Energy-Efficient Ultra-Dense Network using Deep Reinforcement Learning

Hyungyu Ju*, Seunghyun Kim†, YoungJoon Kim†, Hyojin Lee†, and Byonghyo Shim*

*Department of Electrical and Computer Engineering, INMC, Seoul National University, Seoul, Republic of Korea
†Advanced Communications Research Center, Samsung Research, Samsung Electronics, Seoul, Republic of Korea

Email: {hkjoo, snkim, bshim}@islab.snu.ac.kr, {yjoon100.kim, hyojin17.lee}@samsung.com

Abstract—With the explosive growth in mobile data traffic, pursuing energy efficiency has become one of key challenges for the next generation communication systems. In recent years, an approach to reduce the energy consumption of base stations (BSs) by selectively turning off the BSs, referred to as the sleep mode technique, has been suggested. However, due to the macro-cell oriented network operation and also computational overhead, this approach has not been so successful in the past. In this paper, we propose an approach to determine the BS active/sleep mode of ultra-dense network (UDN) using deep reinforcement learning (DRL). A key ingredient of the proposed scheme is to use action elimination network to reduce the wide action space (active/sleep mode selection). Numerical results show that the proposed scheme can significantly reduce the energy consumption of UDN while ensuring the QoS requirement of the network.

I. INTRODUCTION

Over the past years, exponential growth of mobile data traffic has triggered a dramatic increase in the energy consumption of cellular networks. Upsurge of energy consumption has now become a major bottleneck in achieving the primary goals, viz., 100-fold increase in the throughput in 5G and beyond. Since 5G operations is far more complicated than 4G LTE systems [1], energy consumption is a heavy burden in the operational expense (OPEX) for the network operators, not to mention the increased carbon emission and the acceleration of global warming. Pursuing balanced efficiency in energy consumption and throughput is, therefore, an inevitable option to ensure the sustainability of the next generation wireless communications [2].

Among several factors contributing to the energy consumption in cellular systems, by far the dominant source is the base station (BS). It has been reported that more than half of the network energy is consumed at BS [3]. To reduce the energy consumption at the BS, a technique that deliberately turns off lightly loaded BSs has been proposed over the years. Key premise of this approach, called the sleep mode energy saving technique [4], is to turn off underutilized BS to save unnecessary energy consumption caused by the power amplifier and RF/Analog circuitry. In spite of the rosy expectation, however, this approach has not been quite successful in the past due to the lack of BS cooperation mechanism (e.g., backhaul support) and slow deployment of the small cell technology.

Recently, this approach has regained attention with the introduction of the ultra-dense network (UDN), an approach to deploy a large number of small cells on top of macro cells (see Fig. 1) [5]. Since the problem to determine the active/sleep mode of BSs is a binary integer program (and hence NP-hard), it is in general very difficult to find out the optimal solution in the UDN environment. Therefore, to develop a technique that can effectively control the active/sleep mode of small cells in UDN while ensuring the quality of service (QoS) of mobile users is important for the success of UDN.

An aim of this paper is to put forth an entirely different approach to achieve energy efficiency in the UDN. Since the problem to control the network power can be well modeled as Markov decision process (MDP), we exploit the deep reinforcement learning (DRL) technique [6], an efficient tool to solve the sequential decision-making problem, to obtain efficient and adaptive control of the network energy consumption.

In DRL, an agent learns the optimal policy through the interactions with the environment. While the conventional RL cannot easily handle the large-scale control problems, DRL overcomes this limitation by replacing Q-table with the deep neural network (DNN). In recent years, we have witnessed the great success of DRL in various applications such as go game (AlphaGo in Google DeepMind) [7]. DRL has also been applied to various wireless systems such as spectrum access, traffic scheduling, and user association [8]. One potential shortcoming of the DRL-based wireless systems is that the action space of the large-scale network is immense...
to handle the complicated tasks. In particular, when deciding the active/sleep mode of BSs in UDN, the number of possible choices increases exponentially with the number of BSs (e.g., $2^{20} \approx 10^6$ actions when we consider 20 cells). To obtain the optimal policy, DRL agent should explore vast amount of action space, which is clearly prohibitive in the practical network operation.

