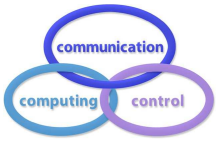
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 2, Month: April, Year: 2021
Article Number: 4189, https://doi.org/10.15837/ijccc.2021.2.4189



CCC Publications

A Resource Allocation Algorithm for Ultra-Dense Networks Based on Deep Reinforcement Learning

H. S. Zhang, T. M. Wang, H. W. Shen

Huashuai Zhang
College of Applied Science and Technology
Beijing Union University
Beijing 100101, China
20202045111106@buu.edu.cn

Tingmei Wang*
College of Applied Science and Technology
Beijing Union University
Beijing 100101, China
*Corresponding author: yykjttingmei@buu.edu.cn

Haiwei Shen
College of Applied Science and Technology
Beijing Union University
Beijing 100101, China
yykjthaiwei@buu.edu.cn

Abstract

The resource optimization of ultra-dense networks (UDNs) is critical to meet the huge demand of users for wireless data traffic. But the mainstream optimization algorithms have many problems, such as the poor optimization effect, and high computing load. This paper puts forward a wireless resource allocation algorithm based on deep reinforcement learning (DRL), which aims to maximize the total throughput of the entire network and transform the resource allocation problem into a deep Q-learning process. To effectively allocate resources in UDNs, the DRL algorithm was introduced to improve the allocation efficiency of wireless resources; the authors adopted the resource allocation strategy of the deep Q-network (DQN), and employed empirical repetition and target network to overcome the instability and divergence of the results caused by the previous network state, and to solve the overestimation of the Q value. Simulation results show that the proposed algorithm can maximize the total throughput of the network, while making the network more energy-efficient and stable. Thus, it is very meaningful to introduce the DRL to the research of UDN resource allocation.

Keywords: ultra-dense networks (UDNs), deep reinforcement learning (DRL), resource allocation, throughput, energy efficiency

1 Introduction

In recent years, wireless resource allocation has become an important issue, with the proliferation of long-term evolution (LTE) systems and fifth generation (5G) communication systems. Meanwhile, the recent surge in business demand of users poses severe challenges to resource optimization. The White Papers of Cisco show that the amount of wireless network data is increasing exponentially, and is expected to grow by 1000% in the next decade.

To improve the service quality and business experience of the network, scholars have shifted their attention towards ultra-dense networks (UDNs). With the rapid development of network technology, lots of small cells could now be deployed on UDNs [8]. However, the growing density of base stations increases the total power consumption of base stations, intensify the interference between small base stations in the network, and make these base stations more unstable [27]. To solve the problems arising from the largescale deployment of small cell base stations, it is of great significance to allocate wireless resources by improving the energy efficiency and stability of the UDNs.

The existing research on UDN wireless resource allocation mainly focuses on increasing the total network throughput and improving network energy efficiency. Liu [17] maximized the total network capacity with a distributed resource allocation algorithm, satisfied the requirement on quality of service (QoS), and optimized the sub-channel allocation and power allocation through geometric programming. Wang et al. [30] adopted a non-cooperative game model with penalty factors, proposed a way to virtualize the local information of small cells for power allocation, and optimized network energy efficiency through effective control of small cell base stations, in the light of their load conditions. Drawing on Nash bargaining and cooperative game theory, Zhang et al. [36], Zhang et al. [35], and Zhang [34] studied the power allocation of sub-channels in small cell networks, and solved the game model by the Lagrangian dual decomposition method. In this way, the model converged to the Pareto optimal solution, maximizing the energy efficiency of the entire system.

In the above research, the wireless resources are mostly allocated by traditional methods, which involve a huge computing load during implementation. In the scenario of UDNs, the growing number of small cell base stations will continuously push up the computing complexity. To simplify the algorithm, many researchers have resorted to wireless resource allocation through reinforcement learning (RL).

So far, fruitful results have been achieved in the application of RL in wireless communication. Amiri et al. [3] applied RL to power allocation of wireless networks, allocated the power of network base stations reasonably through online learning, and thus maximized the total throughput of the network. Chen et al. [6] integrated wireless network with RL, and implemented Q-learning to rationalize the allocation of spectrum resources and increase network throughput. However, these two RL approaches can only optimize network throughput, failing to consider the energy efficiency of the network.

