Wireless Networks
https://doi.org/10.1007/s11276-024-03766-6

ORIGINAL PAPER

A deep reinforcement learning-based D2D spectrum allocation underlaying a cellular network

Yao-Jen Liang[1] · Yu-Chan Tseng[2] · Chi-Wen Hsieh[3]

Accepted: 4 May 2024

Abstract
We develop a deep reinforcement learning-based (DRL) spectrum access scheme for device-to-device communications in an underlay cellular network. Based on the DRL scheme, the base station aims to maximize the overall system throughput of both the D2D and cellular communications by learning an optimal spectrum allocation strategy. While D2D pairs dynamically access the time slots (TSs) of a shared spectrum belonging to a dedicated cellular user (CU). In particular, to ensure that the quality of service (QoS) requirement of cell-edge CUs, this paper addresses the various positions of CUs and D2D pairs by dividing the cellular area into shareable and un-shareable areas. Then, a double deep Q-network is adopted for the BS to decide whether and which D2D pair can access each TS within a shared spectrum. The proposed DDQN spectrum allocation not only enjoys low computational complexity since just current state information is utilized as input, but also approaches the throughput of exhaustive search method since received signal-to-noise ratios are utilized as inputs. Numerical results show that the proposed deep learning-based spectrum access scheme outperforms the state-of-art algorithms in terms of throughput.

Keywords Deep reinforcement learning (DRL) · Device-to-device (D2D) communications · Spectrum access · Double deep Q-network

1 Introduction

Device-to-device communications, as one of the promising techniques for the 5 G and beyond communication systems, leverage the proximity communicating and spectrum reusing. Thus, D2D communication can significantly reduce the latency and improve the spectrum efficiency without traversing the base station (BS) [1]. However, cellular users (CUs) located far away from the BS would suffer strong interference when D2D pairs share their resources without a sophisticated access mechanism.

There are many contributions dedicated to the resource allocation for D2D communications [2–5]. In Liang et al. [2], the authors proposed an algorithm of spectrum allocation and power optimization to overcome the challenges of dynamic D2D channels. A distributed spectrum allocation framework was proposed in Li and Guo [3] by adopting the actor-critic (AC) scheme to handle a decision making problem with state-action spaces. In Najla et al. [4] proposed a sequential bargaining game to determine the coalitions of the D2D pairs mutually reusing multiple channels. Furthermore, Kai et al. [5] considered a joint downlink and uplink resource allocation scheme to maximize the sum data rate of NOMA-enabled D2D groups while guaranteeing the QoS for both CUs and NOMA-enabled D2D groups. Recently, deep learning approaches have been explored in wireless communications [6–10]. Wang et al. [7] investigated the optimal policy for resource allocation in information-centric wireless networks by maximizing the spectrum efficiency based on deep reinforcement learning. Relying on the local user information and observations, multi-agent reinforcement learning

✉ Yao-Jen Liang
yaojen@niu.edu.tw

Yu-Chan Tseng
s1090369@mail.ncyu.edu.tw

Chi-Wen Hsieh
chiwenh@ccu.edu.tw

[1] Department of Electronic Engineering, National Ilan University, Yilan, Taiwan

[2] Department of Electrical Engineering, National Chiayi University, Chiayi, Taiwan

[3] Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan

Published online: 30 May 2024

(MARL) based approaches have been widely applied. Take an instance, Vu et al. [8] proposed a distributed resource allocation algorithm to overcome the dynamic environment issue in vehicular communication systems by leveraging MARL. Double deep Q-network (DDQN) was proposed in Van Hasselt et al. [6] to overcome the overestimation by decomposing action selection and action evaluation into two DQNs. Huang et al., [9] designed a DDQN-based spectrum access (D^4SA) algorithm for D2D pairs to autonomously learn an optimal policy to maximize the sum rate in an underlay cellular network. Furthermore, The authors in Ji et al. [10] combined MARL and DDQN to propose a decentralized DDQN framework for resource allocation at users and a centralized DDQN for reconfigurable intelligent surface optimization at the BS. However, the existing deep learning based allocation methods would result in excessive memory overhead and make the entire network vulnerable.