In order to clear the hurdle, we introduce the notion of the action elimination which is used to eliminate the active/sleep mode decisions that cannot satisfy the QoS requirements (e.g., on/off decision causing too many sleep mode) [9]. By reducing the search space, we can speed up the Q-value search significantly. Further, by considering the transition overhead (on to off and vice versa) in the reward design of DRL, the energy efficiency of UDN can be dramatically improved. From the experimental results, we show that the proposed scheme outperforms the conventional method and basic Deep Q-Network (DQN) method, saving 61% and 21% power, respectively.

II. ULTRA-DENSE NETWORK

In this section, we discuss the system model of UDN and then power consumption model in UDN.

A. System Model

We consider the UDN environment where $M$ BSs cooperatively serve $K$ users. The BSs (a.k.a radio unit (RU) or remote radio head (RRH)) are connected to the DU to share the channel state information (CSI). In the sequel, we denote $m$ as the BS index and $k$ as the user index. In contrast to the conventional cellular networks where a single BS serves the entire users in one cell, a group of BSs cooperatively serves users in UDN [10]. In our work, we consider the fading channel model where the downlink channel vector $h_{m,k}$ from the $m$-th BS to the user $k$ is expressed as $h_{m,k} = \sqrt{\beta_{m,k}}g_{m,k}$ where $\beta_{m,k}$ is the large-scale fading coefficient and $g_{m,k} \sim \mathcal{CN}(0,1)$ is the small-scale fading coefficient. The transmit signal $x_m$ at the $m$-th BS is $x_m = \sum_{k=1}^{K} \sqrt{\gamma_{m,k}}s_k$, where $s_k$ is the data symbol and $\gamma_{m,k}$ is the transmit power from the $m$-th BS to the user $k$. Then, the received signal $y_k$ of the user $k$ is given by $y_k = \sum_{m=1}^{M} \sqrt{\gamma_{m,k}}h_{m,k}^*s_k + \sum_{j \neq k}^{K} \sum_{m=1}^{M} \sqrt{\gamma_{m,j}}h_{m,k}^*s_j + n_k$, where $n_k \sim \mathcal{CN}(0,\sigma_{n}^2)$ is the additive Gaussian noise. The corresponding rate of user $k$ is

$$R_k = \log_2 \left( 1 + \frac{\sum_{m=1}^{M} \gamma_{m,k}\beta_{m,k}}{\sum_{j \neq k}^{K} \sum_{m=1}^{M} \gamma_{m,j}\beta_{m,k} + \sigma_{n}^2} \right).$$

(1)

B. Power Consumption in UDN

In this subsection, we discuss the power consumption of UDN and then formulate the power minimization problem. Basically, the power consumption at the BS is divided into three parts: 1) transmission power $P_{\text{trans}}^m$ consumed by the radio transmission (i.e., power amplifier and RF circuitry), 2) active/sleep mode power $P_{\text{mode}}^m$ consumed by the baseband processing block, power supply, and air conditioning, and 3) transition power $P_{\text{trans}}^m$ consumed due to the mode transition (sleep to active mode and vice versa). We use $\alpha = [\alpha_1, \cdots, \alpha_M]^T$ to represent the active/sleep mode of the BSs. That is, $\alpha_m = 1$ and 0 indicate that the $m$-th BS is in active mode and sleep mode, respectively. Then, the transmit power of the $m$-th BS is

$$P_{\text{tx}}^m = \frac{\alpha_m}{\eta} \mathbb{E} |x_m|^2 = \frac{\alpha_m}{\eta} \sum_{k=1}^{K} \gamma_{m,k},$$

(2)

where $\eta$ is the power amplifier efficiency. The active/sleep mode power of the $m$-th BS is

$$P_{\text{mode}}^m = \alpha_m P_{\text{on}}^m + (1 - \alpha_m) P_{\text{off}}^m,$$