Nevertheless, the RL could not effectively handle the ultra-large state space of UDNs. This gives prominence to deep reinforcement learning (DRL). Many scholars have probed extensively into the DRL [21]. For example, Chang et al. [5] developed a DRL-based architecture for fog network, which offloads the transcoding tasks in network services to fog nodes under the dynamics of wireless networks, strikes a balance between high QoS and low delay, and thereby offers the best user experience. Xiong [32] investigated the slice-level resource reservation and physical resource allocation based on the DRL algorithm, and constructed an automatic resource management system for wireless virtual networks; the constructed system can maximize the resource utilization of the whole system, without sacrificing the quality of user services. He et al. [11] discussed the actual channel state under the interference alignment mechanism, and obtained the optimal user selection strategy under the cache condition, using the DRL algorithm; their strategy eliminates the interference to the signal, and significantly improves the total throughput and energy efficiency of the network.

Currently, the UDNs mainly face the following problems due to the dense deployment of small cell base stations and the huge network space: inter-signal interference, slow algorithm convergence, and poor signal stability. These problems could be alleviated by coupling the DRL technology with concepts like empirical repetition and target network.

This paper proposes a network resource allocation algorithm based on DRL in the context of UDNs, aiming to increase network throughput and speed up convergence. Firstly, a model was established to maximize the total system capacity response to the interference caused by the dense deployment of

small cells in UDN. Next, the DRL algorithm was adopted to make the allocation of wireless resources more efficient, with the goal to solve the complex resource allocation problem in UDN. Finally, the deep Q-network (DQN) was employed for resource allocation, and the concepts of empirical repetition and target network were introduced. Our algorithm stores all operation processes in the memory pool, and randomly selects the data in the memory pool for training. This effectively solves the correlations between samples, and overcomes the instability and divergence of the results caused by the previous network state, thereby improving stability [5]. In the iterative process, the estimation network is updated iteratively at any time according to the current state and action. Then, at every interval, the estimation network assigns the network parameters to the target network, such as to avoid the overestimation of the Q-value [32].

The research contents are detailed in four parts below: First, a small cell network scenario was summarized, and the relevant hypotheses and model were established for the scenario. Next, the DRL was introduced to convert the research problem, and to analyze the corresponding algorithm. Afterwards, an application scenario was created to simulate the algorithm, and to compare our algorithm with other algorithms; the simulation results were analyzed reasonably. Finally, several findings were drawn from the whole process of the research.

2 Hypotheses and modeling

This paper summarizes the system model of an ultra-dense small cell network as Figure 1.

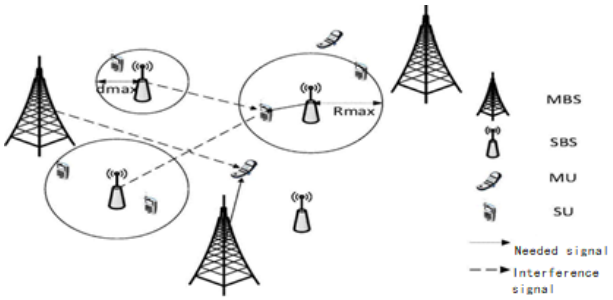


Figure 1: System model

As shown in Figure 1, the base stations are distributed densely, so that the networks of the same frequency cannot be used by users in the same cell, but by those in other cells. On the architecture of communication network, the network system should consider the interference caused by all base stations to users, due to the dense distribution of multiple base stations.

Suppose a core controller can collect relevant information in the entire network, including transmission power and signal-to-noise ratio (SNR); each user can transmit location information, interference, and transmission rate to the core controller through pilot signals, and make a unified plan for spectrum allocation.

Let $n = \{1, 2, 3, \ldots, N\}$ be the $n$small cells in the system, each of which uses k subcarriers $k = \{1, 2, 3, \ldots, K\}$. The distribution of base stations follows the Poisson point process model, multiplexing k orthogonal subcarriers. During the access to the base station, each cell user adopts the transmission power of P_t^n. At time t, the cell user can only connect to one base station; each subcarrier can only be allocated to one user.

At time t, in the small cell n, the signal to interference plus noise ratio (SINR) on the k-th subcarrier can be expressed as:

$$SINK_n^k = \frac{G_n^k P_t^{(n,k)}}{\sum_{n \neq \acute{n}} G_{\acute{n}}^k P_t^{(\acute{n},k)} + \sigma^2} \tag{1}$$

where, $P_t^{(n,k)}$ is the total transmission power of small cell n on subcarrier k; G_n^k is the channel gain; σ^2 is white Gaussian noise (WGN). Hence, the channel capacity of small cell n on channel k can be

obtained as:

