Article:

Xu, Y.-H., Tian, Y.-B., Searyoh, P.K. et al. (2 more authors) (2020) Deep reinforcement learning-based resource allocation strategy for energy harvesting-powered cognitive machine-to-machine networks. Computer Communications, 160. pp. 706-717. ISSN 0140-3664

https://doi.org/10.1016/j.comcom.2020.07.015

Titled: Deep Reinforcement Learning-based Resource Allocation Strategy for Energy Harvesting-Powered Cognitive Machine-to-Machine Networks

Author: Yi-Han Xu[1,2], Yong-Bo Tian[1], Prosper Komla Searyoh[1], Gang Yu[3] and Yueh-Tiam Yong[4]

Affiliations:
[1]College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney 2052, Australia
[3]Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, UK
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Samarahan Campus, Kota Samarahan 94300, Malaysia

Abstract: Machine-to-Machine (M2M) communication is a promising technology that may realize the Internet of Things (IoTs) in future networks. However, due to the features of massive devices and concurrent access requirement, it will cause performance degradation and enormous energy consumption. Energy Harvesting-Powered Cognitive M2M Networks (EH-CMNs) as an attractive solution is capable of alleviating the escalating spectrum deficient to guarantee the Quality of Service (QoS) meanwhile decreasing the energy consumption to achieve Green Communication (GC) became an important research topic. In this paper, we investigate the resource allocation problem for EH-CMNs underlaying cellular uplinks. We aim to maximize the energy efficiency of EH-CMNs with consideration of the QoS of Human-to-Human (H2H) networks and the available energy in EH-devices. In view of the characteristic of EH-CMNs, we formulate the problem to be a decentralized Discrete-time and Finite-state Markov Decision Process (DFMDP), in which each device acts as agent and effectively learns from the environment to make allocation decision without the complete and global network information. Owing to the complexity of the problem, we propose a Deep Reinforcement Learning (DRL)-based algorithm to solve the problem. Numerical results validate that the proposed scheme outperforms other schemes in terms of average energy efficiency with an acceptable convergence speed.

Keywords: Energy harvesting, M2M communication, Resource allocation, Deep reinforcement learning

I. Introduction

Machine-to-Machine (M2M) communication as a promising technology to realize Internet of Things (IoTs) has attracted great attention from both industry and academia. Different with conventional Human-to-Human (H2H) communication, M2M communication is expected to provide ubiquitous connectivity among various heterogeneous devices by means of autonomous communication and networking technologies without human intervention [1-2]. However, such type of communication further poses challenges to the issues of spectrum scarcity and high energy consumption due to it normally involves massive and concurrent access requirement. Although Third Generation Partnership Project (3GPP) continues to promote prospective communication technologies to alleviate the escalating spectrum deficient and decrease the energy consumption, the resource allocation strategy for a large number of devices which provide various types of service in heterogeneity has not been well investigated. At the meantime, several pioneering efforts and researches relevant to resource allocation problem in M2M communication have been conducted in [3-7], but these works are mainly focusing on the network performance such as packet loss ratio, delay and throughput. Seldom works take into account of energy consumption and the influence on the H2H communication [8]. Therefore, conceiving an energy-efficient and interference-manageable resource allocation strategy for M2M communication is essential.

Cognitive M2M communication is a novel technology that integrates cognitive radio into M2M communication to enable devices learn from the environment and utilize the unoccupied licensed spectrum to improve the spectrum efficiency meanwhile avoiding the interference to primary human users. Along with spectrum efficiency, another major concern in M2M communication is the energy efficiency issue. M2M communication as a key enabler of realizing IoTs has involved a massive number of sensor-likewise devices. These devices have the inherent nature of limited energy supplies and the difficulty of batteries recharging. In addition to further improving energy efficiency, Energy Harvesting (EH) is an appealing solution. EH is a technology that enables devices to collect energy from ambient sources [9]. Various types of energy sources can be exploited as energy supplies, for instance, solar, thermal, wind and electromagnetic wave [10], [11]. However, owing to the fluctuation of ambient energy and the immaturity of energy conversion technology, the available energy of each device will become a vital factor in the designing of resource allocation strategy in Energy Harvesting-Powered Cognitive M2M Networks (EH-CMNs).

To response this, we propose an energy efficient resource allocation strategy for EH-CMNs in this paper. The goal of the strategy is to maximize the average energy efficiency of devices in EH-CMNs by jointly consider the transmission power control, time slot allocation, transmission mode and relay selection with the constraints of conventional H2H communication and the energy status of EH-devices. We formulate the problem as a decentralized Discrete-time and Finite-state Markov Decision Process (DFMDP), in which each device acts as agent and effectively learns from the environment to make allocation decision without having complete and global network information. Owing to the complexity of the problem, we also propose a Deep Reinforcement Learning (DRL) algorithm to solve the problem and find the optimal allocation strategy in the formulated model. Numerical results validate that the proposed scheme outperforms other schemes in terms of average energy efficiency. Meanwhile, the proposed DRL algorithm can obtain higher convergence speed as compared to the classical Q-Learning algorithm.

The remainder of this paper is organized as follows. Section II gives a detailed literature survey on the most relevant existing works. After that, our network model is presented in Section III. Section IV provides a high-level description of the corresponding energy efficiency maximization problem and the proposed DRL algorithm. In Section V, the simulation setting and results are discussed. Finally, we give the conclusions in Section VI.

II. Related Work

Conventionally, resource allocation strategy plays a significant role in improving spectrum efficiency and energy efficiency. However, due to the different features between H2H and M2M communications, the resource allocation schemes designed for H2H networks (either IEEE-based networks or 3GPP/3GPP2-based networks) cannot be directly applied to M2M communication. In this section, we review a number of previous research activities related to the issues and the enabling technologies. When a massive number of devices attempt to access a spectrum simultaneously will result collisions. The collided devices will wait for a random time period before next attempt to access. 3GPP in [12] investigated the radio access network improvements for devices in M2M communication underlaying LTE and several potential efforts are proposed to address the overload problem in Physical Random Access Channel (PRACH). In [13], a group-based M2M access scheme is proposed to enhance the efficiency in random access network by using multiple connections among different devices in the same group. Simulation results shown that this scheme enables to improve the random access performance in the condition of the workload is high. Similarly, another group-based random access scheme for cellular M2M communications is proposed in [14] to reduce collisions during the random access procedure. The core idea of this scheme is to make use of multiple beams to divide M2M devices into different groups and utilizing the spatial selectivity of beams to limit the interference among different groups. In [15], an information-centric networking for M2M communications is investigated from design and deployment perspectives. The goal of this scheme is to ensure the easy interoperability with the European Telecommunications Standards Institute (ETSI) M2M specifications. Therefore, a test-bed is also developed to showcase the viability of this scheme. Experimental results shown that the device resources consumption has been improved. Moreover, as a key technology to overcome the spectrum efficiency problem, cognitive M2M communication has attracted interests from researchers worldwide. A comprehensive survey on the major characteristics, research issues, and challenges in cognitive M2M communication from a practical design and implementation perspective is provided in the works of [16] and [17]. In addition, authors of [18] studied the value of cognitive M2M to traditional cellular networks from the prospective of economic. However, these above-mentioned works mainly concentrate on the enhancement of spectrum efficiency, and the energy efficiency issue is ignored. Generally, M2M communications have the characteristics of limited power supply and a massive number of machine-type communication devices deployed in heterogeneity scenarios, therefore the shortcoming of energy efficiency should be highly considered. According to the investigation in [19], the network throughput of M2M communication is mainly limited by the energy budget in each device. Furthermore, some researchers intended to investigate the integration of energy harvesting and M2M communications. An EH-assisted and social-aware transmission protocol for M2M communication is proposed in [20]. The authors of [21] proposed three different spectrum access schemes for EH-M2M communication with the goal of improving the performance in terms of throughput, delay and energy efficiency. However, this work did not consider the co-channel interference caused by spectrum sharing. In [22], a joint power control and time allocation scheme is proposed to minimize the energy consumption for M2M communication. The authors formulated the problem to two strategies: Non-Orthogonal Multiple Access (NOMA) and Time Division Multiple Access (TDMA). However, this work did not take into account of transmission mode and relay selection. In [23], a joint channel selection, peer discovery, power control and time allocation scheme is proposed to maximize the energy efficiency of the transmitter in M2M communication. However, the high computation complexity against the original intention of saving energy in this work. Furthermore, the convergence speed of the proposed algorithm is not evaluated as well.

III. Network Model Descriptions

In this section, we first depict the network model of the proposed EH-CMNs, which is then followed by the details on data transmission model, energy harvesting model and energy efficiency model in EH-CMNs.

A. Network model

In this treatise, we consider a scenario of EH-CMNs underlaying a single cellular network, as illustrated in Figure 1. Base Station (BS) is located at the center of the cell with radius R, while N Cellular Users (CUs) are denoted as c_i $(i \in \{1, 2, \dots, N\})$ and M machine-type communication devices are denoted as d_j $(j \in \{1, 2, \dots, M\})$ are uniformly distributed in the coverage area. Each M2M pair has a transmitter (DU_Tx) and a receiver (DU_Rx). For simplicity, we only consider machine-type communication devices are equipped with EH function, and CUs are still supported by traditional battery power. Moreover, in order to improve the network resource efficiency, we assume that both direct transmission and cooperative transmission modes are supported by the devices. For simplicity, we suppose that only two-hop transmission is supported by the cooperative transmission mode in this model. There are three main reasons for making this assumption: 1) as the number of transmission hops increase, the network throughput will be increased. However, it will lead to the network resource allocation problem becomes more complex. Although the proposed DRL algorithm in this paper enables to find the optimal allocation strategy in such case, it may involve extra computational latency, which is a tradeoff issue between energy efficiency and latency; 2) each device is powered by the harvested energy, the status of the available energy of each device is changing dynamically, if a transmission link includes more hops, it may increase the probability of transmission instability; 3) unlike conventional multi-hops wireless sensor networks, the proposed M2M network is underlaying the cellular network, if a specific device cannot access the network via 2 hops or accessing the network at the cost of more energy consumption, this device can be treated as a CU and it will be assigned with a fixed cellular spectrum. The relay device is denoted as DU_Rly. We define a binary parameter $\alpha_{d_j} \in \{0, 1\}$ $j \in (1, 2, \dots, M)$ to indicate that which transmission mode that is utilized recently by the j-th device. $\alpha_{d_j} = 1$ denotes that the j-th device is in direct transmission mode, while $\alpha_{d_j} = 0$ indicates that the j-th device is in cooperative transmission mode. In MAC layer, TDMA-based access mechanism is employed, in which each transmission frame can be divided into multiple time slots. These time slots can be assigned to the devices, whether they operate in direct transmission or cooperative transmission modes. Meanwhile, it should be noted that the transmission mode of each device is determined by the resource allocation strategy in the proposed scheme. For example, after the proposed scheme finds the optimal resource allocation strategy, in which a specific device is determined to transmit data in direct mode, then the device will use the certain time slot (which is also determined by the resource allocation strategy) to transmit data. We suppose that each transmission frame includes K number of time slots and the time slot set is denoted as $\psi = (1, 2, ..., K)$. We set $t_0 = 0$ and $t_K = T$. The duration of each slot is denoted as $\tau_k = t_k - t_{k-1}$ $\forall k \in \psi$.