To overcome the aforementioned issue, in this work, we consider a dynamic time division duplex (TDD) network where TSs are assigned to CUs orthogonally. D2D pairs act as agents to learn whether to access a TS based on the condition of both the CUs' positions and the communication status in an underlaying manner.[1] The challenges come from two aspects. First, the channel of D2D communications varies fast and makes conventional resource allocation approaches based on alternating optimization hard to converge. Second, D2D pairs are assumed to have no information of the access behaviours of CUs. Thus,based on the deep reinforcement learning (DRL) philosophy, we adopt the double deep Q-network (DDGN) and propose a maximal throughput algorithm (MTA) for D2D spectrum access. Compared with the D^4SA algorithm proposed in Huang et al., [9], our method directly regards throughput as the target function rather than the number of accessed user links. The relationship between the Q-function and the throughput is linked through the well-known Shannon capacity. In D^4SA, previous state information is utilized as inputs, while our method just uses current state information as inputs and thus enjoys less inputs, less computational complexity, and smaller memory requirement. Furthermore, the received signal-to-noise ratios (SNRs) are taken as inputs to the DDQN in our method. Simulation results demonstrate that the proposed algorithm can achieve a much higher throughput compared with the state-of-art contributions.

The remainder of the paper is organized as follows. In Sect. 2, the system model and problem formulation are presented. In Sect. 3, the proposed DDQN MTA algorithm is investigated. In Sect. 4, simulation results and complexity analysis are described. Finally, some conclusions are given in Sect. 5.

2 System model

2.1 System description

We consider a scenario where D2D communications underlaying a cellular network with I D2D pairs share the uplink spectrum resources of K CUs. In this letter, both D2D and cellular communications are assumed to follow the TDD principal, that is, the reciprocity holds between uplink and downlink channel state information (CSI). Thus, the overhead of timely uploading CSI and other related information to the BS can be reduced. The CUs and D2D pairs access the network through allocated TSs within each frame in a repetitive way. One frame is assumed to contain T TSs. The TSs allocated to CUs are assumed orthogonal for avoidance of co-channel interference among CUs. D2D pairs learn to share the spectrum in proper TSs to ensure the QoS requirements of CUs.

Let $\alpha_{i,k}$ be the resource reuse factor of the i-th D2D pair and CU k, and $\alpha_{i,k} = 1$ if the i-th D2D pair reuses the spectrum assigned to CU k; otherwise, $\alpha_{i,k} = 0$. Set $\mathcal{I} = \{1, 2, \ldots, I\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the set of D2D pairs and the set of CUs, respectively. For clarity of explanation, at most one D2D pair can be allowed to access each of the shared TSs.[2] The signal-to-interference-plus-noise ratio (SINR) of the CU k can be expressed as

$$\gamma_k^C = \frac{P^C G^C_{B,k}}{\sigma^2 + \sum_{i\in\mathcal{I}} \alpha_{i,k} P^D G^D_{B,i}}, \tag{1}$$

where P^C and P^D are the transmit power of the CU and D2D transmitter (DT), respectively. $G^C_{B,k}$ and $G^D_{B,i}$ are the channel power gains from the CU k to BS and from DT i to BS, respectively. σ^2 is the variance of the additive white Gaussian noise. On the other hand, the SINR of the sharing the i-th D2D receiver (DR) can be described as

$$\gamma_i^D = \frac{P^D G^D_i}{\sigma^2 + \sum_{k\in\mathcal{K}} \alpha_{i,k} P^C G^C_{k,i}}, \tag{2}$$

where G^D_i and $G^C_{k,i}$ are the channel power gain from DT i to DR i and from CU k to DR i, respectively. In order to ensure the requirement of QoS of CUs, the SINR of the CU γ_k^C should be kept above a predefined threshold γ_{th}, *i.e.*, $\frac{P^C G^C_{B,k}}{\sigma^2+P^D G^D_{B,i}} \geq \gamma_{th}$. Due to the fact that $G^C_{B,k} \propto d_k^{-n}$ (e.g., $n = 3$ in

[1] The transmit powers of both the D2D and cellular users also affect the D2D-CU pairing, while this issue is not considered here but left for a future topic.