$$R_n^{k=}\frac{B}{K}\log_2\left(1+SINK_n^k\right) \tag{2}$$

where, B is the total bandwidth of the system. Thus, the maximum total throughput of the system can be expressed as:

$$\begin{aligned}&\text{argmaxR}=\sum_{n=1}^{N}\sum_{k=1}^{K}R_n^k b_n^k\\&\text{s.t.C1: } P_t^{(n,k)}\geq P_{\min},\forall \text{n, k}\\&\text{C2: } \sum_{n,k}P_t^{(n,k)}\leq P_{\max},\forall \text{n, k}\end{aligned} \tag{3}$$

where, $b_n^k=1$ or 0 is an indicator of whether the small cell base station n allocates subcarrier k to the user (if $b_n^k=1$, the subcarrier k of small cell $n$is occupied; if $b_n^k=0$, the subcarrier k of small cell n is not occupied); P_{min} and P_{max} are the minimum and maximum transmission powers required by the subcarrier, respectively. The overall capacity of the entire system can be enhanced by adjusting the transmission power $P_t^{(n,k)}$of the base stations on the subcarriers. Note that the water injection algorithm was adopted to allocate the initial power between the subcarriers, for this algorithm can maximize the transmission rate by reasonably allocating the transmission power as per the planned criteria and the real-time channel conditions [11].

During channel allocation, the power consumption of each small cell base station is divided into two parts: the minimum fixed power consumed to maintain the operation of the station, and the power consumed to serve users. In the period t, the power consumption of the small cell base station n can be expressed as:

$$P'_{(t,n)}=P_n^{op}+P_t^{(n,k)}b_n^k \tag{4}$$

where, P_n^{op} is the fixed operating loss of small cell base station n; $P_t^{(n,k)}$ is the transmission power of small cell base station n in period t. Energy efficiency, as the ratio of total throughput to the total power consumption, can effectively measure the network performance. Therefore, the total energy efficiency of the system at time t can be expressed as:

$$EE_{\text{total}}^t=\frac{\max R^t}{\sum_{n=1}^{N}P'_{(t,n)}} \tag{5}$$

To optimize the energy efficiency of the network, it is necessary to control the carrier allocation and transmission power of small cell base stations. However, this non-convex optimization problem cannot be efficiently solved by traditional algorithms in a short time. What is worse, the traditional algorithms cannot make real-time adjustment according to the network environment. To solve the problem, artificial intelligence (AI) techniques like DRL provides a possible tool to intelligently and effectively enhance user satisfaction with QoS and the utilization rate of network resources [25].

3 Problem transformation and algorithm analysis

3.1 Basic model of RL

The RL is an important tool widely used to solve Markov dynamic programming and other problems [39]. During the RL, the learner can acquire the optimal strategy by interacting with the complex environment. According to the network state, the learner takes an action by a certain strategy. Once taken, the action will change the network state, and receive a reward. Then, the strategy will be further optimized. This process will be repeated until the strategy is optimal. Hence, network state, action, and reward are three key factors of the RL framework. In this paper, the three factors are defined as follows:

(1) State space

The state space is defined as $S_{(t)}=\left\{B_1^{(t)},C_1^{(t)},P_1^{(t)},\ldots,B_n^{(t)},C_n^{(t)},P_n^{(t)}\right\}$, where $B_1^{(t)}$ is the traffic volume of small cell base station at time t; $C_1^{(t)}$ is the channel state at time t; $P_1^{(t)}$ is the transmission

power of small cell base station at time t. To reduce the complexity of the algorithm and the state space of the network, the transmission power of the base station needs to be discretized by:

$$\begin{array}{c} P_n^{(t)} = \tau \\ \left(P_{\max}^f - A_\tau\right) \leq \sum_{k=0}^{k} P_t^{(n,k)} < \left(P_{\max}^f - A_{\tau+1}\right) \end{array} \tag{6}$$

where, $\tau \in \{0,1,2,3,4,5\}$; $A_0 = P_{max}^f$; $A_6 = 0$; $A_1 \sim A_5$ are arbitrary thresholds.

(2) Action space

The action space is defined as $A_{(t)} = \left\{U_{in}^{(t)}, P_{in}^{(t)}, T_{nk}^{(t)} \mid \mathrm{n} \in \mathrm{N}, \mathrm{k} \in \mathrm{K}\right\}$, where $U_{in}^{(t)}$ is the indicator of whether small cell base station needs to be connected with user i at time t; $P_{in}^{(t)}$ is the power allocation by small cell base station n to user i at time t; $T_{nk}^{(t)}$ is the adjusted value of the energy efficiency on the k-th subcarrier of base station n at time t.

