Deep Reinforcement Learning-Based Resource Allocation for 5G Machine-Type Communication in Active Distribution Networks

Qiyue Li[1,2], Hong Cheng[1,2], Yangzhao Yang[3], Haochen Tang[1,2], Zhi Liu[4(✉)], Yangjie Cao[5], and Wei Sun[1,2]

[1] School of Electrical Engineering and Automation, Hefei University of Technology, Anhui, China
[2] Engineering Technology Research Center of Industrial Automation, Hefei, Anhui, China
[3] Shenzhen Cyberaray Network Technology Co., Ltd, Shenzhen, China
[4] The University of Electro-Communications, Chofu, Japan
liu@ieee.org
[5] Zhengzhou University, Zhengzhou, China

Abstract. With the development of smart grids and active distribution networks (ADNs), reliable and low-latency communication is the key to advanced applications such as energy management and situation awareness (SA). However, with the increasing amount of data and location information to be collected, ensuring the real-time transmission of sampling data has become a challenge. In addition, the operating environment of ADNs is complex, and external interference will affect the reliability of transmission. In particular, the occurrence of power emergencies is random, and the high reliability of emergency data transmission caused by emergencies has attracted much attention. Although repeated data transmission in 5G machine-type communication (MTC) can improve the reliability, how to dynamically allocate communication resources according to the transmitted data and external interference remains a problem. To this end, we propose a scheme of repeated data transmission to eliminate the influence of external interference on the outage probability of emergency data transmission. Our scheme is modeled as a dynamic programming problem to maximize the energy efficiency. First, external interference is considered in the calculation of the transmission outage probability of smart meters (SMs), and the number of repeated transmissions of emergency data is placed in the position of the index, which is determined by reaching the target outage probability. Then, to allocate

This work is supported in part by grants from the National Natural Science Foundation of China (52077049, 51877060), Anhui Provincial Natural Science Foundation (2008085UD04), Fundamental Research Funds for the Central Universities (PA2020GDJQ0027, JZ2019HGTB0089, PA2019GDQT0006), and the 111 Project (BP0719039).

C. T. Calafate et al. (Eds.): MONAMI 2021, LNICST 418, pp. 39–59, 2022.
https://doi.org/10.1007/978-3-030-94763-7_4

dynamic resource in real time in a changing environment, we propose a deep reinforcement learning method, which has fast computing speed, can more quickly allocate resources and reduce the delay of data transmission. Simulation results have verified the superiority of the proposed scheme.

Keywords: Situation awareness · Reliable and low-delay data communication · Resource allocation · Deep reinforcement learning

1 Introduction

An active distribution network (ADN) can actively use the adjustable resources in the distribution network to achieve active planning, management and control services. Its purpose is to solve the problem of grid compatibility and renewable resource consumption. An ADN has two key components: I) massive information collection and II) effective monitoring and accurate diagnosis [1]. Figure 1 describes a structure of ADN. Advanced metering infrastructure (AMI) is composed of many SMs to collect a tremendous amount of data in real time. Simultaneously, ADNs require rapid detection of system events (including system faults and power quality fluctuations) to achieve comprehensive situation awareness (SA), which can be used to monitor and identify normal and abnormal activities of ADN [2]. With the increasing demand of ADNs for situation awareness (SA), how to effectively transmit monitoring data (especially emergency data generated by emergencies) is a significant research topic [3].

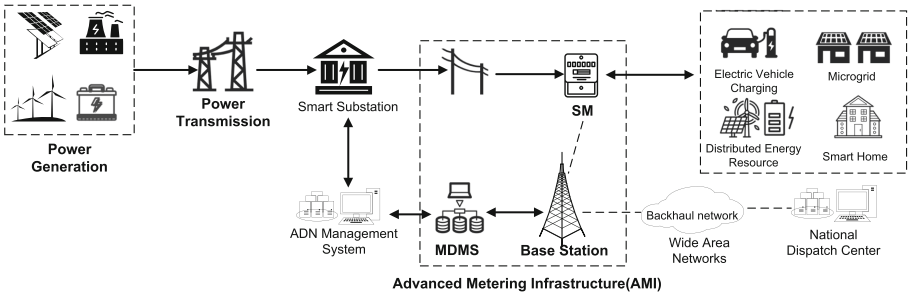

Fig. 1. A typical ADN structure.

Compared with the traditional communication mode, the 5G communication transmission speed is improved, the reliability is more significant, and the energy consumption is reduced. Therefore, we choose to apply 5G communication in AMI for information transmission. However, massive short messages in 5G uplink make it very difficult to schedule and allocate resources (such as time-frequency resources and modulation and coding schemes). In addition, the

occurrence of system failures, power quality fluctuations and other emergencies is random. Low delay and high reliability of emergency data transmission are the keys to ensuring the realization of SA in ADNs. On the one hand, using grant-free scheduling can reduce delay instead of scheduling based on the long-term evolution (LTE) grant-based scheduling method because waiting for grants will increase delay [4]. In this type of fast uplink access without grants, there is no need to send scheduling requests or wait for scheduling grants [5]. On the other, repeated transmission of data is the key factor to improve the reliability performance [4], but using the classic hybrid automatic repeat request (HARQ) retransmission process will introduce additional delay [4]. Therefore, to ensure the reliability of transmission, we repeatedly transmit the emergency data to the instrument data management system (MDMS) without waiting for grant or sending repeated requests. In other words, in data transmission, multiple copies of emergency data are simultaneously transmitted to eliminate the impact of interference on the transmission outage probability. Massive data must be transmitted, and emergency data must be repeatedly transmitted, which requires real-time resource allocation. This paper proposes a resource allocation method based on deep reinforcement learning, which combines with a neural network to speed up the calculation speed, more quickly allocate the resource and reduce the transmission delay.

In 5G networks, high reliability and low delay are the most stringent requirements for communication. In the 3rd-Generation Partnership Project (3GPP) standard, a general URLLC requirement is 99.999% target reliability with 1-ms (two-way) user-plane latency [6]. The reliability here is defined as the percentage of packets that are correctly received within the delay budget. Decreasing the transmission time interval (TTI) length is an efficient method to shorten the latency in the system [7,8]. However, there is a contradiction between low latency and ultra-high reliability. This paper proposes a resource-scheduling method for the repeated transmission of emergency data. This method belongs to grand-free scheduling, i.e., it sends multiple copies of the same packet without waiting for acknowledgment (ACK). Waiting for ACK will increase the delay, and the interference in the transmission process will affect the reliability of communication. Therefore, our method reduces the delay by not waiting for ACK and improves the reliability by increasing the number of emergency data repeated transmissions.