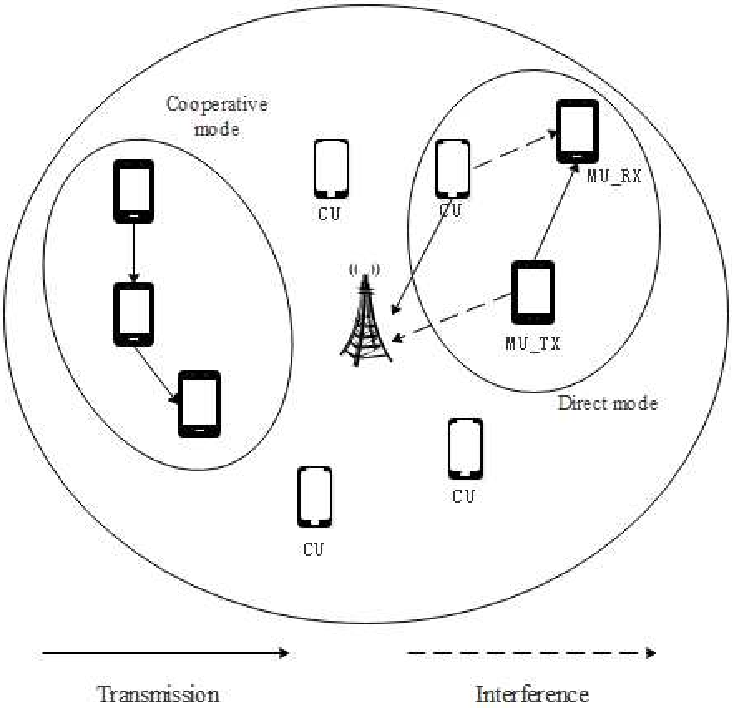

Figure 1 Network model

In case of direct transmission, we define a binary parameter $\beta_{d_j}^k \in \{0, 1\}$, ($j \in (1, 2, ..., M), \forall k \in \psi$) to indicate which time slot is assigned to a specific device. $\beta_{d_j}^k = 1$ denotes that the k-th time slot is assigned to the j-th device for direct transmission, while $\beta_{d_j}^{k} = 0$ means the *k*-th time slot is not assigned to the *j*-th device for direct transmission. More specifically, another two reasonable assumptions are made in this model: 1) each device can only receive data from one device at each time slot; 2) in each time frame, each device only be assigned at most one time slot for transmission. The purpose of these two assumptions is to maintain the fairness of transmission opportunity of each device. Thus, we can derive two constraints as Equations 1 and 2:

$$\sum_{j=1}^{N_M} \beta_{d_j}^{k} \leq 1, k \in \ \psi \tag{1}$$

$$\sum_{k=1}^{K} \beta_{d_j}^{k} \leq 1, j \in (1, 2, ..., M) \tag{2}$$

In case of cooperative transmission, we assume that the *K* time slots in a transmission frame are allocated to both *DU_Tx*-*DU_Rly* and *DU_Rly*-*DU_Rx* links. This assumption is mainly to be used to guarantee the fairness between direct transmission and cooperative transmission, to obtain the optimal resource allocation strategy. Similarly, we define a parameter $\delta_{d_j \to d_r}^{\ k} \in \{0, 1\}, (j,r \in (1, 2, ..., M), \forall k \in \psi$) as an indicator that the *k*-th time slot is allocated to *j*-th device for transmitting data to the *r*-th device, which is selected as the relay of the *j*-th device. Meanwhile, $\delta_{d_j \to d_r \to d_z}^{\ k} \in \{0, 1\}(j,r,z \in (1, 2, ..., M), \forall k \in \psi$) is denoted as the indicator that the *r*-th device forwards the data from the *j*-th device to the *z*-th device at the *k*-th time slot. In this model, we suppose that each *DU_Tx* only can select one *DU_Rly* during any time slot in a transmission frame and each *DU_Rly* can only forward data from one *DU_Tx* during any time slot in a transmission frame. Thus we can obtain two constraints as Equations 3 and 4:

$$\sum_{r=1, r \neq j}^{N_M} \delta_{d_j \to d_r}^{\ k} \leq 1, \qquad \sum_{j=1, j \neq r}^{N_M} \delta_{d_j \to d_r}^{\ k} \leq 1 \tag{3}$$

$$\sum_{j=1, j \neq r}^{N_M} \delta_{d_j \to d_r \to d_z}^{\ k} \leq 1, \qquad \sum_{r=1, r \neq j}^{N_M} \delta_{d_j \to d_r \to d_z}^{\ k} \leq 1 \tag{4}$$

Furthermore, due to each link can only be assigned at most one time slot, we can obtain constraint as Equation 5:

$$\sum_{k=1}^{K} \delta_{d_j \to d_r}^{k} \le 1, \sum_{k=1}^{K} \delta_{d_j \to d_r \to d_z}^{k} \le 1 \quad j \neq r \qquad (5)$$

Another aspect to note is that the data transmission from *DU_Tx* to *DU_Rly* should be prior to the transmission from *DU_Rly* to *DU_Rx*. Therefore, we can obtain Equation 6:

$$\sum_{k=1}^{x} \delta_{d_j \to d_r}^{k} - \sum_{k=x+1}^{K} \delta_{d_j \to d_r \to d_z}^{k} \ge 0, x \in (1, 2, \ldots, K-1) \qquad (6)$$

B. Data transmission model

In this model, each CU in cellular network is pre-assigned uplink spectrum resource with the bandwidth of *B*, which is orthogonal mutually. Reasonably, we suppose that each cognitive M2M pair can multiplex the uplink spectrum that assigned to CUs as the secondary user temporally. We can derive the instantaneous Signal to Interference plus Noise Ratio (SINR) of *i*-th CU as Equation 7.

$$SINR_{c_i, \ k} = \frac{p_{i,k} \cdot g_{c_i - BS}^{k}}{\sum_{d_j \in M} p_{j,k} \cdot g_{d_j - BS}^{k} + n_0} \qquad (7)$$

According to Shannon's theorem, we can get the instantaneous transmission rate of *i*-th CU as Equation 8.

$$R_{c_i, k} = B \cdot log_2\left(1 + SINR_{c_i, \ k}\right) \qquad (8)$$

Furthermore, we can get the long term average transmission rates of CUs as Equation 9:

$$R_C = \lim_{K \to \infty} sup \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N_C} \mathbb{E}[R_{c_i, k}] \qquad (9)$$

Where, $p_{i,k}$ and $p_{j,k}$ are the instantaneous transmission powers of the *i*-th CU and *j*-th M2M device in *k*-th time slot, respectively. g^k denotes the channel gain among i, j and BS ($i \in N, j \in M$), and n_0 is the noise power, which equals to $B \cdot \rho_n$, where ρ_n is the density of noise. Moreover, in order to guarantee the transmission rate of the primary CUs, the value of R_c should attain the minimum transmission rate threshold TR_{th}.

In M2M communication, various devices with different functions may have different transmission rate requirements. In this network model, we denote R_j as the transmission rate of the *j*-th device and it can be expressed as Equation 10:

$$R_j = \alpha_{d_j} \cdot R_j^{\,d} + \left(1 - \alpha_{d_j}\right) \cdot R_j^{\,c}, j \in \left(1, 2, \ldots, N_M\right) \qquad (10)$$

Where, $R_j^{\,d}$ is the transmission rate of the *j*-th device in direct transmission mode, and $R_j^{\,c}$ denotes the transmission rate of the *j*-th device when transmitting data to destination via relay device. Based on the above analysis, we can derive the instantaneous SINR of direct transmission and cooperative transmission. Equations 11, 12 and 13 give the instantaneous SINR of direct link, *DU_Tx-DU_Rly* link and *DU_Rly-DU_Rx* link in the *k*-th time slot, respectively.

$$SINR_{j,k}^{d} = \frac{p_{j,k}^{d} \cdot g_{d_j - d_z}}{\sum_{j_1 = 1, j_1 \neq j}^{N_M} \sum_{r = 1, r \neq j, j_1}^{N_M} \delta_{d_{j_1} \to d_r}^{\;k} \cdot p_{j_1, r, k}^{s} \cdot g_{d_{j_1} - d_r} + n_0} \qquad (11)$$

Where, $p_{j,k}^{d}$ denotes the instantaneous transmission power of the *j*-th device in the *k*-th time slot when transmitting data to the *z*-th device which is the destination device of the *j*-th device, $g_{d_j - d_z}$ is the channel gain between the *j*-th device and the *z*-th device, $p_{j_1, r, k}^{s}$ denotes the instantaneous transmission power of j_1-th device in the *k*-th time slot when transmitting data to *r*-th device, which is selected as the relay of the j_1-th device, $g_{d_{j_1} - d_r}$ is the channel gain between the j_1-th device and *r*-th device.