(3)

where $P_{\text{on}}^m$ and $P_{\text{off}}^m$ are power consumption of the $m$-th BS in active mode and sleep mode, respectively. The mode transition power of the $m$-th BS is expressed as

$$P_{\text{trans}}^m = |\alpha_m - \alpha_{\text{prev}}|P_{\text{trans}},$$

(4)

where $\alpha_{\text{prev}}$ is the active/sleep mode of $m$-th BS in the previous time slot and $\rho_{\text{trans}}$ is the power consumed by the mode transition. Thus, the total power consumption of the network $P_{\text{tot}}$ is given by $P_{\text{tot}} = \sum_{m=1}^{M} (P_{\text{tx}}^m + P_{\text{mode}}^m + P_{\text{trans}}^m)$. In this setting, our goal is to minimize the energy consumption $E_{\text{tot}} = \sum_{l=1}^{L} P_{\text{tot}}^l$ over the operational period of $L$ time slots (not the instantaneous power $P_{\text{tot}}^l$). Note that due to the time-varying nature of the channel and user demands, minimization of the instantaneous power will result in suboptimal active/sleep mode decision in the next time slot, limiting the energy saving of UDN.

The network energy minimization problem is formulated as

$$\mathcal{P}_1: \min_{\{\alpha_{\text{on}}^m,\alpha_{\text{off}}^m\}} \sum_{l=1}^{L} P_{\text{tot}}^l$$

s.t.  \hspace{1cm} R_k^l \geq R_{k_{\text{min}}}^l, \hspace{1cm} \forall k, \forall l$

$$P_{\text{on}}^m \leq P_{m_{\text{max}}}, \hspace{1cm} \forall m, \forall l$$

where $R_{k_{\text{min}}}^l$ is the rate constraint of the user $k$ at time slot $l$ and $P_{m_{\text{max}}}$ is the maximum transmit power of the $m$-th BS. Interestingly, this problem can be readily extended to energy harvesting problem.

III. BASICS OF REINFORCEMENT LEARNING

Reinforcement learning (RL) is a goal-oriented algorithm that learns how to achieve a goal using trials and errors. Key ingredients of RL are agent, environment, state, action, and reward. Let $S = \{s_1, s_2, \ldots, s_N\}$ be the state space and $A = \{a_1, a_2, \ldots, a_M\}$ be the action space. When the state $s_i \in S$ is given, an agent takes an action $a_k \in A$ and then the environment returns the next state $s_{i+1} \in S$ and the immediate reward $r_k$ to the agent. An optimal policy maximizing the expectation of cumulative reward is given by

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[ \sum_{l=0}^{\infty} \lambda^l r_{l+1} | \pi \right],$$

(6)

where $\lambda$ is a discount factor ($0 < \lambda < 1$) to reduce the weight to the future reward.
An aim of Q-learning is to find out an optimal policy maximizing the expected cumulative reward when taking action at state s. That is, \( Q^*(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a] \).

If the transition probability \( P_{sa} \) is given, the optimal Q-value function \( Q^*(s, a) \) can be obtained from the Bellman’s equation as
\[
Q^*(s, a) = r(s, a) + \lambda \sum_{s' \in S} \max_{a' \in A} Q^*(s', a')
\]
where \( r(s, a) \) is the reward corresponding to the current state \( s \) and action \( a \). Since it is not possible to acquire the transition probability in a priori, Q-function is updated recursively. That is, \( Q(s, a) = Q(s, a) + \epsilon[r(s, a) + \lambda \max_{a' \in A} Q(s', a') - Q(s, a)] \) (\( \epsilon \) is the learning rate).

One well-known problem of Q-learning is that the size of state space increases exponentially with the size of the task. When the state space is large, it is very difficult to update the Q-values of all possible state-action pair. To overcome this drawback, DQN has been proposed [6]. Key idea of DQN is to use DNN to approximate Q-function (i.e., \( Q^*(s, a) \approx Q(s, a, w) \)). The weight \( w \) of DQN is updated in a way to minimize the loss function \( L(w) \) given by
\[
L(w) = (Y_t^{DQN} - Q(s, a, w))^2
\]
where \( Y_t^{DQN} = r(s, a) + \lambda \max_{a' \in A} Q(s', a', w) \).