For 5G communication, it is difficult to simultaneously satisfy the requirements of maximal reliability and minimal total resource consumption of data transmission. To maximize the reliability of data scheduling, we consider that in the case of external interference, we can reduce the transmission outage probability of SMs by repeatedly transmitting the emergency data. To reduce the total resource consumption, we establish the energy efficiency formula of the transmission process with the parameters of data packet size, repeated transmission times, SM transmission power and selection of MTC. Then, we model the resource allocation as a dynamic programming problem with the objective function of maximizing the energy efficiency. The Lagrange multiplier method can be used to solve the problem [9]. However, the parameters of data packet size and repeated transmission times are also time-varying. How to adaptively

allocate resources according to the dynamic changes of parameters is also a difficult problem. Therefore, we attempt to use deep reinforcement learning to solve the resource allocation problem of 5G networks. The trained deep reinforcement learning model can quickly solve the problem of resource allocation to achieve resource allocation and scheduling under time-varying conditions. The results show that our method is superior in reliability and calculation speed.

Our contributions are as follows:

1) For emergency data, we propose an algorithm to calculate the number of data repeated transmissions. The algorithm considers noise interference, path loss, and the outage probability of SMs to minimize the number of data transmission repeats while satisfying the target reliability.
2) Aiming at the resource allocation and scheduling problem of a 5G communication network, a deep reinforcement learning method is proposed. This method can allocate and schedule resources in real time according to the network time variance and reduce the total resource consumption to the greatest extent.
3) Extensive simulations are conducted. The simulation results prove the superiority of the designed system and method and provide a new method for 5G communication network resource allocation and scheduling.

The remainder of this paper is organized as follows. Section 2 summarizes the related literature on wireless resource allocation and event detection. Section 3 introduces the system model. Section 4 models the problem. Section 5 uses reinforcement learning to allocate and schedule resources. In Sect. 6, we perform simulations and experiments to verify the effectiveness of our proposed framework. Finally, we summarize the entire paper in Sect. 7.

2 Related Work

A 5G communication network has the characteristics of a large structure and complex interference, which introduces higher requirements for the reliability and low delay of wireless transmission. To satisfy the requirements of low delay and high reliability of wireless communication, we must make full use of scarce bandwidth resources through an appropriate resource allocation algorithm, which can also reduce the energy consumption.

The resource allocation in MTCs is mainly to optimize the allocation of wireless resources in terms of transmission power, time slot, and spectrum. Researchers have proposed many methods for this work. According to the users are mobile and the SMs are static, a two-stage wireless resource allocation method is proposed to maximize the total rate of cellular users while obtaining the minimum transmission power [10]. The method uses machine-to-machine (M2M) communication between the SMs. In [11], a cloud-fog-based model was proposed. This model attempts to summarize the general algorithm for different types of computing services for resource management to achieve load balancing between requests and services. A spectrum allocation technique is designed to

set the minimum BER threshold and evaluate the availability of white holes in unlicensed bands [12]. A wireless resource management technology based on an LTE-A network was proposed [13] to realize automatic meter reading and prevent overload of cellular base stations. However, these resource allocation methods do not consider MCS assignment or the emergency transmission problem in the case of interference.

Reasonable resource allocation is very important to improve the system performance. For example, a 5G-based framework is proposed to reasonably reserve RBS for emergency data and realize the energy-saving resource optimal allocation method [9]. In this method, the resource allocation problem is transformed into a univariate integer programming problem; then, the dual problem is constructed by using Lagrange duality theory. The results show that compared with the traditional greedy algorithm, this algorithm better maximizes the energy efficiency. Although MCS assignment and emergency data transmission are considered in the above method, the reliability of emergency data transmission and dynamic real-time resource allocation are not considered.

To solve the real-time problem of resource allocation, some studies adopt deep reinforcement learning methods. In [14], a real-time adaptive computing resource allocation strategy was designed. In [15], resource allocation methods based on Q-learning and deep reinforcement learning were proposed, but these methods only analyze the real-time performance of resource allocation and do not consider the reliability of resource allocation, i.e., the emergency transmission problem caused by emergencies. Moreover, the Q-learning algorithm will produce a large state space and action space, which greatly increases the computational complexity of the problem.

Because these methods cannot simultaneously satisfy the requirements of high-reliability and low-delay communication, we propose a data scheduling and resource allocation solution for 5G MTC based on deep reinforcement learning. We can improve the reliability of transmission through repeated transmission of emergency data. In addition, we use deep reinforcement learning to achieve dynamic real-time resource allocation and reduce the total resource consumption to maximize the energy efficiency.

3 System Model

AMI is an important component of ADN. To realize the SA of an ADN, we accessed the 5G network to a single-cell uplink channel in AMI. The network structure is shown in Fig. 2. In AMI applications, we assume that all spectrum resources can be allocated to the SMs for 5G uplink data transmission, and the data collected by each sensor in the sampling period can only be sent in a single slot. In addition, because the sampling time of each sensor is much smaller than the 5G time slot, it can be ignored. Simultaneously, we assume that when the sensor detects an emergency, the additional packets containing the emergency event occurrence flag must be immediately sent by SMs.

As shown in Fig. 3, the resource allocation of the 5G network is in the time and frequency domains [16]. In the time domain, the time slot is the

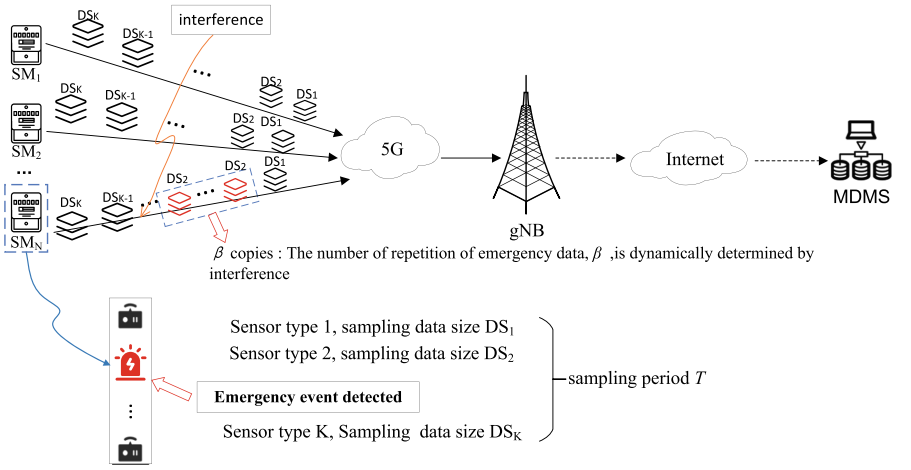

Fig. 2. Network structure (β is the number of repeat transmission times of emergency data).

smallest unit, which contains 14 OFDM symbols. In the frequency domain, the system bandwidth consists of several subchannels, and each subchannel includes 12 consecutive subcarriers. Each resource block (RB) occupies a time slot and a subchannel as the basic unit of data transmission. In addition, 5G NR supports multiple subcarrier spacings and diverse transmission bandwidths. Table 1 shows different transmission bandwidths and corresponding slot durations [17].