$$SINR_{j,r,k}^{s \to r} = \frac{p_{j,r,k}^{s \to r} \cdot g_{d_j - d_r}}{I_{j,r,k}^{s \to r} + n_0} \qquad (12)$$

$$I_{j,r,k}^{s \to r} = \sum_{\substack{j_1 = 1 \\ j_1 \neq j, r}}^{N_M} \sum_{\substack{r_1 = 1 \\ r_1 \neq j, j_1, r}}^{N_M} \delta_{d_{j_1} \to d_{r_1}}^{\;k} \cdot p_{j_1, r_1, k}^{s \to r} \cdot g_{d_{j_1} - d_r}$$

$$+\sum_{\substack{j_1=1\\ j_1\neq j,r}}^{N_M}\beta_{d_{j_1}}^{k}\cdot p_{j_1,k}^{d}\cdot g_{d_{j_1}-d_r}+\sum_{\substack{j_1=1\\ j_1\neq j,r}}^{N_M}\sum_{\substack{r_1=1\\ r_1\neq j,j_1,r}}^{N_M}\delta_{d_{j_1}\to d_{r_1}\to d_z}^{k}\cdot p_{j_1,r_1,k}^{r\to d}\cdot g_{d_{r_1}-d_r}$$

Where, $p_{j,r,k}^{s\to r}$ denotes the instantaneous transmission power of the *j*-th device when transmitting data to the *r*-th device, which is selected as its relay in the *k*-th time slot. $p_{j,r,k}^{r\to d}$ denotes the instantaneous transmission power of *r*-th device in the *k*-th time slot when forwarding data from *j*-th device to the destination. $I_{j,r,k}^{s\to r}$ is the total instantaneous interference of the *DU_Tx-DU_Rly* link in the *k*-th time slot. The expression of $I_{n,m,k}^{s\to r}$ includes three items, the first item indicates the interference from other *DU_Tx-DU_Rly* links, the second item is the interference from direct transmission between *DU_Tx* and *DU_Rx*, and the three item represents the interference from *DU_Rly-DU_Rx* links.

$$SINR_{j,r,k}^{r\to d}=\frac{p_{j,r,k}^{r\to d}\cdot g_{d_r-d_z}}{I_{j,r,k}^{r\to d}+n_0} \tag{13}$$

$$I_{j,r,k}^{r\to d}=\sum_{\substack{j_1=1\\ j_1\neq r}}^{N_M}\sum_{\substack{r_1=1\\ r_1\neq r,j_1}}^{N_M}\delta_{d_{j_1}\to d_{r_1}}^{k}\cdot p_{j_1,r_1,k}^{s\to r}\cdot g_{d_{j_1}-d_r}$$

Where, $I_{j,r,k}^{r\to d}$ is the total instantaneous interference between the relay device and destination device when *r*-th device is selected as the relay of *j*-th device in the *k*-th time slot.

According to Shannon's theorem, we can obtain the transmission rate of direct link as given in Equation 14:

$$R_j{}^{d}=\sum_{k=1}^{K}\beta_{d_j}^{k}\cdot B\cdot log_2(1+SINR_{j,k}^{d}) \tag{14}$$

The transmission rate of the cooperative mode $R_j{}^{c}$ can be divided into two parts: one is the transmission rate of *DU_Tx-DU_Rly* link $R_j{}^{c,\quad s\to r}$and another is the transmission rate of *DU_Rly-DU_Rx* link $R_j{}^{c,\quad r\to d}$, as shown in Equations 15 and 16:

$$R_j^{c,\quad s\to r} = \sum_{\substack{r=1\\ r\neq j}}^{N_M} \sum_{k=1}^{K} \delta_{d_j\to d_r}^{\;k} \cdot B \cdot log_2(1+SINR_{j,r,k}^{s\to r}) \qquad (15)$$

$$R_j^{c,\quad r\to d} = \sum_{\substack{r=1\\ r\neq j}}^{N_M} \sum_{k=1}^{K} \delta_{d_j\to d_r\to d_z}^{\;k} \cdot B \cdot log_2(1+SINR_{j,r,k}^{r\to d}) \qquad (16)$$

However, in cooperative transmission mode, the transmission rate of the path between *DU_Tx* and *DU_Rx* is limited by the smaller transmission rate of *DU_Tx-DU_Rly* link and *DU_Rly-DU_Rx* link. Hence, the transmission rate of the cooperative mode is $R_j^c = \min\ \ (R_j^{c,\quad s\to r},\ R_j^{c,\quad r\to d})$.

C. Data serving model

In this scenario, we make the assumption that the data are stored in the form of packets in the buffer of the device. The arrived data at each device follows an independently and identically distributed (i.i.d.) sequence with average rate of λ_d [24]. Practically, we assume that the buffer of device is finite and served in first in first out fashion. We denoted $DQ_{d_j}^k$ as the instantaneous data queue length at the *j*-th device in time slot *k*. The maximum traffic queue length of devices is represented by $DQ_{d_j}^{max}$. Accordingly, we can obtain the update function of the instantaneous data queue length as Equation 17:

$$DQ_{d_j}^k = \min\left\{ DQ_{d_j}^{max}, DQ^{k-1}_{d_j} - min\left\{\left\lfloor \frac{\alpha_{d_j}\cdot R_j^{\,d} + (1-\alpha_{d_j})\cdot R_j^{\,c}}{PS_{data}}\cdot \tau_k \right\rfloor, DQ^{k-1}_{d_j}\right\} + A^{k-1}_{d_j}\right\} \qquad (17)$$

Where, PS_{data} is the traffic packet size with the unit of bits/packet, $\frac{\alpha_{d_j}\cdot R_j^{\,d} + (1-\alpha_{d_j})\cdot R_j^{\,c}}{PS_{data}}\cdot \tau_k$ is the number of instantaneous served packets of transmission link of *j*-th device in time slot *k-1 and* $A^{k-1}_{d_j}$ is the arriving traffic packets of the *j*-th device in time slot *k-1*.

D. Energy harvesting model

We denoted $\mathrm{E}_{j,\,\mathrm{k}}$ as the energy harvested by the *j*-th device in the *k*-th time slot. $\{\mathrm{E}_{j,\,1}, \mathrm{E}_{j,\,2}, ..., \mathrm{E}_{j,\,\mathrm{t}}, ..., \mathrm{E}_{j,\,\mathrm{K}}\}$ is the time sequence of energy harvested in a transmission frame. It is also i.i.d. sequence with average rate of λ_{e}. We denote $EQ_{d_j}^{k}$ as the instantaneous energy queue length at the *j*-th device in the *k*-th time slot. The maximum energy queue length of devices is represented by $EQ_{d_j}^{max}$. Therefore, we can obtain the update function of the instantaneous energy queue length as Equation 18:

$$EQ_{d_j}^{k} = \min\left\{ EQ_{d_j}^{max}, EQ^{k}{}_{d_j}^{-1} - min\left\{\left\lceil\frac{p_{j,k-1}}{PS_{energy}}\cdot\tau_k\right\rceil, EQ^{k}{}_{d_j}^{-1}\right\} + \mathrm{E}_{j,\,k-1}\right\} \quad (18)$$

Where, PS_{energy} is the energy packet size with the unit of Joules/packet. $p_{j,k-1}$ denotes the transmission power of the device in the *k*-1-th time slot. According to the transmission mode, $p_{j,k-1}$ can be set to one of $p_{j,k-1}^{d}$, $p_{j,r,k-1}^{s\to r}$ and $p_{j,r,k-1}^{r\to d}$.

It is worth noting that, because the capacity of the energy storage device is finite, two constraints can be derived from Equation 18 as expressed in Equations 19 and 20:

$$\sum_{k=1}^{K}\left\lceil\frac{p_{j,k-1}}{PS_{energy}}\cdot\tau_k\right\rceil \le \sum_{k=1}^{K} EQ_{d_j}^{k}, \forall K \in \{1,2,...\} \quad (19)$$

$$\sum_{k=1}^{K} EQ_{d_j}^{k} - \sum_{k=1}^{K}\left\lceil\frac{p_{j,k-1}}{PS_{energy}}\cdot\tau_k\right\rceil \le EQ_{d_j}^{max}, \forall K \in \{1,2,...\} \quad (20)$$

Equation 19 depicts that the current available energy cannot exceed the total energy in the battery. Equation 20 expresses that the total energy stored in the battery cannot exceed the maximum battery capacity.

E. Energy efficiency model

In this paper, we define the energy efficiency of EH-CMNs as the ratio of the transmission rate to the consumed transmission power. Equation 21 gives the energy efficiency of the *j*-th device in time slot *k*.

$$EE_{d_j}^{k} = \frac{\alpha_{d_j} \cdot {R_j}^{d} + (1 - \alpha_{d_j}) \cdot R_j^{c}}{p_{j,k}} \quad \forall j \in (1, 2, ..., M), \forall k \in \psi \tag{21}$$

Therefore, the average energy efficiency of the overall EH-CMNs is presented as follows:

$$EE = \frac{1}{M} \cdot \sum_{k=1}^{K} \sum_{j=1}^{M} \quad EE_{d_j}^{k} \tag{22}$$

The corresponding EE maximization problem can be formulated as follows:

$$\underset{\alpha_{d_j}, \beta_{d_j}^{k}, \delta_{d_j}^{k}, p_{j,k},}{maximize} \quad EE \tag{23}$$

s.t.

$$\sum_{j=1}^{M} \beta_{d_j}^{k} \leq 1, k \in \ \psi, \qquad \sum_{k=1}^{K} \beta_{d_j}^{k} \leq 1, j \in (1, 2, ..., M)$$

$$\sum_{r=1, r \neq j}^{M} \delta_{d_j \to d_r}^{k} \leq 1, \qquad \sum_{j=1, j \neq r}^{M} \delta_{d_j \to d_r}^{k} \leq 1$$

$$\sum_{j=1, j \neq r}^{M} \delta_{d_j \to d_r \to d_z}^{k} \leq 1, \qquad \sum_{r=1, r \neq j}^{M} \delta_{d_j \to d_r \to d_z}^{k} \leq 1$$

$$\sum_{k=1}^{K} \delta_{d_j \to d_r}^{k} \leq 1, \sum_{k=1}^{K} \delta_{d_r \to d_z}^{k} \leq 1 \quad j \neq r$$

$$\sum_{k=1}^{x} \delta_{d_j \to d_r}^{k} - \sum_{k=x+1}^{K} \delta_{d_j \to d_r \to d_z}^{k} \geq 0, x \in (1, 2, ..., K-1)$$

$$\lim_{K \to \infty} sup \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{E}[R_{c_i, k}] \geq TR_{th}$$

$$\sum_{k=1}^{K} \left\lceil \frac{p_{j,k-1}}{PS_{energy}} \cdot \tau_k \right\rceil \leq \sum_{k=1}^{K} EQ_{d_j}^{k}, \forall K \in \{1,2,...\}$$

$$\sum_{k=1}^{K} EQ_{d_j}^{k} - \sum_{k=1}^{K} \left\lceil \frac{p_{n,k-1}}{PS_{energy}} \cdot \tau_k \right\rceil \leq \ EQ_{d_j}^{max}, \forall K \in \{1,2,...\}$$

$$p_{j,k}^{d} \leq p_{j}^{max} \ \forall j \in (1, 2, ..., M), \forall k \in \psi$$

$$p_{j,r,k}^{s \to r} \leq p_{j}^{max} \ \ j,r \in (1, 2, ..., M), j \neq r, \forall k \in \psi$$

$$p_{j,r,k}^{r \to d} \leq p_{j}^{max} \ \ j,r \in (1, 2, ..., M), j \neq r, \forall k \in \psi$$