[2] The extension to the case of multiple D2D pairs is left for further research.

fully scattered environment) where d_k stands for the distance from CU k to BS. The following relationship can be acquired

$$d_k \leq \left(\frac{P_k^C}{\sigma^2 + P^D G_{B,i}^D} \right)^{1/n} \triangleq d_{th}, \tag{3}$$

where d_{th} is the threshold to guarantee the QoS of CU with shared resources if the CU is located close enough to the BS ($d_k \leq d_{th}$). Only when a CU is located in the "sharable area", a circular area within d_{th} away from the BS, D2D pairs can be admissible to access the spectrum by sharing this CU's allocated TSs. Otherwise, if the CU is far away from the BS, no D2D pair is allowed to reuse the spectrum resource of this CU. Then, the ergodic capacity of CU k and the i-th D2D pair sharing CU k at TS t can be expressed as

$$C_k^C[t] = W[k] \log_2(1 + \gamma_k^C[t]), \tag{4}$$

$$C_i^D[t] = W[k] \log_2(1 + \gamma_i^D[t]), \tag{5}$$

respectively, where $W[k]$ is the allocated bandwidth of the k-th CU. The overall capacity of the underlay D2D network can be described as

$$C_{tot} = 1/T \sum_{t=1}^{T} \left(\sum_{k=1}^{k} C_k^C[t] + \sum_{i=1}^{I} C_i^D[t] \right) \tag{6}$$

2.2 Problem formulation

Our target is to maximize the overall capacity C_{tot} in (6) of the system by optimizing the reuse vector, $\boldsymbol{\alpha} = [\alpha_{1,1}, \dots, \alpha_{1,K}, \dots, \alpha_{I,1}, \dots, \alpha_{I,K}]^T$ as follows.

$$\max_{\boldsymbol{\alpha}} C_{tot} \tag{7a}$$

$$\text{subject to } \gamma_k^C \geq \gamma_{th}, \forall k \in \mathcal{K} \tag{7b}$$

$$\alpha_{i,k} \in \{0, 1\}, \forall i \in \mathcal{I}, k \in \mathcal{K} \tag{7c}$$

$$\sum_{k \in \mathcal{K}} \alpha_{i,k} \leq 1, \tag{7d}$$

$$\sum_{i \in \mathcal{I}} \alpha_{i,k} \leq 1, \tag{7e}$$

where constraints (7d) and (7e) assume that each D2D pair only reuses one spectrum and at most one D2D pair can be allowed to transmit information at each TS. Constraints (7c)–(7e) make the above optimization problem non-convex. For solving this non-covex problem, one has to use an exhaustive searching, which is impractical on a large number of D2D pairs and/or CUs. This motivates us to alternatively leverage the DDQN framework for dynamic spectrum access, which is applicable to solve the above problem with a large number of D2D pairs and/or CUs as well as many states and action dimensions.

