIS in terms of achieved sum rate, which demonstrates the effectiveness of RIS. As the D2D transmit power is relatively low, the sum rate is mainly contributed by cellular links. With D2D transmit power increasing from 5dBm to 12dBm, the sum rate R decreases as the interference from D2D transmitters to the BS becomes severer. When D2D transmit power becomes more than 14dBm, the sum rate R increases with higher D2D transmit power as D2D links contribute more to the sum rate. It is also noted that our proposed CNN based DQN algorithm outperforms the Q-learning and random scheme in all the cases.

In Fig. 4(b), the sum rate versus transmit power of the cellular user shows the similar trend. The sum rate reach the bottom when the cellular transmit power is 35dBm, which means the both of the D2D receiver and the BS experience severe interference. As the transmit power of the cellular user is beyond 35dBm, the sum rate turns to increase with the higher transmit power from the cellular user.

C. Impact of the number of RIS elements

Fig. 5 demonstrates the impact of the number of RIS elements on the sum rate. In the random phase shift, we optimize the position of RIS, while randomly adjusting the phase shift. It is observed that the sum rate of the considered network increase with more number of RIS elements implemented. This is because the sum rate increases due to the improved gain. However, the performance improvement slows down as

(a) Sum rate over D2D transmit power. $p_k = 30$dBm.
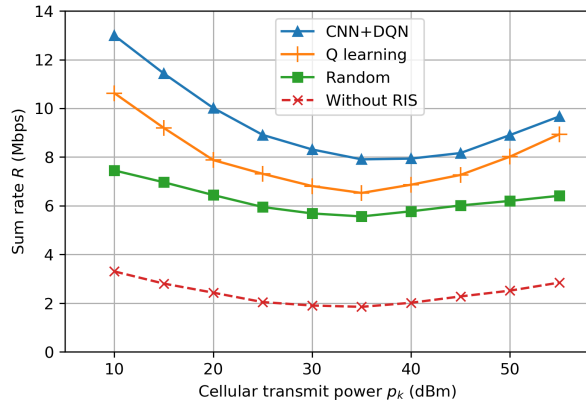
(b) Sum rate over cellular transmit power. $p_i = 15$dBm.
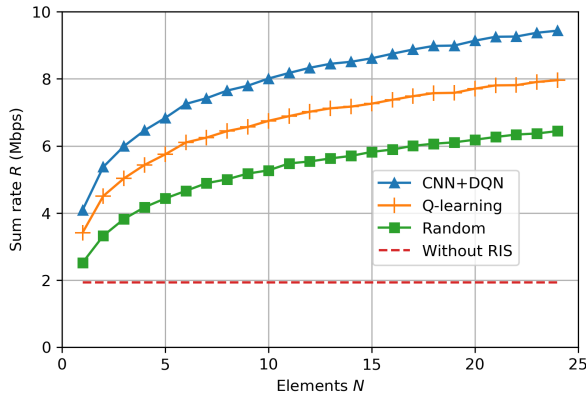
Fig. 4. Sum rate over D2D and cellular transmit power. $N = 16$.

Fig. 5. Sum rate over number of elements. $p_k = 30$dBm, $p_i = 15$dBm.

the number of elements becomes larger, which is caused by the server interference.

V. CONCLUSION

In this paper, we consider the joint optimization of position and phase shift to maximize the sum rate of the D2D network. To solve the non-convex problem, a novel CNN-combined DQN was proposed to reduce the number of weights to be

trained. By interacting with wireless communication environment and receiving real-time feedback, the RIS controller was to be able to learn an policy to optimize the installation position and phase shift of the RIS. The simulation results demonstrated that the proposed algorithm outperforms the benchmark schemes.

REFERENCES

[1] Y. Kai, J. Wang, H. Zhu, and J. Wang, "Resource allocation and performance analysis of cellular-assisted OFDMA device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 416–431, Jan. 2019.

[2] M. Robat Mili, P. Tehrani, and M. Bennis, "Energy-efficient power allocation in OFDMA D2D communication by multiobjective optimization," *IEEE Wireless. Commun. Lett.*, vol. 5, no. 6, pp. 668–671, Dec. 2016.

[3] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-Enabled vehicular communications," *IEEE Trans. on Commun.*, vol. 65, no. 7, pp. 3186–3197, Apr. 2017.

[4] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[5] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. D. Renzo, C. Chae, and L. Hanzo, "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45 913–45 923, Mar. 2020.

[6] X. Tan, Z. Sun, J. M. Jornet, and D. Pados, "Increasing indoor spectrum sharing capacity using smart reflect-array," in *Proc. IEEE Internat. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May. 2016, pp. 1–6.

[7] B. Wu, A. Sutinjo, M. E. Potter, and M. Okoniewski, "On the selection of the number of bits to control a dynamic digital MEMS reflectarray," *IEEE Antennas Wirel. Propag. Lett.*, vol. 7, pp. 183–186, Mar. 2008.

[8] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. on Commun.*, vol. 68, no. 3, pp. 1838–1851, Dec. 2020.

[9] M. A. El Mossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. on Cogn. Commun. and Netw.*, p. 11, 2020.

[10] Y. Cao and T. Lv, "Sum rate maximization for reconfigurable intelligent surface assisted device-to-device communications," *arXiv preprint arXiv:2001.03344*, Jan. 2020.

[11] M. Fu, Y. Zhou, and Y. Shi, "Reconfigurable intelligent surface for interference alignment in MIMO device-to-device networks," *arXiv preprint arXiv:2005.06766*, May. 2020.

[12] C. Pradhan, A. Li, L. Song, J. Li, B. Vucetic, and Y. Li, "Reconfigurable intelligent surface RIS-enhanced two-way OFDM communications," *arXiv preprint arXiv:2005.01910*, May. 2020.

[13] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[14] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for low-latency vehicular communications with packet retransmission," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Kansas City, MO, USA, Dec. 2018, pp. 1–6.

[15] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *arXiv preprint arXiv:1904.10136*, Apr. 2019.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[17] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *arXiv preprint arXiv:2002.10072*, Feb. 2020.

[18] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, 1997.

[19] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *arXiv preprint arXiv:2001.10363*, Jan. 2020.