IV. Problem Formulation and Optimization Algorithm

From the energy efficiency maximization problem, we can see that it is a multi-objectives optimization problem. Simultaneously, because the variables $p_{n,k}$ are continuous, while $\alpha_{S_n}, \beta_{S_n}^{k}, \delta_{S_n}^{k}$ are binary, the problem (23) is a mixed integer nonlinear programming problem, which cannot be directly solved by convex optimization methods. Even if we can transform the original problem into a tractable convex optimization problem, the problem still requires the prior network information. Furthermore, from Equations 17 and 18, we found that both the traffic packet and the energy packet are only related to current arrivals and the previous remainders. Thus, we can formulate problem (23) as the DFMDP [25]. More specifically, our scenario can be formulated to either centralized or decentralized DFMDP. However, in the centralized DFMDP, BS should acquire all information about the network to make the optimal decision. In this situation, BS will face a large-scale state-action exploratory space that may result network signaling overhead and redundancy. Therefore, we formulated the problem to be a decentralized DFMDP. Meanwhile, due to the reason that the massive number of devices will be deployed in M2M network in future IoTs, the RL-based algorithm such as classical Q-learning algorithm cannot satisfy the requirement of delay-sensitive applications. Therefore, in this paper, we intend to propose a DRL-based algorithm to solve the energy efficiency problem. DRL is capable of improving the learning rate by utilizing Deep Neural Networks (DNNs) replaces classical greedy algorithm to train the learning process.

A. DFMDP model

Typically, a DFMDP is defined by a tuple (*S, A, p, r*), where *S* is a finite set of states, *A* is a finite set of actions, *p* is a transition probability from state *s to* state *s'* ($\forall s \in S, \forall s' \in S$) after action *a* ($\forall a \in A$) is performed, and *r* is the immediate reward obtained after *a* ($\forall a \in A$) is executed [26]. We denote π as a policy that is a mapping from a state to an action. Our goal is to find the optimal policy denoted as π^* to maximize the reward function over a finite time in the DFMDP. Therefore, the detailed tuple in our proposed model is designed as follows:

1) The state of each device d_j in the k-th time slot can be denoted as $s_{d_j}^k \in S$. In this model, $s_{d_j}^k$ contains two parts: $\mathrm{D}Q_{d_j}^k$and $EQ_{d_j}^k$. They are the data and energy queue lengths of j-th device at the beginning of the k-th time slot, respectively. To ensure the completeness of the exploration of state space, $\mathrm{D}Q_{d_j}^k$ and $EQ_{d_j}^k$ are specified to be an integer and take the values of $[0, 1, \ldots,\ DQ_{d_j}^{max}]$ and $[0, 1, \ldots,\ EQ_{d_j}^{max}]$, respectively.
2) The action a $(\forall a \in A)$ in this scenario should be the resource allocation strategy, which includes transmission mode α_{d_j}, time slot allocation $\beta_{d_j}^k$, relay selection $\delta_{d_j}^k$ and power allocation $p_{j,k}$. To make sure the integrity of the exploration of action space, $p_{j,k}^d$, $p_{j,r,k}^{s\to r}$ and $p_{j,r,k}^{r\to d}$ should be subject to the maximum transmission power p_j^{max}.
3) The reward r is the immediate reward corresponding to current state-action pair, which is given in Equation 22.

However, the traditional value-based algorithms such as Monte Carlo [27] and Temporal Difference (TD) [28] algorithms have some shortcomings in practical applications, for instance, they cannot handle the tasks in continuous action space efficiently and the final solution may not be global optimal. Therefore, we intend to adopt a policy-based algorithm in this work. The goal of the proposed algorithm is to find out the optimal policy $\pi^*(s_{d_j}^k)\to A$ for each state in each device's complete state-action space. By this way, we can obtain the energy efficiency performance under the influence of random and fluctuant data and energy arriving model.

B. Deep reinforcement learning algorithm

To address the formulated DFMDP problem, the classical Q-learning algorithm is an effective tool [29]. As we mentioned previously, our goal is to find the optimal policy $\pi^*(s_{d_j}^k)\to A$ for each user to maximize the energy efficiency, the Q-learning algorithm also can be a candidate algorithm to obtain the solution. The core idea behind the Q-learning algorithm is to first define the value function $V^{\pi}(s_{d_j}^{k}) \rightarrow r$ that represent the expected value gotten by policy π from each state $s_{d_j}^{k} \in S$. The value function V^{π} for policy π quantifies the goodness of the policy via an infinite horizon and discounted MDP. To simplify the discussion, we use $V^{\pi}(s)$ to represent $V^{\pi}(s_{d_j}^{k})$ and which can be expressed as Equation 24:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma \cdot r^{k}(s^{k},a^{k}) \,\middle|\, s^{0} = s\right] = \mathbb{E}_{\pi}\left[r^{k}(s^{k},a^{k}) + \gamma \cdot V^{\pi}(s^{k+1}) \,\middle|\, s^{0} = s\right] \quad (24)$$

Because we aim to find the optimal policy π^{*}, the optimal action at each state can be found by means of the optimal value function, as Equation 25:

$$V^{*}(s) = \overset{max}{a^{k}}\left\{\mathbb{E}_{\pi}\left[r^{k}(s^{k},a^{k}) + \gamma \cdot V^{\pi}(s^{k+1})\right]\right\} \quad (25)$$

If we denoted $Q^{*}(s,a) \triangleq r^{k}(s^{k},a^{k}) + \gamma \cdot \mathbb{E}_{\pi}[V^{\pi}(s^{k+1})]$ as the optimal Q-function, the optimal value function can be rewritten as $V^{*}(s) = \overset{max}{a}\{Q^{*}(s,a)\}$. The $Q^{*}(s,a)$ can be obtain through iterative process according to the Equation 26.

$$Q^{k+1}(s^{k},a^{k}) = Q^{k}(s^{k},a^{k}) + \alpha\left[r^{k}(s^{k},a^{k}) + \gamma max Q^{k}(s^{k},a^{k+1}) - Q^{k}(s^{k},a^{k})\right] (26)$$

Where, α is the learning rate to determine the impact of new information to the existing Q-value, and $\gamma \in [0,1]$ is the discount factor.

However, the Q-learning algorithm can get the optimal policy when the state-action spaces are small. Practically, such as in our complicated model, the spaces are normally large. As a result, Q-learning algorithm may insufficient to find the optimal policy within the acceptable time. Hence, we implement a Deep Q-Network (DQN) to replaces the Q-table in the classical Q-learning algorithm as a DRL-based algorithm to derive the approximate value of $Q(s^{k}, a^{k})$. Therefore, the Q-value of DQN in *k*-th time slot can be rewritten as $Q(s^{k}, a^{k},\omega)$, where ω is the weight of DNN. After the approximation, the optimal policy $\pi^{*}(\mathrm{s})$ will be presented by Equation 27:

$$\pi^{*}(\mathrm{s}) = \arg\overset{max}{a^{k}} Q^{*}(s^{k}, a^{k+1},\omega) \quad (27)$$

Where, $Q^{*}(s,a)$ is the optimal Q-value via DNN approximation. DQN will choose the approximated action $a^{k+1} = \pi^*(s^{k+1})$. Then the approximated $\widetilde{Q}(s^k, a^k)$ can be given as Equation 28:

$$\widetilde{Q}(s^k, a^k, \omega) = r(s^k, a^k, \omega) + \gamma max_{a^{k+1}}[\mathrm{Q}(s^{k+1}, a^{k+1}, \omega)] \qquad (28)$$

The value of ω is updated by minimizing the loss as expressed in Equation 29. We present the proposed DRL-based resource allocation algorithm in our formulated model in Algorithm 1.

$$\mathrm{L} = \mathrm{E}\left[\left(\widetilde{Q}(s^k, a^k, \omega) - \mathrm{Q}(s^{k+1}, a^{k+1}, \omega)\right)^2\right] \qquad (29)$$

Algorithm 1. The DRL-based resource allocation algorithm

```
initialize replay memory $D$ to the number of devices $M$
initialize the Q-network Q with random weights ω
**for** *episode = 1 to U* **do**
    Initialize the EH-CMNs scenario, receive initial observation state $s_1$
    **for** *k = 1 to K* **do**
        select a random action $a^k$ (energy harvesting and traffic served
        time in time slot $k$, $\alpha_{d_j}$, $\beta^k_{d_j}$, $\delta^k_{d_j}$, $p_{j,k}$) with the probability ε
        Otherwise select $a^k = argmax\ Q^*(s^k, a^k, \omega)$
        perform action $a^k$ and observe immediate reward $r^k$ $(EE^k_{d_j})$ and
        next state $s^{k+1}$ $(DQ^{k+1}_{d_j}$ and $EQ^{k+1}_{d_j})$
        store transition $(s^k, a^k, r^k, s^{k+1})$ in $D$
        select randomly samples $c(s_i, a_i, r_i, s_{i+1})$ from $D$
        the weights of the of DNN are updated by using stochastic
        gradient descent with respect to the ω to minimize the loss
        as Equation 29
        update the policy $\pi(s^k) = \arg\overset{max}{_{a^{k+1}}} Q^*(s^k, a^{k+1}, \omega)$ after every
        a fixed number of steps
    **end for**
```

14. end for

V. Simulation Results and Analysis

In this section, we compare the proposed scheme with other three schemes: (1) direct transmission-only scheme; (2) random power allocation scheme; and (3) classical Q-learning resource allocation scheme. To verify the effectiveness of the proposed scheme, we evaluate the performance in terms of energy efficiency and the convergence speed.