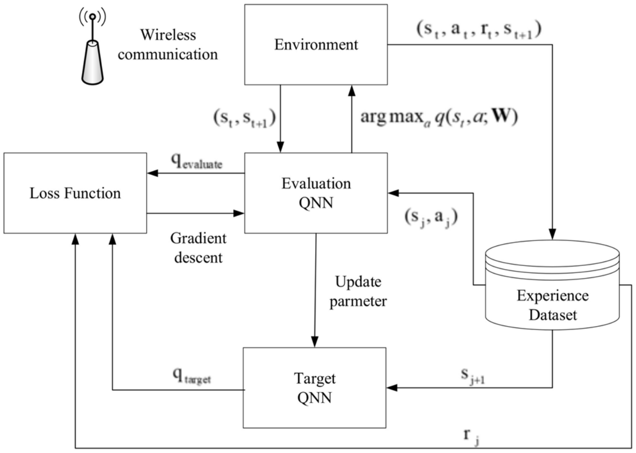


Fig. 1 The interaction of the DDQN at the centralized agent

3 Proposed algorithm

3.1 Reinforcement learning and DQN

Reinforcement learning (RL) can be modeled as a Markov decision process [10], including an environment state $\mathcal{S}$, an action $\mathcal{A}$, and a reward $\mathcal{R}$ which is evaluated from each state-action pair. At each training step p, the agent observes the state s_p and responses an action a_p according to a certain policy π. Then, the agent receives a corresponding reward r_p and transfers to the next state s_{p+1}, which is determined by the current state s_p and action a_p. This process can be denoted by a transfer tuple $e_p = (s_p, a_p, r_p, s_{p+1})$. Figure 1 shows the interaction process for the centralized DDQN agent at the BS. This centralized agent decides whether to access and which DT to transmit at the current TS. If a DT is selected to transmit, it will receive feedback after the transmission and then inform the agent with the results. During the training step p, RL agent aims to learn an optimal policy to maximize the cumulative weighted reward, which is expressed as

$$R_p = \sum_{\tau=0}^{\infty} \beta^\tau r_{p+\tau}, \tag{8}$$

where $0 \leq \beta \leq 1$ is the discount factor to indicate the impact of the future rewards. The expected reward of a state-action pair (s, a), a.k.a. action-value function, can be defined as

$$q^{\pi}(s,a) = E_{\pi}[R_p | s_p = s, a_p = a], \tag{9}$$

where the policy π represents a mapping from state $\mathcal{S}$ to the probability that each action in $\mathcal{A}$ is selected. Then, the optimal policy π^* can be described as $\pi^* = \arg\max_{\pi}\{q^{\pi}(s,a)\}$.

Q-learning, one of the typical RL algorithms, maintains a Q-value table of the action-value function to find the optimal policy π^*, *i.e.*, $Q^*(s,a) = \max_{\pi} q^{\pi}(s,a)$. However, as the cardinality of state set $\mathcal{S}$ and/or action set $\mathcal{A}$ is large, the resources for the convergence of the Q-learning will be huge and difficult for implementation. Alternately, deep Q-network (DQN) introduces a Q neural network (QNN) to approximate the Q-value instead of to maintain a Q-table. That is, given an input state s_p, QNN outputs an estimated Q-value for all possible actions, that is, $q(s_p, a; \mathbf{W}) \approx Q^*(s_p, a), \forall a \in \mathcal{A}$, where $\mathbf{W}$ stands for the parameters of QNN. The training data set $\mathcal{D} = [\mathbf{e}_1, \mathbf{e}_2, \ldots]$ for QNN is stored according to the agent's experience $\mathbf{e}_p = (s_p, a_p, r_p, s_{p+1})$ at each training step p. In the sequel, a training batch of experience $\mathbf{e}_p$ is sampled randomly from the data set $\mathcal{D}$. The training process is used to minimize the loss function, which is defined as the mismatch between the target Q-value and the realistic Q-value,

$$Loss(\mathbf{W}) = E[(q_{target} - q(s,a;\mathbf{W}))^2], \tag{10}$$

$$\text{where } q_{target} = r_p + \beta \max_{a'} q(s_{p+1}, a'; \mathbf{W}'). \tag{11}$$

Here q_{target} is the target Q-value of the target network with weights $\mathbf{W}'$. Moreover, the weights $\mathbf{W}$ ($\mathbf{W}'$) can be iteratively updated by the gradient descent method [11] as

$$\mathbf{W} = \mathbf{W} + \delta E[(q_{target} - q(s,a;\mathbf{W}))\nabla q(s,a;\mathbf{W})], \tag{12}$$

where δ stands for the updating constant and $\nabla q(s,a;\mathbf{W}) = q(s_p, a_p; \mathbf{W}) - q(s_{p-1}, a_{p-1}; \mathbf{W})$ is the backwards difference operator.

3.2 DDQN algorithm

It is well-known that DQN can achieve near-optimal solution in some scenarios, while it sometimes companies with the issue of overestimation, *i.e.*, the target Q-value may be higher than the true optimum action-value. To overcome this issue of overestimation, DDQN is proposed by decomposing the original deep Q-network into an action selection network and an action evaluation network [6], that is, DDQN uses a target Q-network for the action evaluation and an evaluation Q-network for the action selection. Whereas the target Q-value of DDQN can be acquired as

$$q_{target} = r_p + \beta q(s_{p+1}, \arg\max_{a' \in \mathcal{A}} q(s_{p+1}, a'; \mathbf{W}); \mathbf{W}'). \tag{13}$$

Here we adopt DDQN to design our proposed algorithm for D2D underlay networks. Rather than distance based information used in D^4SA of [9], CSI based information can enhance the robustness of DDQN model of a underlay D2D network [10]. Then, the definitions of "state", "action", and "reward" of our proposed DDQN are given in the following.