(3) Reward

To optimize energy efficiency, the reward function can be set as the energy efficiency at time t:

$$r_t = EE_{total}^t = \frac{\max R^t}{\sum_{n=1}^{N} P'_{(t,n)}} \tag{7}$$

3.2 DRL strategy

To maximize the throughput of the entire network, it is a must to maximize the cumulative reward by choosing a suitable method for allocating network resources. The Q-learning presents an effective way for RL, and has been widely applied by researchers [29]. The significance of Q-learning is to search for the optimal strategy π^* that maximizes the cumulative discount return in the long term.

In each period t, the learner belongs to the state of $S_{(t)}$, and executes the action $A_{(t)} = \pi\left(S_{(t)}\right)$ corresponding to strategy π. He/she will be rewarded with $r_t\left(S_{(t)}, A_{(t)}\right)$, depending on the current state and action. Then, the next state will be obtained. The strategy will be optimized through iterative learning by the small cell base station.

Based on strategy π, the expected long-term return $V^\pi\left(S_{(t)}, A_{(t)}\right)$ of the state-action strategy can be expressed as:

$$V^\pi\left(S_{(t)}, A_{(t)}\right) = E_\pi\left[\sum_{i=1}^{I} \lambda_i r_i\left(S_{(i)}, A_{(i)}\right) \mid S_{(1)} = S_{(t)}, A_{(1)} = A_{(t)}\right] \tag{8}$$

where, $r_i\left(S_{(i)}, A_{(i)}\right)$ is the instant reward generated by performing action $A_{(i)}$ in state $S_{(i)}$; λ is a discount factor $(0 < \lambda < 1)$.

To maximize the long-term return $V^\pi\left(S_{(t)}, A_{(t)}\right)$, the optimal state value function can be expressed as:

$$\pi^*\left(S_{(t)}\right) = \max \mathrm{Q}\left(S_{(t)}, A_{(t)}\right) \tag{9}$$

The optimal strategy can be obtained through iterative update of the optimal Q value:

$$\mathrm{Q}\left(S_{(t)}, A_{(t)}\right) \leftarrow \left(1-\alpha\right)\mathrm{Q}\left(S_{(t)}, A_{(t)}\right) + \alpha\left[r_t\left(S_{(t)}, A_{(t)}\right) + \lambda \max \mathrm{Q}\left(S_{(t+1)}, A_{(t+1)}\right)\right] \tag{10}$$

where, α is the learning rate that affects the update speed of Q value $(0 < \alpha < 1)$.

4 DQN

In the real-world UDN, both state space and action space are huge. It is difficult for Q-learning to achieve desirable results. Therefore, the DQN algorithm that combines deep learning with Q-learning can effectively overcome this defect. Apart from the DQN, the authors also introduced two novel concepts: empirical repetition and target network.

During Q-learning, the Q-value might be overestimated, for the strategy selection and update are based on the same Q function table [2]. The target network was introduced to prevent the overestimation. There are two neural networks with different parameters in DQN. One of them is an estimation network, i.e., an optimal value function $\mathrm{Q}\left(S_{(t)}, A_{(t)} \mid \theta\right) \approx \mathrm{Q}\left(S_{(t)}, A_{(t)}\right)$, with network weight θ, fitted from the currently inputted state and action; the other is the target network, whose target value can be calculated by:

$$\mathrm{r} = r_t\left(S_{(t)}, A_{(t)}\right) + \lambda \max \mathrm{Q}\left(S_{(t+1)}, A_{(t+1)} \mid \theta^{-}\right) \tag{11}$$

where, θ^{-} is the parameter of the target network. Any small deviation of the function value will affect the entire strategy. To stabilize $\mathrm{Q}\left(S_{(t)}, A_{(t)} \mid \theta\right)$, the estimation network needs to be trained in each step to minimize the loss function, so as to approximate the actual $\mathrm{Q}\left(S_{(t)}, A_{(t)}\right)$. Thus, the deviation between Q values of the target network and estimation network can be minimized by gradient descent:

$$\mathrm{Loss}(\theta) = \mathrm{E}\left[\left(\mathrm{r} - \mathrm{Q}\left(S_{(t)}, A_{(t)} \mid \theta\right)\right)^2\right] \tag{12}$$

During the training, only the estimation network is trained, and updated to the target network through multiple online trainings. In this way, the correlation between the two Q values is greatly weakened, avoiding the overestimation of Q value.