A. Simulation setting

In simulation, we consider a scenario of EH-CMNs underlaying cellular network, in which cognitive-enabled devices and CUs are deployed randomly in a cellular cell with the radius of 800m. BS is located at the center of this topology. The communication range between two devices is randomly set between [20, 50] m and a minimum distance between CU and M2M pairs is set to 200m in order to avoid serious interference. Simultaneously, we suppose that only M2M devices are equipped with the energy harvesting function, and the energy harvesting process is Poisson-distributed with a rate λ_e at arrival instants t^k. The traffic arriving process is also Poisson-distributed with a rate λ_d at arrival instants t^k. The DQN framework has no prior knowledge about them. We set 150 time instants for each episode and the energy efficiency will be averaged to reduce the instability. The DNN utilized in DQN framework contains two fully connected hidden layers, in which 64 neurons and 32 neurons are set respectively. The implementation of DNN is carried out by using Tensorflow 1.0. For each configuration, we generate 100 independent runs and average the performance of energy efficiency. Moreover, the confidence intervals with 95% probability are also provided in each performance evaluation figure. All of the detailed simulation variables used in this paper are summarized in Table 1.

Table 1 Simulation parameters setting

Parameters	Value
R	800 m
Distance of two devices	Random distributed in [20, 50]

N	[1:1:30]
M	[6:2:60]
B	180 KHz
ρ_n	-174 dBm/Hz
p_i^{max}	20 dBm
p_j^{max}	17 dBm
λ_d	[1:1:8] packet/time slot
λ_e	[1:1:8] packet/time slot
ψ	200
τ_k	0.5 ms
PS_{data}	8 bits/packet
PS_{energy}	0.0005 J/packet
$DQ_{d_j}^{max}$	50 packets
$EQ_{d_j}^{max}$	50 packets
TR_{th}/B	8 bps/Hz and 12 bps/Hz

B. Results and analysis

(1) The influence of learning rate α and discount factor γ on energy efficiency

In order to avoid other factors influencing the performance, we first evaluate the influence of learning rate α and discount factor γ on energy efficiency. We implement a scenario in which one CU and one direct transmission-only M2M pair are deployed. The M2M pair multiplexes the uplink spectrum resource of the CU with $\lambda_e = 3$ and $\lambda_d = 5$. Figures 2(a), 2(b) and 2(c) show the average energy efficiency under different values of α and γ. From the results, we can see that either the decrease of learning rate α or the increase of discount factor γ will cause the instability of energy efficiency in the proposed resource allocation algorithm. This is because a smaller α leads to less exploration. In such case, the proposed algorithm increasingly concentrates on the DNN which has more immediate effect in increasing the users' utility. Contrarily, a smaller γ means that the policy gives priority to the immediate reward and a larger value of γ causes more foresight in the policy updating. Therefore, from the long term perspective, a larger γ will increase the average utility in the long term [30]. Moreover, another interesting finding can be obtained from Figure 2(c) is that although a large value of γ can increase energy efficiency from the long term perspective, in the case of conjunction with α, a larger α will get a fast convergence speed, but the energy efficiency fluctuates largely after convergence and meanwhile a smaller α will cause a slow convergence speed, but the energy efficiency is more stable. Furthermore, we also tried some more complex scenarios in which more M2M devices are deployed, but the influences of learning rate α and discount factor γ are similar. For simplicity and ease of understanding, we only demonstrate this scenario and we can obtain a vivid result that the proposed algorithm performs better in the case of a higher α and lower γ. Consequently, we set $\alpha = 0.9$ and $\gamma = 0.1$, respectively, in the following simulations.

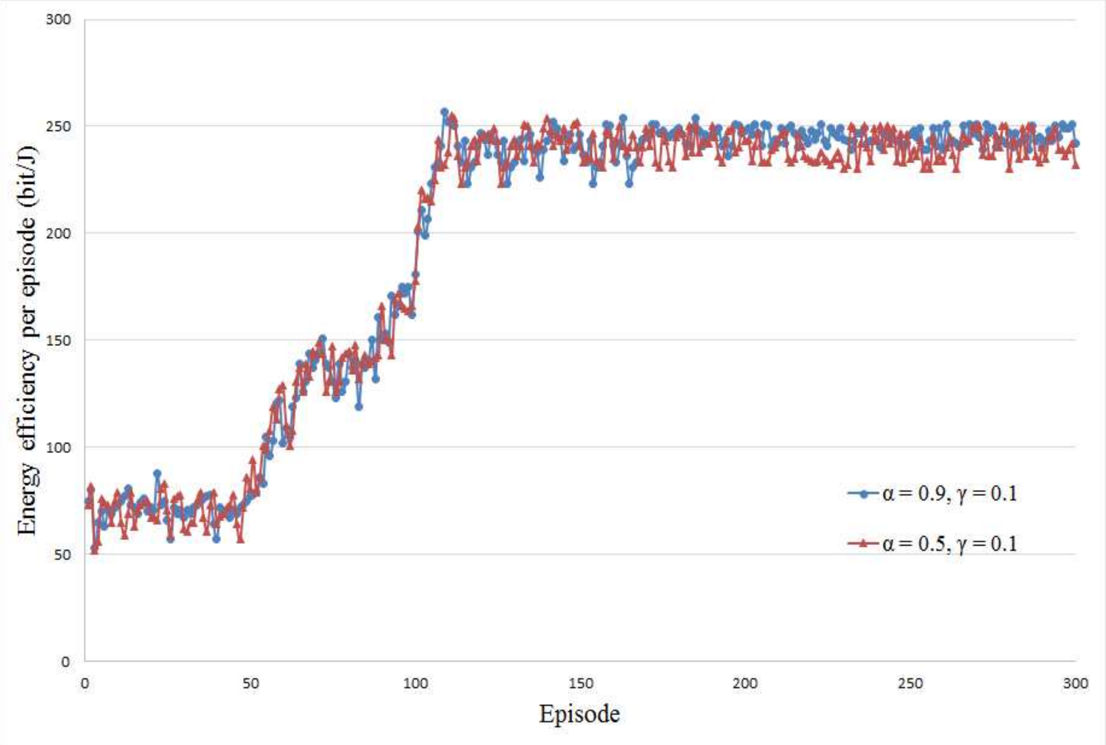

Figure 2(a) Influence of α and γ on energy efficiency ($\alpha = 0.9$ and 0.5, $\gamma = 0.1$)

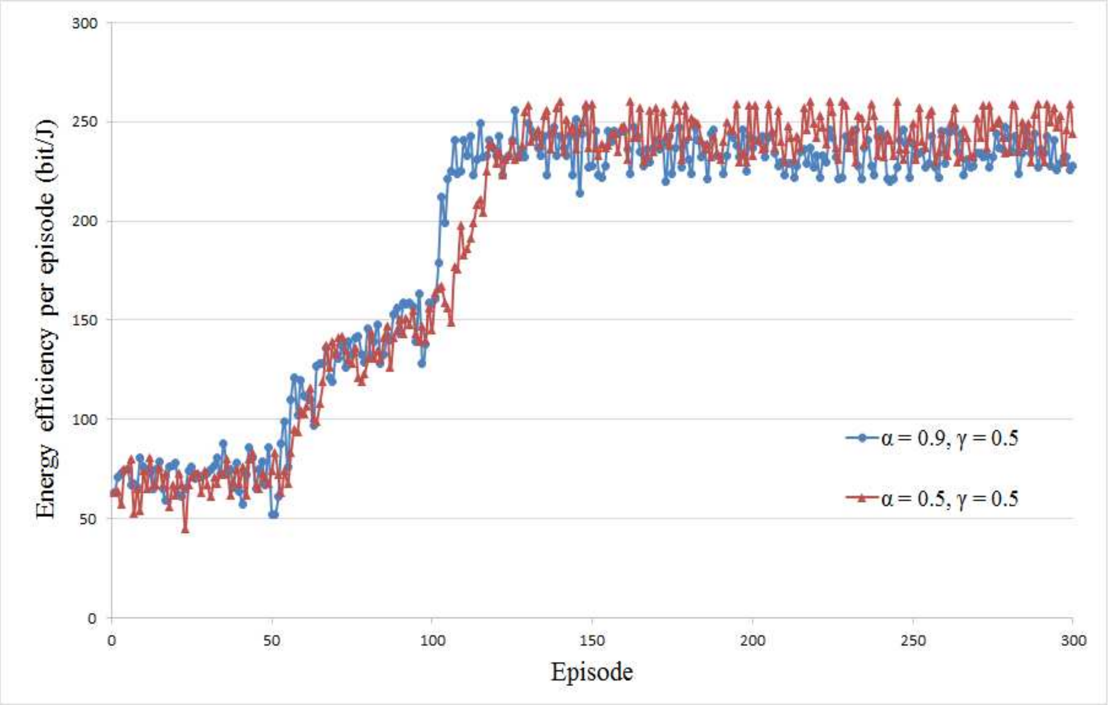

Figure 2(b) Influence of α and γ on energy efficiency (α = 0.9 and 0.5, γ = 0.5)