(1) state: The agent observes the wireless environment by listening to the channel state c_p after taking action a_p. The channel state is defined as $c_p \in \{\mathcal{I}, \mathcal{S}, \mathcal{R}, \mathcal{F}\}$, where $\mathcal{I}$ means no transmission, $\mathcal{S}$ means just one transmission at the instant TS, $\mathcal{R}$ means a sharing between a D2D pair and a CU, while $\mathcal{F}$ represents that D2D pairs reuse a TS and cause the QoS requirement of the CU unsatisfied due to severe interference. The observation space of the centralized agent includes: the channel state c_t, the channel power gain $G^C_{B,k}, G^C_{i,k}, G^D_i, G^D_{k,i}$ and noise variance σ^2 in (1) and (2). Thus, the environment state at step p can be expressed as $s_p = (c_p, G^C_{B,k}, G^C_{i,k}, G^D_i, G^D_{k,i}, \sigma^2)$.

(2) action: The action set $\mathcal{A}$ is defined to reflect which DT to transmit at the current TS as $\{0, 1, \ldots, I\}$, where I is the number of D2D pairs. The action element $a_p = 0$ means no transmission from D2D pairs, while $a_p = i \neq 0$ means that DT i transmits signals at step p.

(3) reward: To ensure the SINR requirement of CUs, *i.e.*, (7b)–(7e) are guaranteed, the reward vector at time t is defined as $\mathbf{r}[t] = [r^D_1[t], \ldots, r^D_I[t], r^C_1[t], \ldots, r^C_K[t]]^T$, where the element $r^D_i[t]$ and $r^C_k[t]$ represent of the reward of the i-th D2D pair and of the k-th CU, respectively. Here, they are defined as $r^C_k[t] = C^C_k[t], \forall k \in \mathcal{K}$ and

$$r^D_i[t] = \begin{cases} C^D_i[t], \forall i \in \mathcal{I}, & \text{if } (7b) - (7e) \text{ are satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

During the training phase, each epoch contains several training steps wherein the agent interacts with the environment and stores the experience in the training data set. The pseudo codes of the proposed DDQN scheme are shown as in Algorithm 1.

Algorithm 1 DDQN for Resource Allocation

```
Initialize $\mathbf{W}$ and $\mathbf{W}'$.
for each training step do
    Input state $s_p$.
    Choose action $a_p$ according to the state $s_p$ and
     $\epsilon$-greedy algorithm.
    Form the action $a_p$, reward $\mathbf{r}_p$ and the next state $s_{p+1}$.
    Store the experience $e_p = (s_p, a_p, \mathbf{r}_p, s_{p+1})$
     in memory $\mathcal{D}$.
    $\mathbf{W}' \leftarrow \mathbf{W}$ every $C$ steps.
    Sample random $N_E$ batch of experiences from $\mathcal{D}$ as $\mathcal{E}$.
    Set $e = (s, a, \mathbf{r}, s')$.
    for each $e$ in $\mathcal{E}$ do
        $a' = \arg\max_a \sum_{l=1}^{L+K} \log(q^{(i)}(s', a, \mathbf{W}))$
        Calculate $q_{target}$ in (11).
    end for
    Update $\mathbf{W}$ and $\mathbf{W}'$ as in (12).
end for
```

Table 1 Simulation parameters

Parameter	Value
carrier frequency	2 GHz
bandwidth of each sub-channel	1MHz
CU Tx power, P^C	23 dBm
D2D Tx power, P^D	20 dBm
radius of BS coverage, R	300 m
distance between D2Ds per pair	Uniformly distributed in (5, 15) m
small-scale fading	i.i.d. complex Gaussian distributed with zero mean and unit variance
noise power spectral density	− 175 dBm/Hz
SINR requirement for CU, γ_{th}	5 dB
number of CUs, K	4, 6
number of D2D pairs, I	2