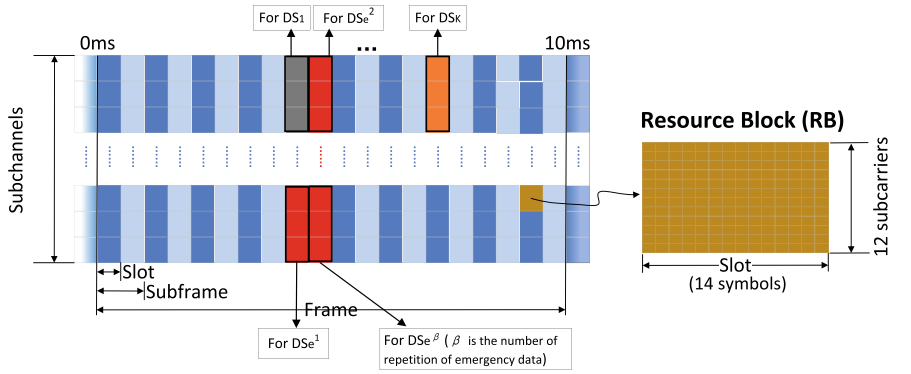

Fig. 3. 5G-based resource block structure in our framework.

According to [18], the next generation of SMs is equipped with voltage sensors, current sensors, and temperature and humidity sensors, which can be used to monitor different types of emergencies. We collect the data sent by SMs

in a gNB, including the data generated by normal power activities and emergency data generated by emergencies. When an emergency occurs, in the data repeated transmission scheme, the number of emergency data repeats is determined according to the average signal-to-noise ratio and the outage probability of SMs; appropriate resource blocks are allocated for normal data packets and emergency data packets that must be repeated in 5G wireless resources. The optimal allocation decision can usually be obtained by solving an optimization problem. However, due to the huge search space and long solving time of this type of resource allocation and scheduling problem, it cannot keep up with the rapid change of sampling data. Therefore, we use the DRL agent to make real-time allocation decisions by collecting the sensor location and packet length information.

4 Problem Formulation

4.1 Repeated Data Transmission Analysis

In this section, we discuss how to reduce the outage probability of SMs [19] by increasing the time of emergency data repeat transmission β in the case of interference. As shown in (1), with the increase in β, the outage probability of SMs will gradually decrease.

$$P_a = \left[0.05(1 + erf(0.1 * (\frac{\gamma - \overline{\gamma}}{\sigma\sqrt{2}})))\right]^{\beta} < a \tag{1}$$

$$\beta = \lceil \frac{\log a}{\log\left[0.05(1 + erf(0.1 * (\frac{\gamma - \overline{\gamma}}{\sigma\sqrt{2}})))\right]} \rceil \tag{2}$$

In Eqs. (1) and (2), σ and a are fixed parameters, γ is the signal-to-noise ratio threshold, $erf(\cdot)$ is the error function, and $\overline{\gamma}$ is the average signal-to-noise ratio of the sensor. Additionally:

$$\overline{\gamma} = \frac{P_r}{N} = \frac{P_t(n)\cdot}{N}(\frac{d_0}{d})^{\lambda} \tag{3}$$

In Eq. (3), $P_t(n)$ is the transmit power of the nth sensor, λ is the path loss exponent, d_0 is the reference distance from the system to the base station, and d is the actual distance from the system to the base station. N is the interference variance, which includes thermal noise N_0 and other interference N_1. In other words, the change in interference will affect the number of repetitions.

4.2 Problem Formulation

In an AMI application system of ADN, it is assumed that there are N SMs, each SM has K sensors, and the emergency packets must be repeatedly transmitted; i.e., there are $L(L = N \times K)$ packets to allocate resources, including a_e emergency packets that need to repeatedly transmit. To maximize the energy efficiency and

optimize the resource allocation, we take the maximal ratio of the size of all transmitted data (bytes) to the energy consumed by all RBs (joules) as the objective function to form problem P0:

$$\begin{aligned} &maxE = \\ &[\sum_{l=1}^{L-a_e}\sum_{m=1}^{M}\sum_{t=1}^{T}(x_{l,m,t} \times DS_{l,m,t}) \\ &+\sum_{1}^{\beta}\sum_{1}^{a_e}\sum_{m=1}^{M}\sum_{t=1}^{T}(x_{\beta,a_e,m,t} \times DS_{\beta,a_e,m,t})] \\ &\div [\sum_{l=1}^{L-a_e}\sum_{m=1}^{M}\sum_{t=1}^{T}(x_{l,m,t} \times \lceil\frac{DS_{l,m,t}}{R_l}\rceil \times P_{l,m,t}) \\ &+\sum_{1}^{\beta}\sum_{1}^{a_e}\sum_{m=1}^{M}\sum_{t=1}^{T}(x_{\beta,a_e,m,t} \times \lceil\frac{DS_{\beta,a_e,m,t}}{R_{\beta,a_e,m,t}}\rceil \times P_{\beta,a_e},m,t)] \\ &\beta \geq 1, a_e \geq 1 \end{aligned} \tag{4}$$

where $x_{l,m,t}$ is a binary variable. When data DS_l are transmitted at time slot t with MCS selection m, $x_{l,m,t} = 1$; otherwise, $x_{l,m,t} = 0$. $P_{l,m,t}$ is the transmission power when transmitting normal data DS_l at time slot t with MCS selection m. $R_{l,m,t}$ is the transport block size (TBS) at time slot t with MCS selection m. β is the number of data repeated transmission. Additionally, when the emergency data $DS_{\beta,a_e,m,t}$ are transmitted at time slot t with MCS selection m, $x_{\beta,a_e,m,t} = 1$; otherwise, $x_{\beta,a_e,m,t} = 0$. $P_{\beta,a_e,m,t}$ is the transmission power when transmitting emergency data $DS_{\beta,a_e,m,t}$ at time slot t with MCS selection m. $R_{\beta,a_e,m,t}$ is the TBS at time slot t with MCS selection m.