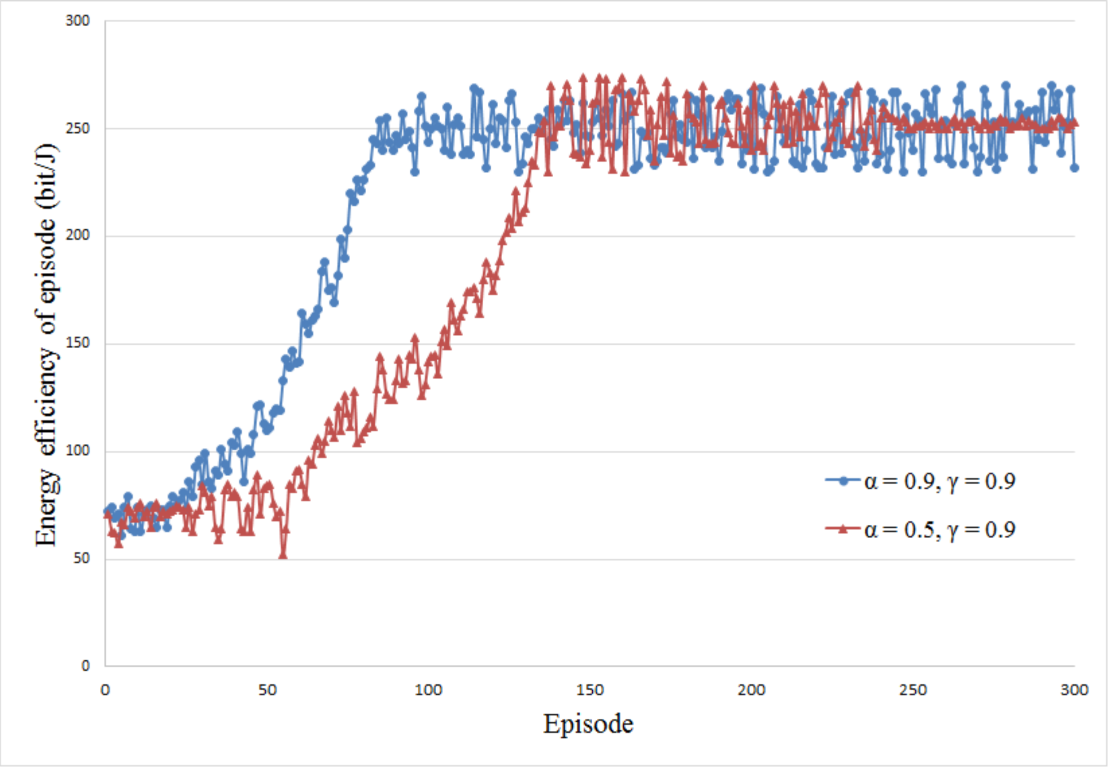

Figure 2(c) Influence of α and γ on energy efficiency (α = 0.9 and 0.5, γ = 0.9)

Comparison between the proposed DRL algorithm and Q-learning algorithm

Figure 3 illustrates the optimization processes for energy efficiency of the proposed DRL algorithm and Q-learning algorithm. The simulation result gives two observations. First, Q-learning algorithm performs better than DRL algorithm before

70 episodes. This is because the fact that in the first 70 episodes, DRL algorithm also selects actions randomly and stores the feedbacks into replay memory. After 70 episodes, DRL algorithm starts to learn from the experience. It is worth noting that the proposed DRL algorithm is unstable initially. However, as the episodes increase, the performance trends to stable. Second, Q-learning algorithm performs quite stable after 50 episodes rather than 100 episodes for the proposed algorithm in this scenario, which indicates that Q-learning algorithm achieves convergence faster than the proposed DRL algorithm. Nevertheless, the proposed DRL algorithm remain obtains the better energy efficiency performance within an acceptable time.

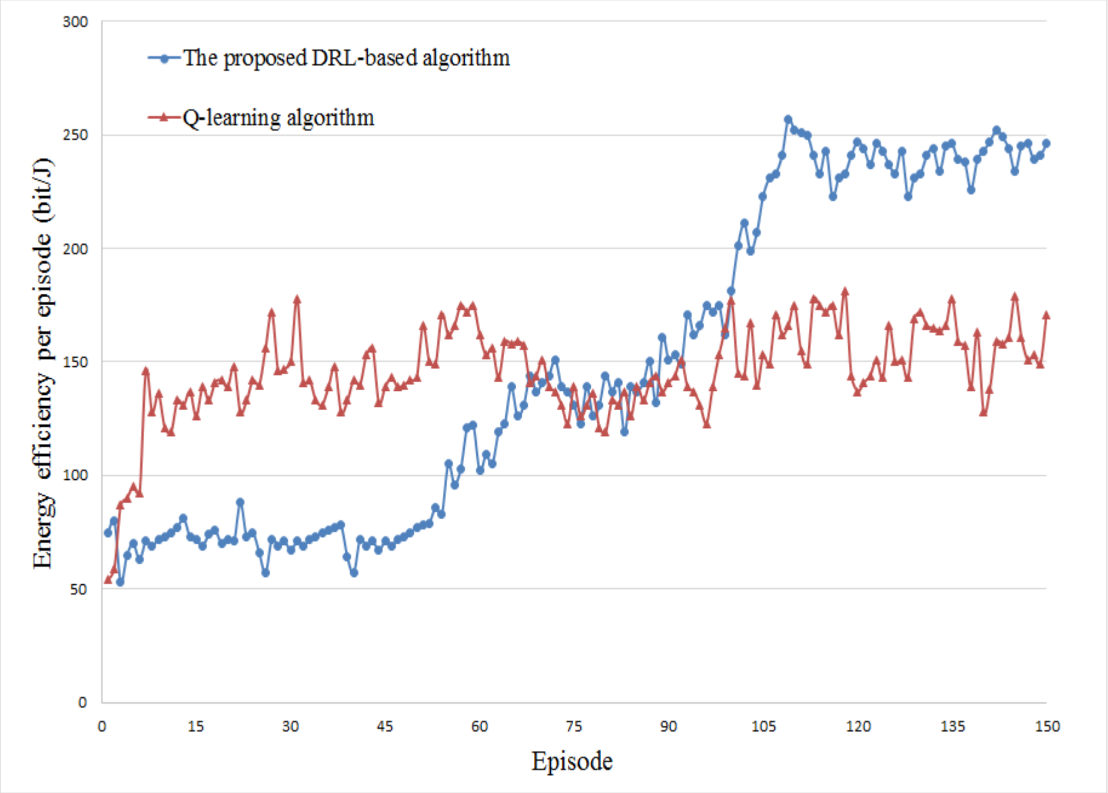

Figure 3 The optimization process for energy efficiency

(2) The influence of the number of CUs with different QoS constraints

Figures 4 and 5 present the energy efficiency for different number of CUs with the QoS constraints of spectral efficiency $TR_{th}/B = 8bps/Hz$ and $TR_{th}/B = 12bps/Hz$, respectively. From the results, it can be observed that the proposed algorithm has the higher energy efficiency as compared to the direct transmission-only scheme. This is because the proposed scheme takes into consideration of both direct and cooperative transmission modes and the optimal transmission mode can be selected by the DRL algorithm. Compared with the random power allocation scheme, the proposed scheme jointly considers the transmission mode, relay selection and allocated time slot to determine the level of transmission power rather than the random allocation in the random power allocation scheme. It also can be observed that the random power allocation scheme has the worst energy efficiency. Remarkably, the performance of Q-learning algorithm initially is better than the proposed DRL algorithm. However, as the number of CUs increases to 10, the DRL algorithm outperforms Q-learning algorithm. There are two reasons for this observation: 1) when the number of CUs is small, the resource allocation problem is simpler. However, the DRL algorithm has the more computation complexity as compared to the Q-learning algorithm. Thus, the energy efficiency is lower; 2) as the number of CUs increases meanwhile the DRL algorithm starts to learn from the experience rather than the replay memory, the performance of the DRL algorithm goes up. Another interesting find in this simulation is that as the number of CUs increases to 14, the performance of the direct transmission mode-only scheme overs Q-learning-based scheme. This is because more devices are implemented will reduce the distance between two devices. It is worth noting that even if the direct transmission mode-only scheme does not support cooperative transmission, it still adopts DRL algorithm to obtain relative optimized energy efficiency. Moreover, with the conjunction of Figures 4 and 5, we can see that the higher energy efficiency can be obtained with the lower CUs QoS constraint ($TR_{th}/B = 8bps/Hz$). This is due to the fact that smaller of TR_{th}/B means less $p_{i,}$, which results less interference to M2M communication, the energy efficiency will increase.

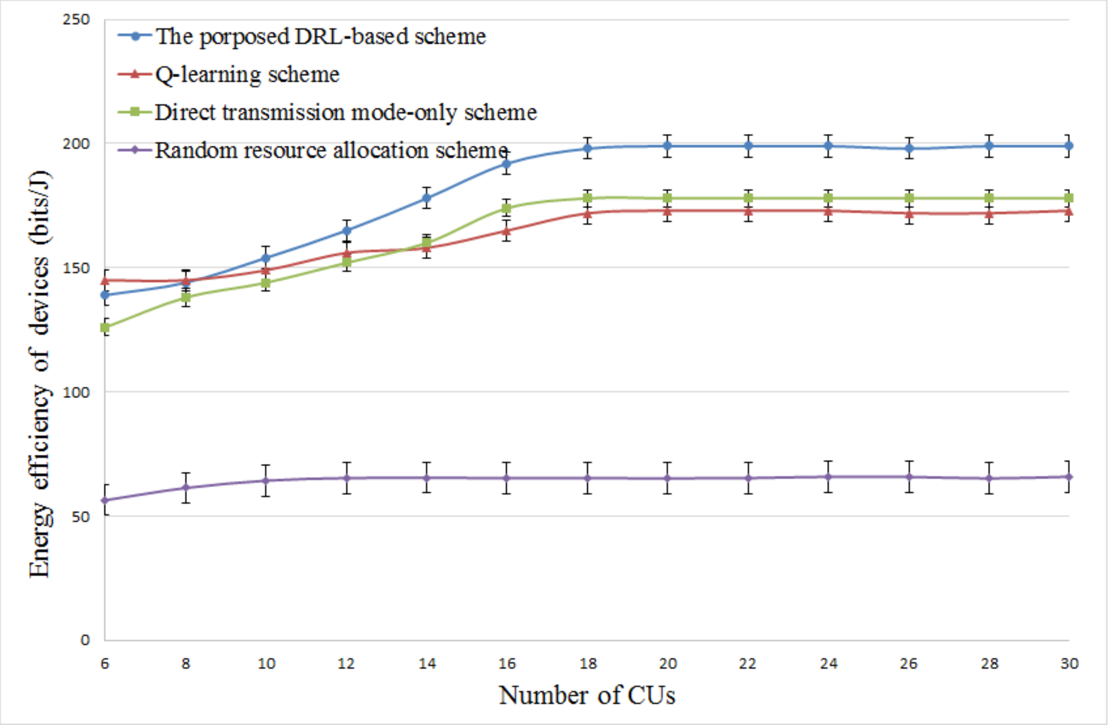

Figure 4 Energy efficiency versus different number of CUs with $TR_{th}/B = 8bps/Hz$