4 Simulation results

This section demonstrates the performance of the proposed algorithm and compare it with four benchmarks: overlay approach, D^4SA algorithm in [9], IF algorithm of our previous work [12] and exhaustive search method. Before describing the environment settings, we briefly discuss the considered algorithms for comparison. In the overlay approach, CUs and D2D pairs exclusively use orthogonal resource blocks and thus no resource sharing is allowed. In [12], the IF method aims to maximize the number of admissible D2D pairs by examining whether the SINR requirements of all the accessed CUs and D2D users can be met if a new D2D pair is admitted to access the network. While the exhaustive search method combinatorially select the optimal pairing among the possible arrangements which meet all the interference requirements of the network. In the considered environment, the locations of CUs and D2D pairs are randomly deployed in the cell coverage at each TS and the probability of CUs in the "shareable" area is approximated as $\frac{d_{th}^2}{R^2}$, where R denotes the radius of the cell. The simulation parameters are listed in Table 1.

Each DQN of the proposed DDQN consists of a 5-layer fully connected neural network (FCNN) with 3 hidden layers and each hidden layer has $(8KI)$ neurons.[3] The rectified linear unit (ReLU) function, *i.e.*, $f(x) = \max(0, x)$, is applied as the activate function. For other parameters, the updating constant in (12) δ is set to 0.01, and the discount factor β is 0.95. The exploration ϵ in ϵ-greedy algorithm is set to 1 at the beginning and decreases to 0.005 with step-size of 0.005. The update period C for the target DQN is 10. The batch size of experiences from $\mathcal{D}$, N_E, is set to 30.

4.1 Loss function and throughput

In Fig. 2, the values of loss function in (10) for various transmission probabilities of D2D users are compared. We can find that the loss values converge in the training phase as the

[3] The number of neurons of the hidden layer should be greater than the number of inputs to the DQN to prevent information loss during training, however the optimal tradeoff between the number of neurons and the computational complexity is beyond the scope of this work.

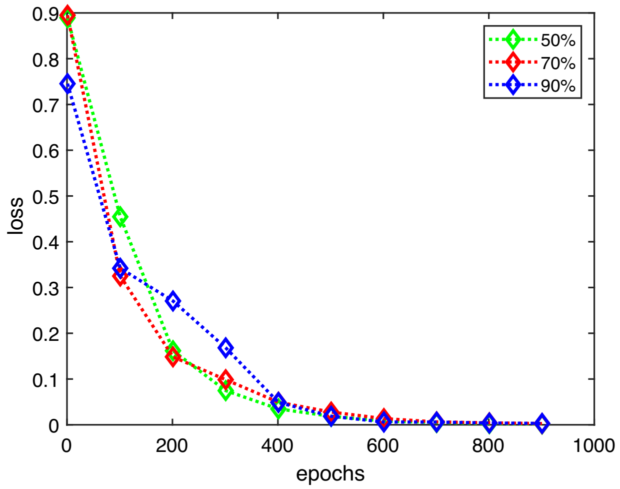


Fig. 2 Training performance comparison among various transmission probabilities of D2D users ($K = 4, I = 2$)

number epoches larger than 700 while the effect of various transmission probabilities is little. In Fig. 3, the achievable throughput among the above considered five methods is demonstrated for various D2D transmission probabilities. From this figure, one can find that the considered four underlay methods all outperform the overlay counterpart. Furthermore, the exhaustive search method enjoys the best performance at the cost of huge computational complexity, which is given in Table 2. However, our proposed DDQN outperforms D^4SA [9] and overlay approach with the same order of computational complexity. The main reason comes from that throughput is directly used as the reward in our proposed DDQN method rather than the number of accessed user links in D^4SA. It is noteworthy that D^4SA utilizes dozens of length of state history,