The constraints are as follows:

$$P + 10 \times \lg\left\lceil\frac{DS}{R}\right\rceil + (\alpha - 1) \times PL_n - IoT \geq SINR \tag{5}$$

$$P \times \left\lceil\frac{DS}{R}\right\rceil \leq P_{max} \tag{6}$$

$$\begin{aligned} &\sum_{l=1}^{L-a_e}\sum_{m=1}^{M}(x_{l,m,t} \times \left\lceil\frac{DS_{l,m,t}}{R_{l,m,t}}\right\rceil) \\ &+\sum_{1}^{\beta}\sum_{1}^{a_e}\sum_{m=1}^{M}(x_{\beta,a_e,m,t} \times \left\lceil\frac{DS_{\beta,a_e,m,t}}{R_{\beta,a_e,m,t}}\right\rceil) \leq Y \\ &\forall t \in [1,T] \end{aligned} \tag{7}$$

$$\sum_{m=1}^{M}\sum_{t=1}^{T} x \leq 1, \forall n \in [1,N] \tag{8}$$

$$x \in \{0,1\} \tag{9}$$

Constraint (5) is the power control model of the cellular network uplink channel [20], where α is the path-loss compensation factor that can be set from 0.0 to 1.0 in steps of 0.1. PL_n is the downlink path loss measured by the base station, and it can be considered a constant value for fixed SMs; IoT is interference over thermal, which can be ignored in our system. $SINR$ is the signal-to-noise ratio (SINR) requirement of MCS selection m.

Constraint (6) indicates that the total power consumed by each packet is less than the maximum transmission power of each SM.

Constraint (7) indicates that the number of RBs in each time slot must be less than Y.

Constraint (8) indicates that each data packet can only have one MCS selection and transmission in a single slot.

Constraint (9) indicates that x is a binary variable.

Clearly, P0 is a complex mixed integer programming problem, which takes a long time to solve. In addition, to solve P0, we must calculate the parameter of data repetition times, so we cannot obtain the solution in a short time. Therefore, we use the deep reinforcement learning method based on the actor-critic (AC) algorithm to build a real-time scheduling and resource allocation framework. First, the data sent by SMs are classified; then, parameters such as the distance from SM to gNB and the number of emergency data repeats are collected. Finally, the data packet and its related parameters are input into the neural network to output the decision results.

5 Deep Reinforcement Learning Assisted Scheduling and Resource Allocation

In this section, we will introduce the structure of our proposed deep reinforcement learning method for scheduling and resource allocation based on a pointer network. This method is a low-complexity approximation method and can be used to solve our problems.

5.1 Pointer Network

To solve our problem, this paper proposes a scheduling resource allocation algorithm based on a pointer network. The basis of the pointer network is the seq2seq framework [21,22]. Seq2seq is an encoder-decoder structure network, and both its input and output are a sequence. The encoder transforms a variable-length signal sequence into a fixed-length vector expression, the decoder transforms the fixed-length vector into a variable-length target signal sequence, and the conditional probability $p(Y|X,\pi)$ solves the output sequence. Here, $X = x_1, x_2, \cdots, x_n$ is the input sequence, $Y = y_1, y_2, \cdots, y_m$ is the output sequence, and the conditional probability conforms to the probability chain rule, which can be expressed as:

$$p(Y|X;\pi) = \prod_{i=1}^{I} p(y_i|y_1, ..., y_{i-1}, X) \tag{10}$$

The resource allocation problem of the 5G single-cell uplink channel in this paper is also a mapping problem and requires that the input sequence and output sequence of the pointer network are exactly identical, but the sequence order changes. Therefore, we add an attention mechanism [23] to seq2seq to form a new pointer network structure (see Fig. 4).

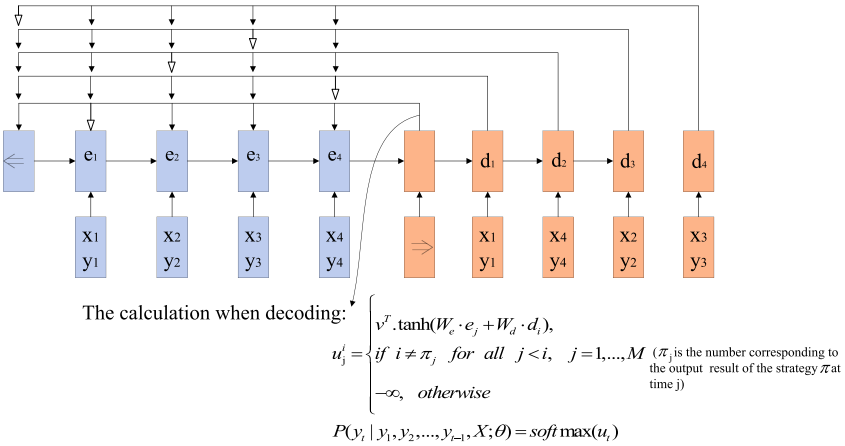

Fig. 4. Pointer network architecture.

The conditional probability model of the attention mechanism is established as follows:

$$u_j^i = \begin{cases} v^T \cdot tanh(W_e \cdot e_j + W_d \cdot d_i), \\ if \ \ i \neq \pi_j \ \ for \ \ all \ \ j < i \qquad for \ \ j = 1, ..., M \\ -\infty, otherwise \end{cases} \tag{11}$$

$$p(y_i|y_1, ..., y_{i-1}, X) = softmax(u^i) \tag{12}$$

where e_j and d_i are the jth and ith hidden layer outputs of the time sequence, respectively; v^T, W_e and W_d are parameters of the neural network that can be trained; π_j is the number that corresponds to the output result of strategy π at time j. The result of the softmax function is a probability distribution, i.e., the weight assigned to the input sequence. The probability distribution can be used as a pointer to the input sequence so that when predicting the elements, we can find the element with the largest weight in the input sequence. According to 11, to ensure that our model only points to the unselected packets, we set the logits of selected package to $-\infty$. During decoding, the pointer network points to the packet in the input packet and stops when the 5G-based resources are exhausted.