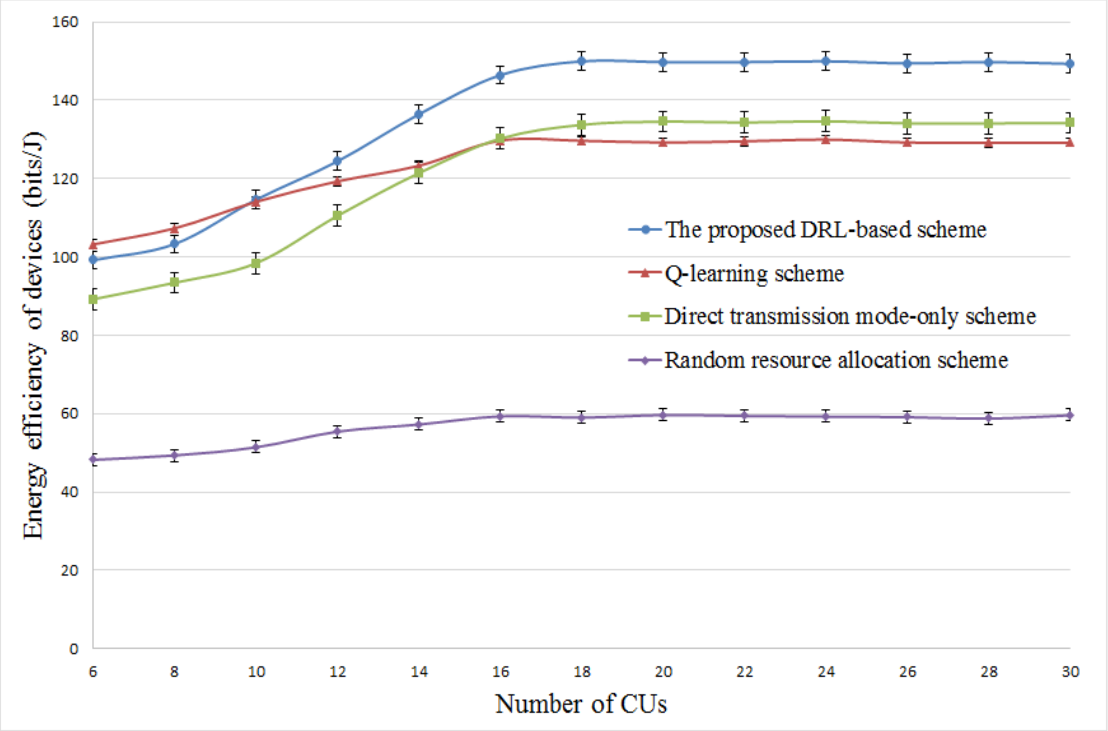

Figure 5 Energy efficiency versus different number of CUs with $TR_{th}/B = 12bps/Hz$

(3) The influence of the number of devices deployed

Figure 6 gives the average energy efficiency for a different number of devices deployed in EH-CMNs. In this evaluation, the number of CUs constantly set to 10 and they are randomly deployed in the scenario. Simulation results depict that the proposed DRL algorithm has the highest energy efficiency among the Q-learning scheme, the direct transmission mode-only scheme and the random power allocation scheme. Initially, when the number of devices is small, the proposed DRL scheme, Q-learning scheme and direct transmission mode-only scheme have the similar performance. However, as more devices are deployed, the DRL scheme outperforms other two schemes. Meanwhile, from the results we can see that the average energy efficiency reaches the highest point for most algorithms (except the random power allocation algorithm) as the increase of the devices. However, while the number of devices further increases, the energy efficiency is reduced. Interestingly, it can be observed that the proposed DRL scheme enables to maintain the widest range of the number of deployed devices with the highest energy efficiency. Another valuable finding is that the average energy efficiency of devices under direct transmission mode-only scheme reduces drastically and goes to the lowest value approximate at 31 bits/J. This is because when the number of devices is larger, the distance between *DU_Tx* and *DU_Rx* is larger, which causes higher energy consumption. Thus, a significant conclusion can be obtained is that the transmission mode selection makes an important contribution to the energy efficiency improvement.

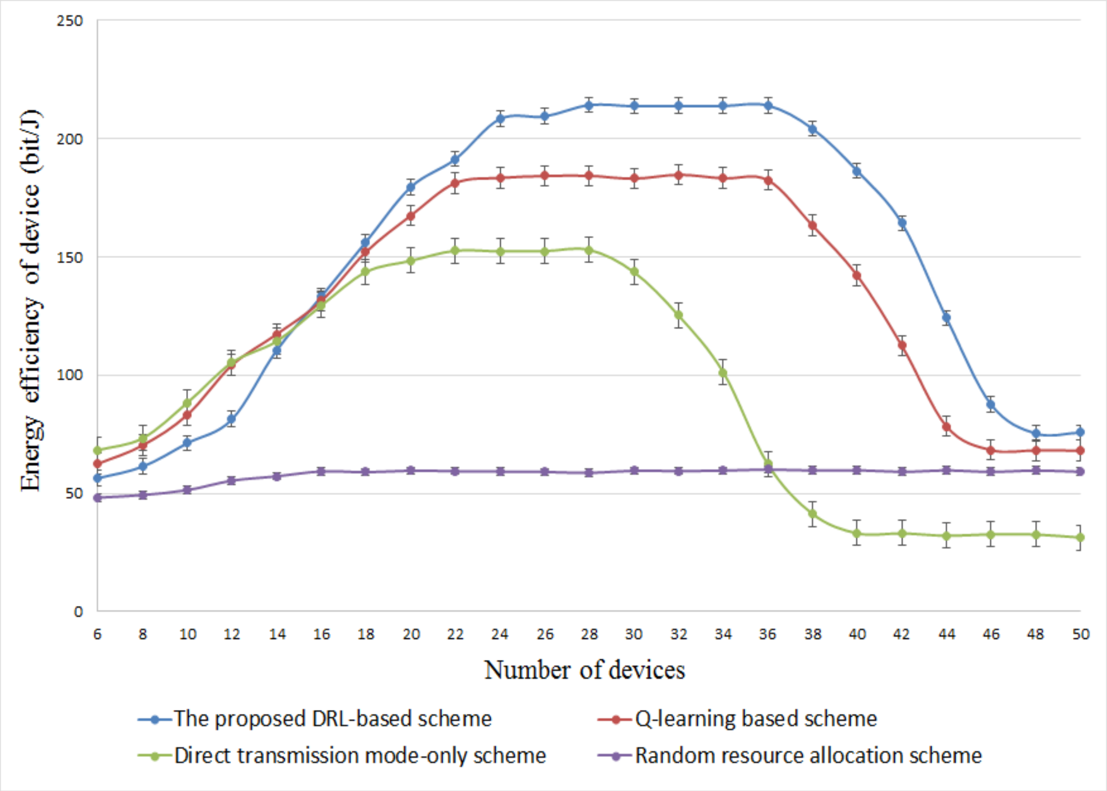

Figure 6 Energy efficiency versus different number of devices deployed

(4) The influence of energy harvesting rate λ_e

Figure 7 presents the energy efficiency with different energy harvesting rates λ_e. In this simulation, the data arrival rate λ_d is set to 3 to emulate the small bursty data of M2M communication in IoTs. From the results, it is clear that the proposed DRL scheme and Q-learning scheme can obtain relatively higher energy efficiency. With the increase of λ_e, the energy efficiency is improved sharply. This is because more energy can be harvested in each time slot with the higher λ_e. Meanwhile, we found that DRL scheme always has the highest energy efficiency along with the increase of λ_e, the reason is due to that it enables to obtain an optimal correlation between energy harvesting time, transmission mode, relay selection, and power allocation. Finally, we found that the random resource allocation scheme does not change the energy efficiency. This is because it does not take into account the available energy when allocating transmission power.

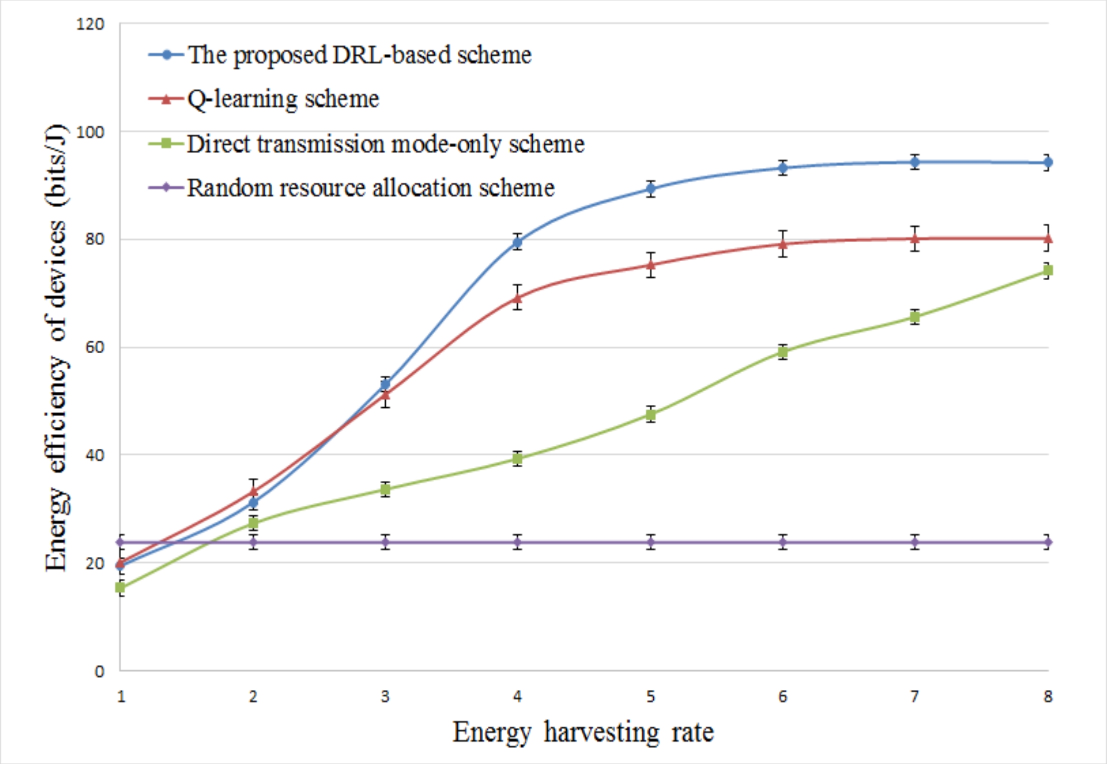

Figure 7 Energy efficiency versus different energy harvesting rate λ_e

VI. Conclusion

The main motivation of this paper is to study the resource allocation scheme for EH-CMNs. Unlike the traditional M2M communications, the available energy will be another vital issue that should be considered in the resource allocation. Specifically, with the goal of maximizing the average energy efficiency, we formulate the resource allocation problem to be a decentralized DFMDP, in which the transmission mode, relay selection, allocated time slot, power allocation, and energy constraint of each device are considered. Owing to the high complexity of the problem, we also propose a DRL algorithm to solve the maximization problem. Through extensive simulations, it is shown that the proposed scheme enables each agent adaptively learns from environment to enhance the energy efficiency significantly for different network settings. Additionally, the proposed DRL algorithm with the low convergence speed is more suitable for the scenarios of EH-CMNs.