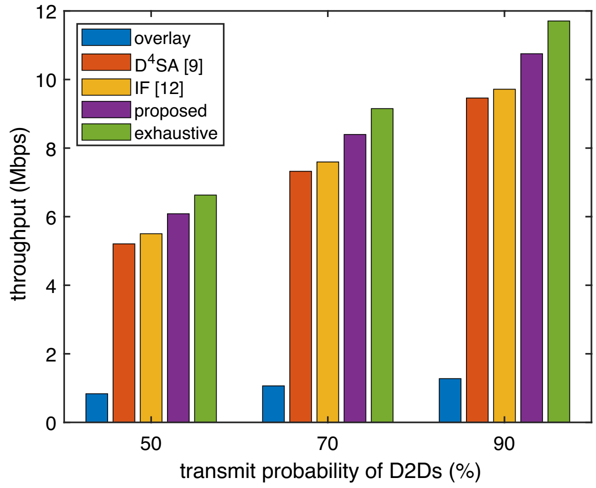


Fig. 3 Training throughput among various methods ($K = 4, I = 2$)

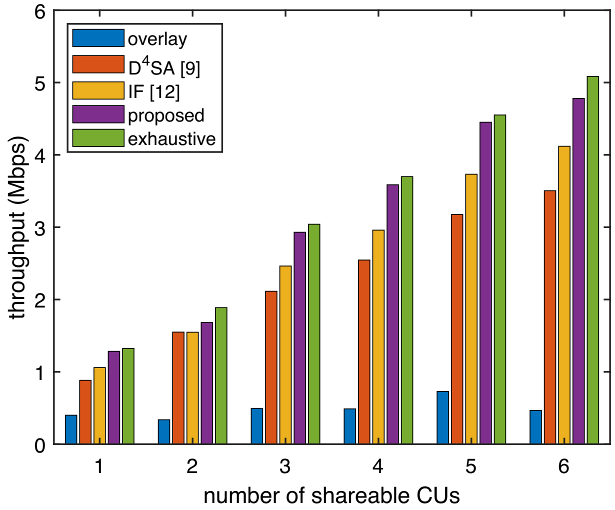


Fig. 4 Training throughput among various methods for different shareable CU numbers. ($I = 4$)

Table 2 Complexity comparison among various spectrum allocation schemes

	Proposed	D^4SA	IF	Exhaustive
Offline training	$KI^2N_EN_{epo}$	$KI^2N_EN_{prv}N_{epo}$	–	–
Online testing	KI	KIN_{prv}	K^2I	K^I

while the proposed DDQN just uses the current state for much less inputs and neurons within each layer.

In Fig. 4, the transmission probability is set to 0.5. From this figure, one can also find that the gap of throughput between the proposed DDQN and D^4SA increases as the number of shareable CUs increases. The improvement of the proposed DDQN over D^4SA is more than 30% as the number of sharable CUs is 6. This phenomenon reveals that the reward function plays a critical role of resource allocation in deep learning approaches.

4.2 Complexity analysis

The computational complexity of the exhaustive search method is of order $\mathcal{O}(K^I)$, while the complexity of IF is on the order of $\mathcal{O}(K^2I)$ [12]. On the contrary, the complexity of the proposed method in training phase is seen to be dominated by the evaluation of q_{target} in (11), which is of order $\mathcal{O}(KI^2N_E)$ per epoch. Furthermore, the complexity of the proposed method in testing phase is just by the multiplication of FCNN, which is of order $\mathcal{O}(KI)$. The complexity comparison is listed as in Table 2, where N_{epo} and N_{prv} stand for the number of training epochs and the length of previous state information required in D^4SA, respectively. It should be noted that the computational complexity of the IF and exhaustive methods is only evaluated in "Online Testing", since no training procedure is needed for these two approaches.

5 Conclusion

In this letter, we aim to maximize the overall throughput of the D2D pairs and the cellular users for the D2D communication underlay cellular networks. Based on the DRL philosophy, a novel centralized double deep Q-network (DDQN) is proposed to solve the non-convex problem with low complexity. Moreover, leveraging of the CSI-based information, the simulation results show that our proposed algorithm can outperform other DQN and non-learning approaches in terms of the achievable throughput. For further research, power control and more complex sharing principles among D2D pairs and CUs can be included to enrich the communication environment.