5.2 Beam Search

In the decoding process, for reducing the time complexity of the decoding truncation and improving the accuracy, we use the beam search method to decode; i.e., in each step of the decoding stage, according to the probability distribution calculated by softmax, we reserve the top h optimal sequences. A larger h corresponds to a higher calculation cost, but the relative accuracy also improves. The decoding process is described in Algorithm 1.

5.3 Problem Redefinition

Considering the proposed dynamic programming problem of resource allocation in this paper, we combine reinforcement learning with a deep neural network and propose a neural combination optimization model. The model builds a policy network based on the pointer network to output the policy of the problem. To estimate the expected value of the objective function and reduce the gradient variance, an estimation network is constructed. A reinforcement learning framework actor-critic network based on a strategy gradient is used to train model parameters instead of a supervised manner.

For problem P0, we assume that the sensor captures $L(L = N \times K)$ packets. The 3D parameter sequence of L data packets $x = \{(DS_i, \beta_i, d_i)\}_{i=1}^{L}$ in one sampling period T is given. If the data is generated by emergencies, β_i is calculated by the formula in Sect. 4.1. Otherwise, $\beta_i = 1$. The optimization objective of the model is to learn random strategy $p(\pi|x)$ to output the 5G uplink channel resource allocation result with higher energy efficiency under the condition of a given three-dimensional parameter sequence of M data packets $\pi = (s_1, a_1, s_2, a_2, ..., s_T, a_T)$. a_j represents the data packet that is output by the pointer network selecting the MTC and time slot at time j. $a_j = \{R_j, P_j\}$, R_i and P_j are the transport block size (TBS) and transmission power at a certain time slot with MCS selection. s_j depends on the state of the previous time s_{j-1}, MTC and time slot selected by the data packet, which is output by pointer network a_j at time j. The following reward is defined to evaluate the action under a state: We express the energy efficiency formula as the value function $r(\pi)$ of the model.

$$r(\pi) = \sum_{i=1}^{D}[DS_i \times P_i] \div \sum_{i=1}^{D}[(DS_i \div R_i) \times P_i] \tag{13}$$

Among them, D is the number of selected output packets of the pointer network among M data packets, DS_i is the data packet size, P_i is the transmission power, and R_i is the transport block size in one MCS and a specific time slot. We sample the data transmitted through SMs for a sampling period to fit the parameters.

Algorithm 1:Beam Search

Procedure decoding(Number of steps $g = 0$,hash_table = start,BEAM = start)
1. **While** $BEAM \neq \phi$ **do**
2. SET=ϕ
3. **For** each state in $BEAM$ **do**
4. **For** each successor of *state* **do**
5. **if** successor == goal **Return** g + 1
6. SET = SET $\cup \{successor\}$
7. **end for**
8. **end for**
9. **end**
10. BEAM$\neq \phi$
11. g = g + 1
12. **While**(BEAM$\neq \phi$) and (B >| BEAM|) **do**
13. state = successor in SET with smallest h value
14. SET = SET $\setminus \{state\}$
15. **if** state $\notin$ hash_table
16. **if** hash_table is full **Return** ∞
17. hash_table = hash_table $\cup \{state\}$
18. hash_table = hash_table $\cup \{state\}$
19. **end**
20. **Return** ∞
21. **end Procedure**

5.4 Optimize Using the Strategy Gradient

The iterative process of traditional reinforcement learning methods must collect a large amount of data to update the strategy gradient. However, in many complex real scenes, it is difficult to obtain massive training data, so local optimal solutions emerge. This problem can be solved by using the strategy gradient method based on the AC framework to optimize the parameters of the pointer network. The AC-based strategy gradient algorithm is divided into two parts: the actor network selects a behavior based on probability, the critic network determines the score of behavior based on the actor's behavior, and the actor modifies the probability of selecting behavior according to the critic's score. The policy gradient network can only be updated at the end of the round. Each step of the selection affects the network, so an array is required to store the state, probability distribution of the next selection action, and reward of the selection action. At the end of the round, parameters θ update to θ_{t+1}.

First, the actor network is introduced. Because the parameters and repeated transmission times of the packets are different, the order of RBs allocated to the packets by agents is not necessarily identical. We assume that strategy π will lead the agent along path $\pi = (s_1, a_1, s_2, a_2, ..., s_T, a_T)$.

$$p_\theta(\pi) = p(s_1) \prod_{t=1}^{T} p_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t) \tag{14}$$

$r(\pi)$ is the value function, which is a random variable. If the expectation of the reward is obtained, then (15):

$$L(\theta) = E_{\pi \sim p_\theta(\pi)} r(\pi) = \sum_{\pi} r(\pi) p_\theta(\pi) \tag{15}$$

To find the optimal parameters θ, let (16):

$$\max L(\theta) = max \sum_{\pi} r(\pi) p_\theta(\pi) \tag{16}$$

Calculate the gradient of the objective function as shown in (17):

$$\begin{aligned} \nabla L(\theta) &= \sum_{\pi} r(\pi) \nabla p_\theta(\pi) \\ &= \sum_{\pi} r(\pi) p_\theta(\pi) \frac{\nabla p_\theta(\pi)}{p_\theta(\pi)} \\ &= \sum_{\pi} r(\pi) p_\theta(\pi) \nabla log p_\theta(\pi) \end{aligned} \tag{17}$$

The calculation of the gradient is converted to solving for the expectation of $r(\pi) P_\theta(\pi) \nabla log p_\theta(\pi)$. Then, the Monte Carlo method can be used for approximate estimation to obtain the B sampling results of the current strategy π, and we have (18):

$$\begin{aligned} \nabla J(\theta) &= E_{\pi \sim p_\theta(\pi)} [r(\pi) P_\theta(\pi) \nabla log p_\theta(\pi)] \\ &\approx \frac{1}{B} \sum_{i=1}^{B} r(\pi^i) P_\theta(\pi^i) \nabla log p_\theta(\pi^i) \\ &= \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{T_i} (r(\pi^i) P_\theta(\pi^i) \nabla log p_\theta(a_t^i | s_t^i) \end{aligned} \tag{18}$$

Since the sum of the actions taken and the probability is 1, there may be a situation in which the probability value of a good action decreases, and the probability of a bad action increases after normalization. Therefore, it is necessary to introduce a baseline b to make $\nabla L(\theta)$ positive and negative and rewrite it as shown in (19):

$$\nabla J(\theta) = \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{T_i} (r(\pi^i) - b_i) \nabla log p_\theta(a_t^i | s_t^i) \tag{19}$$