To prevent the target strategy from falling into the local optimum trap and search for a better Q value, the ε-greedy strategy was introduced to the DQN. The random selection of the ε-greedy strategy was employed to find the action with the greatest value.

However, a huge amount of data is needed for neural network fitting [10, 13, 16, 22, 24, 26, 38]. Besides, it may take a long time for the action taken under a strategy to produce benefits. Thus, empirical repetition was implemented to store each step of computation into a preset memory pool. Then, the stored data were extracted from the pool for training:

At time t, the algorithm belongs to the state $S_{(t)}$. The corresponding action $A_{(t)}$ is chosen as per the optimal function $\pi^*\left(S_{(t)}\right)$. Then, the instant reward r_t and the next state $S_{(t+1)}$ will be obtained. Next, the data $\left(S_{(t)}, A_{(t)}, r_t, S_{(t+1)}\right)$ will be stored in the memory pool with the capacity of N . After the sample size accumulates to a certain level, the DQN will randomly select data from the pool for iterative computation, and then update the parameters of the estimation network. After certain iterations, the estimation network parameters will be synced to the target network [7, 9, 12, 15, 19, 28]. This approach helps to alleviate the instability and divergence of the results caused by the past states, thereby improving the algorithm stability.

5 Simulation and results analysis

Our simulation was carried out under a UDN with 11 macro base stations and 80 small cell base stations. The area of UDN is 1,000m × 1,000m, and each base station has a radius of 250 m. The users are uniformly distributed within the coverage areas. Each user is allocated at least one channel [18, 23, 31, 37]. If one base station is interrupted, then the service of the user connected to this base station will be interrupted. During the simulation, the power of the small cell base stations was discretized and divided into 5 levels, corresponding to $P = \{10, 15, 20, 25, 30\}$. A neural network with two hidden layers was designed, with 200 hidden layer nodes. Meanwhile, the memory pool capacity N, number of selected samples S, and learning rate α were set to 10,000, 300, and 0.0005, respectively.

Due to the instability of UDN, the energy efficiencies obtained by simulation were all mean values. To verify the DQN performance, Q-learning algorithm was selected as the contrastive algorithm. The simulation results are shown in Figure 2.

In the first 600 iterations, DQN performed poorer than Q-learning, for the DQN requires a preliminary training and applies function fitting to training. At the initial phase of training, the DQN has a much poorer fitting effect than the table look-up method of Q-learning.

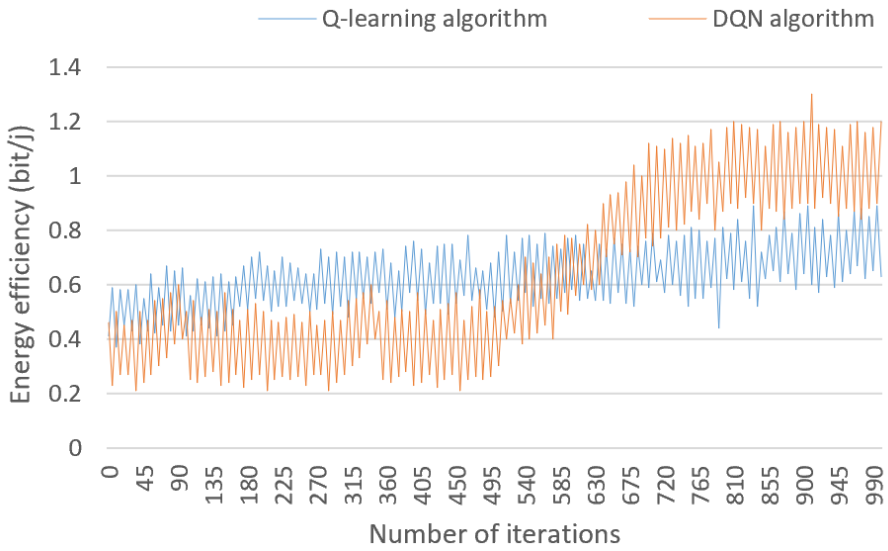


Figure 2: Simulation results

After the 600-th iteration, DQN far outperformed Q-learning, owing to the following reason: Q-learning relies on the table look-up method; In the UDN, however, not all scenarios are stored in the table; thus, Q-learning involves lots of random explorations, which drag down its performance [1, 4, 14, 20, 33, 40].

In addition, the number of base stations has an impact on the overall energy efficiency of the system. Table 3 compares the energy efficiency of base stations at different numbers of base stations. Note that the random power algorithm was introduced as another contrastive algorithm against DQN algorithm.