Acknowledgement

This project was sponsored by China Scholarship Council, Introduction of high-level talents and overseas returnee’s scientific fund in Nanjing Forestry University

(No. GXL015).

References

[1] Z. Zhou, G. Jie, Y. He, & Z. Yan, Software defined machine-to-machine communication for smart energy management. IEEE Communications Magazine, 55(10) (2017) 52-60.

[2] C. Zhang, Z. Zhou, P. Liu, & B. Gu, Resource allocation for energy harvesting based cognitive machine-to-machine communications. IEEE ICC workshops, 2019, pp. 1-6.

[3] C. Y. Oh, D. Hwang, & T. J. Lee, Joint access control and resource allocation for concurrent and massive access of m2m devices. IEEE Transactions on Wireless Communications, 14(8) (2015) 4182-4192.

[4] H. Y. Hsieh, C. H. Chang, & W. C. Liao, Not every bit counts: data-centric resource allocation for correlated data gathering in machine-to-machine wireless networks. ACM Transactions on Sensor Networks, 11(2) (2015) 38.

[5] F. Ghavimi, Y. W. Lu, & H. H. Chen, Uplink scheduling and power allocation for m2m communications in sc-fdma based lte-a networks with qos guarantees. IEEE Transactions on Vehicular Technology, 66(7) (2017) 6160-6170.

[6] Z. Zhou, Y. Guo, Y. He, X. Zhao, & W. M. Bazzi, Access control and resource allocation for m2m communications in industrial automation. IEEE Transactions on Industrial Informatics, 15(5) (2019) 3093-3103.

[7] H. S. Jang, H. S. Park, & D. K. Sung, A non-orthogonal resource allocation scheme in spatial group based random access for cellular m2m communications. IEEE Transactions on Vehicular Technology, 66(5) (2017) 4496-4500.

[8] M. G. Kibria, G. P. Villardi, K. Ishizu, & F. Kojima, Throughput enhancement of multicarrier cognitive m2m networks: universal-filtered ofdm systems. IEEE Internet of Things Journal, 3(5) (2017) 830-838.

[9] J. A. Paradiso, & T. Starner, Energy scavenging for mobile and wireless electronics. IEEE Pervasive Computing, 4(1) (2005) 18-27.

[10] M. L. Ku, L. Wei, C. Yan, & K. J. R. Liu, Advances in energy harvesting

communications: past, present, and future challenges. IEEE Communications Surveys & Tutorials, 18(2) (2017) 1384-1412.

[11] H. Gao, W. Ejaz, & M. Jo, Cooperative wireless energy harvesting and spectrum sharing in 5g networks. IEEE Access, 4 (2016) 3647-3658.

[12] 3GPP TR 37.868 V11.0.0, Study on ran improvements for machine-type communications. September 2011.

[13] K. Lee, J. S. Shin, Y. Cho, K. S. Ko, & H. Shin, A group-based communication scheme based on the location information of MTC devices in cellular networks. IEEE ICC workshops, 2012, pp. 4899-4903.

[14] X. Xiong, L. Hou, & L. Zhao, A group-based massive multiple access scheme in cellular m2m networks. Computer Communications, 121 (2018) 44-49.

[15] M. Amadeo, O. Briante, C. Campolo, A. Molinaro, & G. Ruggeri, Information-centric networking for m2m communications: design and deployment. Computer Communications, 89-90 (2016) 105-116.

[16] P. Rawat, K. D. Singh, & J. M. Bonnin, Cognitive radio for m2m and internet of things: a survey. Computer Communications, 94 (2016) 1-29.

[17] A. Aijaz, & A. H. Aghvami, Cognitive machine-to-machine communications for internet-of-things: a protocol stack perspective. IEEE Internet of Things Journal, 2(2) (2015) 103-112.

[18] H. K. Lee, D. M. Kim, Y. Hwang, S. M. Yu, & S. L. Kim, Feasibility of cognitive machine-to-machine communication using cellular bands. IEEE Wireless Communications, 20(2) (2013) 97-103.

[19] K. Zheng, F. Hu, W. Xiang, M. Dohler, & W. Wang, Radio resource allocation in lte-advanced cellular networks with m2m communications. IEEE Communications Magazine, 50(7) (2015)184-192.

[20] S. Huang, Z. Wei, Y. Xin, Z. Feng, & P. Zhang, Performance characterization of machine-to-machine networks with energy harvesting and social-aware relays. IEEE Access, 5 (2017) 13297-13307.

[21] M. J. Shih, H. Y. Wei, & G. Y. Lin, Two paradigms in cellular internet-of-things access for energy-harvesting machine-to-machine devices: push-based versus

pull-based. IET Wireless Sensor Systems, 6(4) (2016) 121-129.

[22] Z. Yang, X. Wei, Y. Pan, C. Pan, & C. Ming, Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for iot. IEEE Internet of Things Journal, 5(1) (2017) 229-245.

[23] Z. Y. Zhou, C. T. Zhang, J. W. Wang, & B. Gu, Energy efficiency resource allocation for energy harvesting based cognitive machine-to-machine communications. IEEE Transactions on Cognitive Communications and Networking, 5(3) (2018) 595 – 607.

[24] P. Mitran, On optimal online policies in energy harvesting systems for compound poisson energy arrivals. IEEE International Symposium on Information Theory, 2012, pp. 960-964.

[25] L. Baxter, Markov decision processes: discrete stochastic dynamic programming. Technometrics, 37(3) (1995) 1.

[26] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, & Y. C. Liang, et al. (2019). Applications of deep reinforcement learning in communications and networking: a survey. IEEE Communications surveys & Tutorials, 21(4) (2019) 3133-3174.

[27] K. H. Quah, & C. Quek, Mces: a novel monte carlo evaluative selection approach for objective feature selections. IEEE Transactions on Neural Networks, 18(2) (2007) 431-448.

[28] W. Caarls, & E. Schuitema, Parallel online temporal difference learning for motor control. IEEE Transactions on Neural Networks & Learning Systems, 27(7) (2016) 1457-1468.

[29] X. Cai, J. Zheng, & Y. Zhang, A Graph-coloring based resource allocation algorithm for D2D communication in cellular networks. IEEE International Conference on Communications, 2015, pp. 5429-5434.

[30] Y. Zhang, F. Fu, & D. S. M. Van, On-line learning and optimization for wireless video transmission. IEEE Transactions on Signal Processing, 58(6) (2010) 3108-3124.

BIOS

Yi-Han Xu received his Ph.D degrees in Telecommunications Engineering from University of Malaya, Malaysia in 2014. He is currently a conjoint associate professor in the College of Information Science and Technology, Nanjing Forestry University, China and School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. His general research interests include statistical signal processing, Internet of Things, machine learning for various wireless communications.

Yong-Bo Tian received his bachelor degrees in computer science and technology from Nanjing Forestry University, China in 2019. He is currently pursuing master degree in the College of Information Science and Technology, Nanjing Forestry University. His research field includes Internet of Things and wireless and mobile communications.

Prosper Komla Searyoh received his bachelor degrees in computer science and technology from Nanjing University of Information Science and Technology, China in 2018. He is currently pursuing master degree in the College of Information Science and Technology, Nanjing Forestry University. His research field includes Internet of Things and wireless and mobile communications.

Gang Yu received his bachelor degrees in Internet of Things engineering from Nanjing Forestry University, China in 2019. He is currently pursuing master degree with the Department of Electronic and Electrical Engineering, The University of Sheffield. His research field includes Internet of Things, protocol design of wireless networks and data coding techniques.

Yueh-Tiam Yong received the B.Sc. degree and MEng.Sc. degree in Electrical Engineering from the University of Malaya, Malaysia on 2008 and 2011, respective. He received the PhD degree from the Universiti Malaysia Sarawak (UNIMAS) in 2018. Currently, he is a Lecturer of Computer Science at the Faculty of Computer Science and Mathematics, Universiti Teknologi Mara (UiTM). He has more than 10 years' experience in wireless networks and data communication related research. He is also the member of BEM (Board of Engineering Malaysia) and a member of the Institution of Engineers, Malaysia (IEM). His major interests are Wireless Sensor Network and Internet of Things.

Conflicts of Interest: The authors declare no conflict of interest.

Author Contributions: Yi-Han Xu and Yong-Bo Tian, Conceptualization; Yong-Bo Tian, Data curation; Yi-Han Xu and Prosper Komla Searyoh, Formal analysis; Yi-Han Xu, Funding acquisition; Gang Yu and Yueh-Tiam Yong, Investigation; Yi-Han Xu, Methodology; Yi-Han Xu, Project administration; Yi-Han Xu, Resources; Yong-Bo Tian and Prosper Komla Searyoh, Software; Yi-Han Xu, Supervision; Yi-Han Xu, Validation; Yi-Han Xu, Visualization; Yi-Han Xu, Writing – original draft preparation; Yi-Han Xu and Prosper Komla Searyoh Writing – review and editing.