Then, the parameters are optimized by using the strategy gradient method and stochastic gradient ascending method. The gradient of (19) is confirmed using the well-known REINFORCE algorithm [24]:

A common baseline choice is the exponential moving average of network rewards over time. Using parameter benchmarks to estimate the expected

resource size $E_{\pi\sim p_\theta(\pi)}r(\pi)$ can usually improve learning. Therefore, an auxiliary network was introduced, which is called the critic network, whose parameters are θ_ν. The critical network uses the mean square error of objective predictions $E_{\pi\sim p_\theta(\pi)}r(\pi)$ and actual resource $D(\theta_\nu)$ to train with stochastic gradient ascent, as shown in (20):

$$D(\theta_\nu) = \frac{1}{B}\sum_{i=1}^{B} \parallel b_{\theta_\nu} - E_{\pi\sim p_\theta(\pi)}r(\pi) \parallel_2^2 \tag{20}$$

Critical network structure for resource allocation: First, we map the input sequence to baseline prediction b_{θ_c}. The critic network consists of three network modules: (1) an LSTM encoder, (2) an LSTM process block, and (3) a neural network decoder. The encoder of this network has an exactly identical structure to the pointer network encoder, and it encodes the input sequence into a sequence of latent storage states and hidden states. A process block of a critic network, such as [21], is essentially an attention mechanism. At the end of the process block, the hidden state obtained by the neural network decoder is decoded into the baseline prediction. Our training algorithm is described in Algorithm 2:

5.5 Resource Scheduling Framework

This paper proposes a resource allocation algorithm based on reinforcement learning. The algorithm framework is shown in Fig. 5. First, sample data containing interference signals are sampled at time t, and these data are sent to the 5G network by SMs. The 5G network will classify normal data and emergency data according to the additional data packet with emergency data flags.

Algorithm 2:Actor-Critic

Procedure Train(number of training steps T,batch size B, training set S)
Input actor network $p_{\theta_p}(\pi|s)$ and critical network $V_{\theta_\nu}(s)$
1.Initialize actor network and critical network parameters θ_p, θ_ν
2. **For** $i \in [1, T]$ **do**
3. $s_1, s_2, ..., s_B \sim$ SampleInput(S)
4. $\pi_1, \pi_2, ..., \pi_B \sim$ SampleSolution($(p_{\theta_p}(.|s))$)
5. $b_1, b_2, ..., b_B \sim (V_{\theta_\nu}(s)$
6. Update actor network parameters θ_p:
$\nabla_{\theta_p} J \leftarrow \frac{1}{B}\sum_{i=1}^{B}(L(\pi_i|s_i) - b_{\theta_\nu}(s_i))\nabla_{\theta_p} log p_{\theta_p}(\pi_i|s_i)$
$\theta_p \leftarrow Adam(\theta_p, \nabla_{\theta_p} J)$
7. Calculate the objective function value of the critical network:
$\mathrm{L}(\theta_\nu) \leftarrow \frac{1}{B}\sum_{i=1}^{B} ||b_{\theta_\nu}(s_i) - L(\pi_i|s_i)||_2^2$
8. Update critical network parameters θ_ν:
9. $\theta_\nu \leftarrow Adam(\theta_\nu, \nabla_{\theta_\nu} L)$
10.**End for**
11.**Return** θ_p, θ_ν

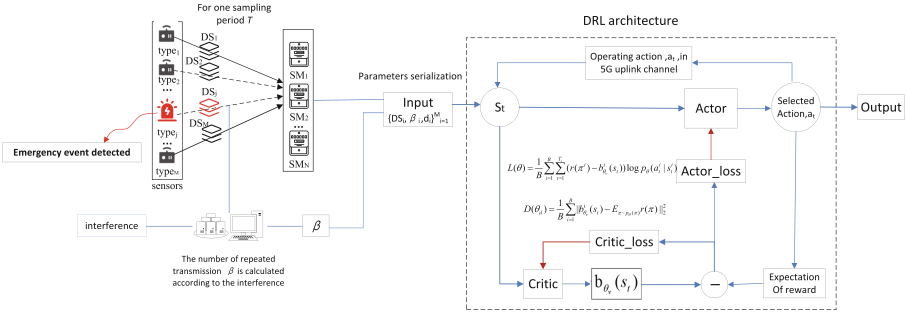

Fig. 5. Algorithm framework-based deep reinforcement learning.

In the case of interference, the number of repeated transmissions of emergency data β is determined by Formulas 1, 2 and 3 in Sect. 4.1. Finally, the number of repeated transmissions β, packet size DS, and distance between SM and gNB d are taken as the input of the pointer network in deep reinforcement learning, and the maximum energy efficiency is the objective function. Through the training of samples, the output of the pointer network is the result of resource allocation.

6 Simulation

6.1 Setup

In this section, we will evaluate the performance of the scheduling algorithm. An AMI application consists of massive SMs, a 5G communication network for information interaction with gNB, and an MDMS. Table 1 lists different configurations of the 5G network system bandwidth and subcarrier interval [17]. We set the system bandwidth and subcarrier interval to 20 MHz and 15 kHz, respectively. Accordingly, the number of subchannels Y is 106, and the duration of each time slot is 1 ms. Detailed parameters for simulation are listed in Table 2. In addition, there are 16 MCS options in the channel, and each MCS corresponds to different TBS and SINR ranges [25], as shown in Table 3.

Table 1. Transmission bandwidth configuration and time slot duration

Time slot duration	SCS (kHz)	10 MHz	15 MHz	20 MHz	...	100 MHz
		NRB	NRB	NRB	...	NRB
1 ms	15	52	79	106	...	N/A
0.5 ms	30	24	38	51	...	273
0.25 ms	60	11	18	24	...	135

Table 2. Simulation parameters

Parameters	Values
Number of packets, M	50–500
Network topology	Random deployment
Maximum transmission power, P_{max}	23 dBm
Path loss model, PL_N	$13.6 + 35 \times \lg d(dB)$
Sampling period of sensors, T_k	100–500 ms
Reference distance from SM to gNB, d_0	50 m
Actual distance from SM to gNB, d	50–500 m
Sampling data size, DK	300–500 bytes
Observation period, T	500 ms
SNR threshold	40 dB
σ	2
Outage probability threshold, a	0.01
Path loss exponent, λ	2
Thermal noise variance, N_0	1 dB

We assume that all SMs can detect emergency events in ADNs, and the events detected depend on the type of sensor installed, such as the voltage sag that can be detected by the voltage sensor. Combined with the collected data samples, the method in [9] is used to calculate the number of emergency data in a unit time slot. Then, we reduce the outage probability of SMs by increasing the number of repeat transfers of emergency data; i.e., we determine the number of repeat transfers according to the threshold of outage probability.