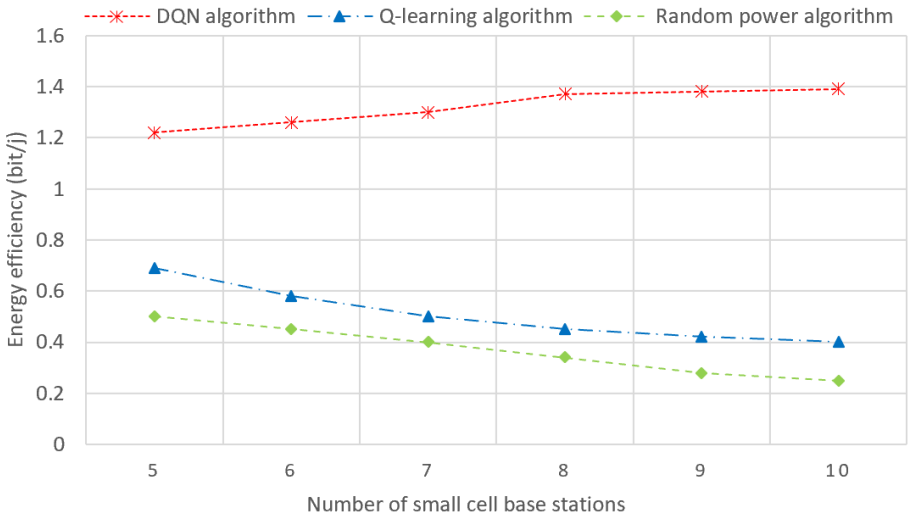


Figure 3: Energy efficiency at different numbers of base stations

As shown in Figure 3, the overall energy efficiency of the system increased with the number of base stations under the DQN algorithm. This is because more small cell base stations mean the algorithm controls more targets, making the scheme more flexible; a flexible scheme pushes up the overall efficiency. By contrast, under Q-learning, the growing number of base stations led to an exponential rise in the number of candidate strategies; then, it is much harder to find the optimal strategy. That is why the Q-learning algorithm had a decline in learning performance. Figure 4 presents the convergence curves of DQN algorithm and Q-learning algorithm.

As shown in Figure 4, our algorithm achieved faster convergence and better energy efficiency than Q-learning.

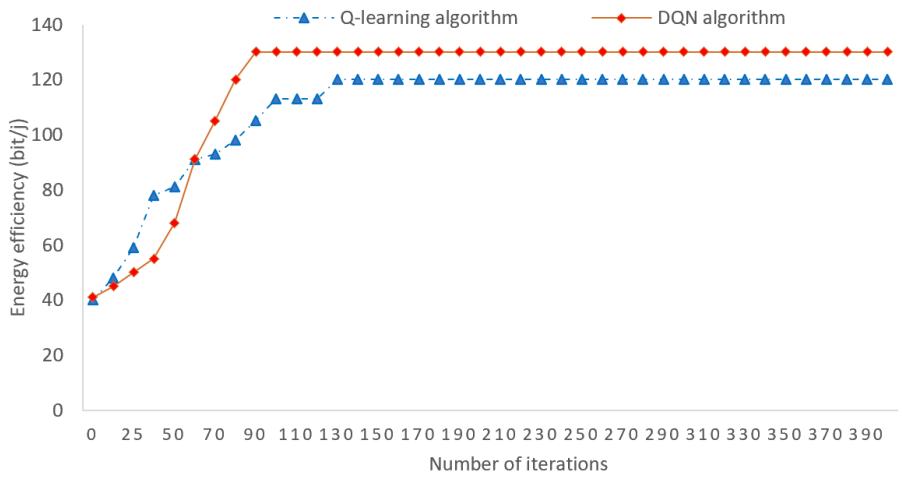


Figure 4: Network convergence

6 Conclusions

This paper proposes a network resource allocation algorithm based on DRL to increase network throughput and energy efficiency of the UDN. Drawing on the theories on deep learning, a long-term reward function was constructed with energy efficiency as the reward. The weight parameters of neural network were updated through repeated training. Simulation results show that, with the growing number of iterations, our algorithm could optimize system energy efficiency more effectively than other algorithms, and provide an effective solution to resource allocation under dynamic complex network environment. The future research will further refine the details and improve the performance of the optimization algorithm, carry out simulations on a larger scale, and apply the algorithm in real-world scenarios.

7 Funding

The APC was funded by R&D center "Cercetare Dezvoltare Agora" of Agora University.

8 Author contributions. Conflict of interest

The authors contributed equally to this work. The authors declare no conflict of interest.