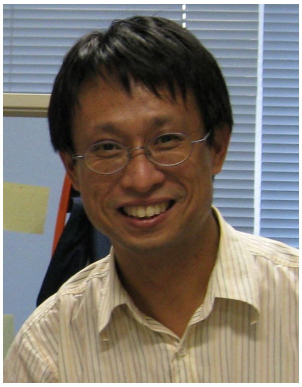
References

1. Cotton, D., & Chaczko, Z. (2021). Gymd2d: A device-to-device underlay cellular offload evaluation platform. In *IEEE wireless communications and networking conference (WCNC), 2021*, 1–7.
2. Liang, L., Li, G. Y., & Xu, W. (2017). Resource allocation for D2D-enabled vehicular communications. *IEEE Transactions on Communications, 65*, 3186–3197.
3. Li, Z., & Guo, C. (2019). Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications. *IEEE Transactions on Vehicular Technology, 69*(2), 1828–1840.
4. Najla, M., Becvar, Z., & Mach, P. (2021). Reuse of multiple channels by multiple d2d pairs in dedicated mode: A game theoretic approach. *IEEE Transactions on Wireless Communications., 20*, 4313–4327.
5. Kai, C., Wu, Y., Peng, M., & Huang, W. (2021). Joint uplink and downlink resource allocation for NOMA-enabled D2D communications. *IEEE Wireless Communications Letters, 10*, 1247–1251.
6. Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 30*(1), 2094–2100.
7. Wang, D., Qin, H., Song, B., Du, X., & Guizani, M. (2019). Resource allocation in information-centric wireless networking with D2D-enabled MEC: A deep reinforcement learning approach. *IEEE Access, 7*, 114935–114944.
8. Vu, H. V., Liu, Z., Nguyen, D. H. N., Morawski, R., & Le-Ngoc, T. (2020). Multi-agent reinforcement learning for joint channel assignment and power allocation in platoon-based c-v2x systems. arXiv:2011.04555.
9. Huang, J., Yang, Y., He, G., Xiao, Y., & Liu, J. (2021). Deep reinforcement learning-based dynamic spectrum access for d2d communication underlay cellular networks. *IEEE Communications Letters, 25*(8), 2614–2618.
10. Ji, Z., Qin, Z., & Parini, C. G. (2022). Reconfigurable intelligent surface aided cellular networks with device-to-device users. *IEEE Transactions on Communications, 70*, 1808–1819.
11. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning, pp. 1–9. arXiv:1312.5602
12. Liang, Y.-J., & Lin, Y.-S. (2017). A non-iterative resource allocation strategy for device-to-device communications in underlaying cellular networks. *Wireless Network, 23*, 2485–2497.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yao-Jen Liang received the Ph.D. degree in Communication Engineering from National Taiwan University, Taiwan, in 2010. He has more than six years experience in industry. From February 2011 to February 2012, he was a visiting scholar at the School of Electrical and Computer Engineering, Georgia Institute of Technology. From 2012 to 2022, he was an assistant and associate professor with the Department of Electrical Engineering, National Chiayi University, Chiayi, Taiwan. He is currently an associate professor with the Department of Electronic Engineering, National Ilan University, Yilan, Taiwan. His current research interests include MIMO and OFDM systems, wireless networks, and statistical signal processing.

Yu-Chan Tseng Yu-Chan Tseng received his B.S. Degree from the Department of Electrical Engineering from Tamkang University, Tamsui, Taiwan, in 2020, and master's degree in Electrical Engineering from National Chiayi University, Chiayi, Taiwan, in 2022. His research interest focuses on spectrum allocation in wireless communications.



Chi-Wen Hsieh received the B.Sc. degree from the Department of Physics, Fu-Jen Catholic University, New Taipei, Taiwan, in 1990, and the master's and Ph.D. degrees from the Department of Electrical Engineering, National Tsing-Hua University, Hsinchu, Taiwan, ROC, in 1993 and 2007, respectively. In 2018–2022, he has been a professor and the Chair of the Department of Electrical Engineering, National Chiayi University, Chiayi, Taiwan. And he got a distinguish professor title in Jan, 2022 and to be a team leader of Lab of digital signal processing of NCYU. He is professor and currently working at Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan, ROC. His research interests include medical signal processing, microwave applications, and image processing.