In the experiment, we set the relevant parameters of the deep reinforcement learning model, and the batch size is 128; i.e., we use a small batch of 128 sequences, including 128 hidden LSTM units, and embed J parameters of each packet into 128 dimensional space. The initial learning rate is 10^{-4}, and the optimization reduction coefficient is 0.96 per 5000 steps. We use Adam optimizer to train the model.

6.2 Experiment Result

To evaluate our proposed algorithm, we compare the deep reinforcement learning algorithm with the optimal Lagrangian algorithm and greedy algorithm. The optimal Lagrange algorithm transforms the problem into the optimal programming problem in the mathematical model to obtain the global optimal solution. In theory, the greedy algorithm does not consider global optimization and only considers the local optimal selection. The results show that the calculation result of this method is close to the optimal value, which verifies the correctness of the algorithm and is obviously better than that of the greedy algorithm. We provide

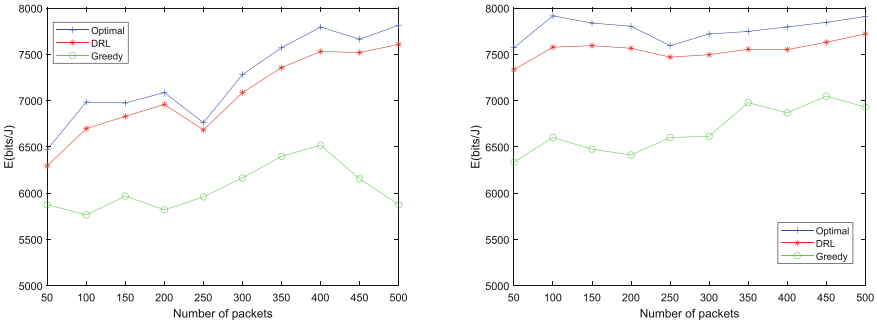

Fig. 6. Energy efficiency comparison of three algorithms. (a) N = 4; (b) N = 8.

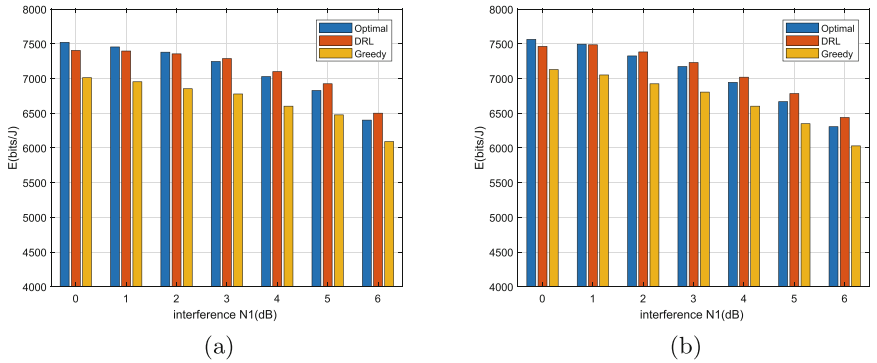

Fig. 7. Energy efficiency under interference. (a) N = 4, L = 300; (b) N = 8, L = 300.

the simulation results in terms of energy efficiency, scheduling time, and resource allocation ratio.

Table 4 compares the operation time between deep reinforcement learning and the traditional Lagrange multiplier method for the same scale problem. The method based on deep reinforcement learning has a much shorter running time than the optimal Lagrange multiplier method, which can be applied to real-time scenes.

Figure 6, we compare the experimental results of our proposed method and two other methods: the optimal Lagrange multiplier method and the greedy algorithm. The results show that the computational results of this method are close to optimal, which verifies the correctness of the algorithm and is obviously better than that of the greedy algorithm. In addition, with the increase in number of packets, the calculation results of the greedy algorithm significantly decrease, while the calculation results of the other two algorithms fluctuate; thus, the greedy algorithm has a worse effect than other algorithms. However, our deep reinforcement learning algorithm performs well at different scales.

Table 3. TBS and SINR range of each MCS

Index	Modulation	TBS (bits)	SINR range (dB)
0	QPSK	56	$(-9, -7]$
1	QPSK	72	$(-7, -5]$
2	QPSK	104	$(-5, -3]$
3	QPSK	120	$(-3, -1]$
4	QPSK	136	$(-1, 1]$
5	QPSK	144	$(1, 3]$
6	QPSK	208	$(3, 5]$
7	16QAM	280	$(5, 7]$
8	16QAM	336	$(7, 8.5]$
9	16QAM	408	$(8.5, 10]$
10	64QAM	440	$(10, 11.5]$
11	64QAM	488	$(11.5, 13.5]$
12	64QAM	520	$(13.5, 15]$
13	64QAM	552	$(15, 17]$
14	64QAM	584	$(17, 19.5]$
15	64QAM	616	$(19.5, \text{inf}]$

Table 4. Running time of different methods

M (Number of data packets)	DRL based	Lagrange
M = 50	0.19 s	15.08 s
M = 100	0.82 s	27.89 s
M = 200	2.03 s	182.64 s
M = 300	10.83 s	901.07 s

Figure 7 compares the energy efficiency of the optimal Lagrange multiplier method, deep reinforcement learning algorithm and greedy algorithm when the number of SMS is different and the number of packets is fixed. The results show that with the increase of interference, the energy efficiency of the three algorithms shows a downward trend, but the decline speed of deep reinforcement learning is lower than that of the other two algorithms. When the external interference $N1 > 3$, the energy efficiency of deep reinforcement learning algorithm is greater than that of the best Lagrangian algorithm, which shows that the deep reinforcement learning algorithm in this paper has strong anti-interference ability.

Figure 8 shows the target value changes of multi-round training for the same datasets under different batch training sizes. The results show that for a larger batch training size, the convergence of the pointer network is better, the training time is shorter, and the optimal value can be more quickly calculated.