References

[1] Abtahi, F.; Zhu, Z.; Burry, A.M. (2015). A deep reinforcement learning approach to character segmentation of license plate images, *In 2015 14th IAPR international conference on machine vision applications (MVA)*, 539-542, 2015.

[2] Abuzainab, N.; Saad, W.; MacKenzie, A.B. (2019). Distributed uplink power control in an ultra-dense millimeter wave network: A mean-field game approach, *IEEE Wireless Communications Letters*, 8(5), 1328-1332, 2019.

[3] Amiri, R.; Mehrpouyan, H.; Fridman, L.; Mallik, R.K.; Nallanathan, A.; Matolak, D. (2018). A machine learning approach for power allocation in HetNets considering QoS, *In 2018 IEEE International Conference on Communications (ICC)*, 1-7, 2018.

[4] Bai, C.J.; Liu, P.; Zhao, W.; Tang, X.L. (2019). Active sampling method for deep Q learning based on TD-error adaptive correction, *Journal of Computer Research and Development*, 56(2), 38-56, 2019.

[5] Chang, Y.; Fu, F.; Zhang, Z.C. (2020). Research on resource allocation based on reinforcement learning in wireless networks, *Journal of Test and Measurement Technology*, 34(2), 152-158, 2020.

[6] Chen, M.; Hua, Y.; Gu, X.; Nie, S.; Fan, Z. (2016). A self-organizing resource allocation strategy based on Q-learning approach in ultra-dense networks, *In 2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, 155-160, 2016.

[7] Deng, L.; Yu, D. (2014). Deep learning: Methods and applications, *Foundations and Trends In signal Processing*, 7(3-4), 197-387, 2014.

[8] Ge, X.; Tu, S.; Mao, G.; Wang, C.X.; Han, T. (2016). 5G ultra-dense cellular networks, *IEEE Wireless Communications*, 23(1), 72-79, 2016.

[9] Goldsmith, A. (2005). Wireless communications, *Cambridge: Cambridge Univ. Press*, 477-480, 2005.

[10] Hasselt, H.; Guez, A.; Silver, D. (2015). Deep reinforcement learning with double Q-learning, *Computer Science*, 14(8), 367-375, 2015.

[11] He, Y.; Zhang, Z.; Yu, F.R.; Zhao, N.; Yin, H.; Leung, V.C.; Zhang, Y. (2017). Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks, *IEEE Transactions on Vehicular Technology*, 66(11), 10433-10445, 2017.

[12] Hinton, G.E.; Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks, *Science*, 313(5786), 504-507, 2006.

[13] Hui, Q.L. (2020). Multi cell power allocation algorithm based on deep reinforcement learning, *Technology and Market*, 27(10), 11-14, 2020.

[14] Khatib, O. (1986). Real-time avoidance for manipulator and mobile robot, *The International Journal of Robotic Research edition*, 5(1), 90-98, 1986.

[15] Levine, S.; Finn, C.; Darrell, T.; Abbeel, P. (2016). End-to-end training of deep visuomotor policies, *The Journal of Machine Learning Research*, 17(1), 1334-1373, 2016.

[16] Liao, X.M.; Yan, S.H.; Shi, J.; Tan, Z.Y.; Zhao, Z.L.; Li, Z. (2019). Deep reinforcement learning based resource allocation algorithm in cellular networks, *Journal on Communications*, 40(2), 15-22, 2019.

[17] Liu, H.Y. (2016). Research on distributed radio resource management in 5G oriented ultra dense networks, *Beijing Jiaotong University*, 2016.

[18] Lozano-Pérez, T.; Wesley, M.A. (1979). An algorithm for planning collision-free paths among polyhedral obstacles, *Communications of the ACM*, 22(10), 560-570, 1979.

[19] Maddumala, V.R., Arunkumar, R. (2020). Big data-driven feature extraction and clustering based on statistical methods, *Traitement du Signal*, 37(3), 387-394, 2020.

[20] Mitchell, M.; Holland, J.H. (1993). When will a genetic algorithm outperform hill-climbing? *Advances in Neural Information Process System*, 9(4), 120-136, 1993.

[21] Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. (2015). Human-level control through deep reinforcement learning, *Nature*, 518(7540), 529-533, 2015.

[22] Nie, J.; Haykin, S. (1999). A Q-learning-based dynamic channel assignment technique for mobile communication systems, *IEEE Transactions on Vehicular Technology*, 48(5), 1676-1687, 1999.