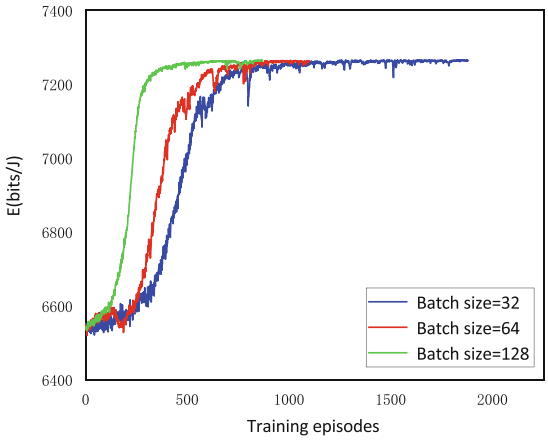

Fig. 8. Energy efficiency with different batch sizes.

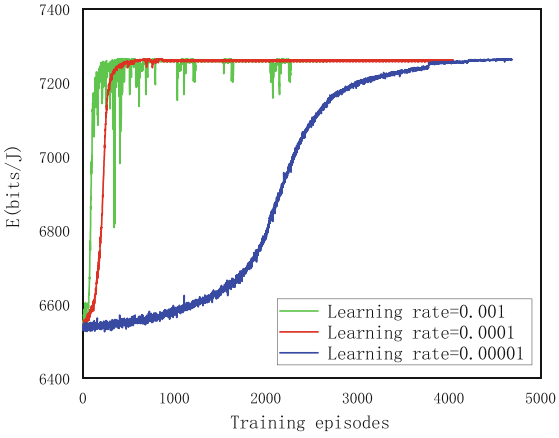

Fig. 9. Energy efficiency with different learning rates.

Figure 9 shows the change in training target values of the same data set under different learning rates. The results show that at a higher learning rate, the convergence is better, but the computational stability is worse. A smaller learning rate corresponds to better computational stability but a slower convergence rate.

7 Conclusion

This paper proposes a data repeat transmission scheme and uses a deep reinforcement learning method for resource allocation. To realize the SA of ADNs, data transmission must satisfy the requirements of low latency and high reliability. A framework based on 5G is proposed. First, by increasing the number of repeated

transmissions of emergency data generated by emergencies, the impact of external interference on the transmission interruption is eliminated to improve the transmission reliability. Then, with massive and real-time data, the 5G uplink resource allocation is modeled as a dynamic programming problem to maximize the energy efficiency, and the deep reinforcement learning method is used to solve the problem to improve the calculation speed and reduce the transmission delay. In addition, we compare the performance of the algorithm with other typical algorithms. The experimental results show that the algorithm can improve the calculation speed and provide nearly optimal transmission energy efficiency.

References

1. Bayram, I.S., Ustun, T.S.: A survey on behind the meter energy management systems in smart grid. Renew. Sustain. Energy Rev. **72**, 1208–1232 (2016)
2. Dong, Z., Xu, T., Li, Y., Feng, P., Gao, X., Zhang, X.: Review and application of situation awareness key technologies for smart grid. In: 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2), pp. 1–6 (2017). https://doi.org/10.1109/EI2.2017.8245450
3. Sun, Y., Chen, X., Yang, S., Tseng, K.J., Amaratunga, G.: Micro PMU based monitoring system for active distribution networks. In: 2017 IEEE 12th International Conference on Power Electronics and Drive Systems (PEDS) (2017)
4. Sesia, S., Toufik, I., Baker, M.: Introduction to LTE-Advanced, pp. 613–622 (2011)
5. Schulz, P., et al.: Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture. IEEE Commun. Mag. **55**(2), 70–78 (2017)
6. 5G; study on scenarios and requirements for next generation access technologies (V15.0.0); 3GPP TR 38.913 version 15.0.0 release 15
7. Pedersen, K.I., Khosravirad, S.R., Berardinelli, G., Frederiksen, F.: Rethink hybrid automatic repeat request design for 5g: five configurable enhancements. IEEE Wirel. Commun. **24**(6), 154–160 (2017)
8. Xiaotong, S., Nan, H., Naizheng, Z.: Study on system latency reduction based on shorten TTI. In: 2016 IEEE 13th International Conference on Signal Processing, ICSP (2016)
9. Li, Q., Tang, H., Sun, W., Li, W., Xu, X.: An optimal wireless resource allocation of machine-type communications in the 5g network for situation awareness of active distribution network. In: 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) (2020)
10. Kong, P.Y., Song, Y.: Joint consideration of communication network and power grid topology for communications in community smart grid. IEEE Trans. Industr. Inf. **16**(5), 2895–2905 (2020)
11. Zahoor, S., et al.: Cloud-fog-based smart grid model for efficient resource management. Sustainability **10**, 2079 (2018)
12. Al-Rubaye, S., Al-Dulaimi, A., Cosmas, J.: Spectrum allocation techniques for industrial smart grid infrastructure. In: IEEE International Conference on Industrial Informatics (2017)
13. Yaacoub, E., Kadri, A.: LTE radio resource management for real-time smart meter reading in the smart grid. In: IEEE International Conference on Communication Workshop, pp. 2000–2005 (2015)

14. Yang, T., Hu, Y., Gursoy, M.C., Schmeink, A., Mathar, R.: Deep reinforcement learning based resource allocation in low latency edge computing networks, pp. 1–5 (2018)
15. Ji, L., Hui, G., Lv, T., Lu, Y.: Deep reinforcement learning based computation offloading and resource allocation for MEC. In: 2018 IEEE Wireless Communications and Networking Conference (WCNC) (2018)
16. 5G; NR; physical channels and modulation (V16.1.0); 3GPP TS 38.211 version 16.1.0 release 16
17. 5G; NR; base station (BS) radio transmission and reception (V1.0.0); 3GPP TS 38.104 version 1.0.0 release 15
18. Albu, M., Sanduleac, M., Stanescu, C.: Syncretic use of smart meters for power quality monitoring in emerging networks. IEEE Trans. Smart Grid **8**, 485–492 (2016)
19. Park, J., Hwang, J.-N., Li, Q., Yiling, X., Huang, W.: Optimal dash-multicasting over LTE. IEEE Trans. Veh. Technol. **67**(5), 4487–4500 (2018)
20. Castellanos, CÚ., Villa, D.L., Rosa, C., Pedersen, K.I., Michel, J.: Performance of uplink fractional power control in UTRAN LTE. In: Vehicular Technology Conference (2008)
21. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Computer Science, 28 (2015)
22. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
23. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computer Science (2014)
24. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (1992)
25. 3GPP: physical layer procedures for data (release 16) (V16.1.0); 3GPP TS 38.214 version 16.1.0 release 16