[23] Saad, H.; Mohamed, A.; ElBatt, T. (2014). A cooperative Q-learning approach for distributed resource allocation in multi-user femtocell networks, In 2014 *IEEE Wireless Communications and Networking Conference (WCNC)*, 1490-1495, 2014.

[24] Sutton, R.S.; Barto, A.G. (1998). Reinforcement learning: An introduction, *Cambridge: MIT Press*, 47-68, 1998.

[25] Tan, C.W.; Palomar, D.P.; Chiang, M. (2005). Solving nonconvex power control problems in wireless networks: Low SIR regime and distributed algorithms, *In GLOBECOM'05. IEEE Global Telecommunications Conference*, 2005.

[26] Tang, L.; Wei Y.N.; Ma R.L.; He, X.Y.; Chen, Q.B. (2019). Online learning-based virtual resource allocation for network slicing in virtualized cloud radio access network, *Journal of Electronics & Information Technology*, 41(7), 1533–1539, 2019.

[27] Teng, Y.; Liu, M.; Yu, F.R.; Leung, V.C.; Song, M.; Zhang, Y. (2018). Resource allocation for ultra-dense networks: A survey, some research issues and challenges, *IEEE Communications Surveys & Tutorials*, 21(3), 2134-2168, 2018.

[28] Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; Freitas, N. (2016). Dueling network architectures for deep reinforcement learning, *In International Conference on Machine Learning*, 1995-2003, 2016.

[29] Wang, X.; Liu, B.; Su, X. (2018). A power allocation scheme using non-cooperative game theory in ultra-dense networks, *In 2018 27th Wireless and Optical Communication Conference (WOCC)*, 1-5, 2018.

[30] Wang, X.; Liu, B.; Su, X. (2018). A power allocation scheme using non-cooperative game theory in ultra-dense networks, *In 2018 27th Wireless and Optical Communication Conference (WOCC)*, 1-5, 2018.

[31] Wang, H.D. (2020). A synchronous transmission method for array signals of sensor network under resonance technology, *Traitement du Signal*, 37(4), 579-584, 2020.

[32] Xiong, K. (2019). Research on resource allocation of wireless virtual network based on deep reinforcement learning, *University of Electronic Science and Technology*, 2019.

[33] Yoon, J.; Arslan, M. Y.; Sundaresan, K.; Krishnamurthy, S.V.; Banerjee, S. (2018). Characterization of interference in OFDMA small-cell networks, *IEEE Transactions on Vehicular Technology*, 67(9), 7937-7954, 2018.

[34] Zhang, G.; Zhang, H. (2008). Adapative resource allocation for downlink OFDMA networks using cooperative game theory, *In 2008 11th IEEE Singapore International Conference on Communication Systems*, 98-103, 2008.

[35] Zhang, G.; Yang, K.; Chen, H.H. (2012). Resource allocation for wireless cooperative networks: A unified cooperative bargaining game theoretic framework, *IEEE Wireless Communications*, 19(2), 38-43, 2012.

[36] Zhang, H.; Jiang, C.; Beaulieu, N.C.; Chu, X.; Wang, X.; Quek, T.Q. (2015). Resource allocation for cognitive small cell networks: A cooperative bargaining game theoretic approach, *IEEE Transactions on Wireless Communications*, 14(6), 3481-3493, 2015.

[37] Zhang, H.; Liu, H.; Cheng, J.; Leung, V.C. (2017). Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks, *IEEE Transactions on Communications*, 66(4), 1705-1716, 2017.

[38] Zhao, W.C.; Wu, J.Q. (2018). Study on the game control based on prior experience replay algorithm, *Journal of Gansu Sciences*, 30(2), 15-19, 2018.

[39] Zia, K.; Javed, N.; Sial, M.N.; Ahmed, S.; Pirzada, A.A.; Pervez, F. (2019). A distributed multi-agent RL-based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets, *IEEE Access*, 7, 6733-6745, 2019.

[40] Zou, Y.; Xing, Q.Z.; Wang, B.C.; Zheng, S.X.; Cheng, C.; Wang, Z.M.; Wang, X.W. (2019). Application of the asynchronous advantage actor–critic machine learning algorithm to real-time accelerator tuning, *Nuclear Science and Techniques*, 30(10), 1-9, 2019.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Zhang, H. S.; Wang, T. M.; Shen, H. W. (2021). A Resource Allocation Algorithm for Ultra-Dense Networks Based on Deep Reinforcement Learning, *International Journal of Computers Communications & Control*, 16(2), 4189, 2021.
https://doi.org/10.15837/ijccc.2021.2.4189