IV. DEEP REINFORCEMENT LEARNING BASED ENERGY-EFFICIENT ULTRA-DENSE NETWORK VIA ACTION ELIMINATION

Primary purpose of this work is to design the DRL-based UDN that consumes as small energy as possible. To make an efficient and quick decision of BSs in a network, we exclude the irrelevant active/sleep mode decisions from the action space. To this end, we exploit the action elimination, derived from the domain knowledge [9]. By eliminating the list of actions that would not necessarily be optimal, we can reduce the computational overhead of action space exploration significantly.

In a nutshell, overall learning process of the proposed AE-assisted DRL (AE-DRL) is as follows: 1) the DRL agent removes the irrelevant active/sleep mode decisions from the action space, 2) the agent determines the active/sleep mode of BSs, and 3) DU performs the transmit power allocation for the active BSs. 4) After that, DU computes the reward based on the total power consumption of the network and returns to the DRL agent (see Fig. 2). 5) Using this reward, the DRL agent adjusts the active/sleep strategy to maximize the cumulative reward.

A. Energy-Efficient UDN Model

In the proposed scheme, a group \( M \) BSs is used as an environment and DU is used as an agent.

1) State Space: we denote the channel state information \( M_t \) and the rate constraints \( R_t \) of users at time slot \( t \) as
\[
M_t = \begin{bmatrix} \beta_{t,1}(t) & \ldots & \beta_{t,M}(t) \end{bmatrix}, R_t = \begin{bmatrix} R_{t,1,min}^t \\ \vdots \\ R_{t,K,min}^t \end{bmatrix}
\]
The state \( s_t \) accounting for the channel characteristics and rate constraints is defined as \( s_t = [M_t M_{t-1} R_t]^T \).

2) Action Space: an action \( a_t \) is defined as
\[
a_t = [\alpha_{1}(t) \ldots \alpha_{M}(t)], \quad \alpha_{m}(t) \in \{0, 1\},
\]
where \( \alpha_{m}(t) = 1 \) (or \( \alpha_{m}(t) = 0 \)) indicates that \( m \)-th BS is in active mode (or sleep mode).

3) Reward: when the action \( a_t \) is decided, one can measure the consumed power of the network. Since the reward should be high when the power consumption is low, we have
\[
r_t = P_{max} - P_{tot} = P_{max} - \sum_{m=1}^{M} (P_{m}^{tx} + P_{m}^{mode} + P_{m}^{trans}),
\]
where \( P_{max} \) is the maximum total power consumption when all BSs are turned on. Since all three components of \( P_{tot} \) (\( P_{tx}, P_{mode}, \) and \( P_{trans} \)) are functions of the active/sleep mode of BSs, maximizing the reward is equivalent to the decision of the set of BSs minimizing \( P_{tot} \).

B. Action Elimination Network

We say the active/sleep mode decision is irrelevant if the allocated transmit power of the active BSs cannot satisfy the rate constraints. The elimination indicator is defined as
\[
e_t(s_t, a_t) = \begin{cases} 1 & \text{if the rate constraints are satisfied} \\ 0 & \text{otherwise} \end{cases}
\]

Using the channel state information (CSI) and the rate constraints \( s_t \) as input, AEN estimates the elimination indicator for each action. If the estimated elimination indicator is smaller than the pre-specified threshold (e.g., \( \hat{e}_t(s_t, a_t, \theta) < 0.5 \)), the corresponding active/sleep mode is excluded from the action space in the current time slot (see Fig. 3 on the next page). Since DQN explores only the reduced action space \( A' \), chance of finding out the optimal active/sleep mode decision increases substantially. After the agent performs the action \( a_t \), it receives the true elimination indicator \( e_t(s_t, a_t) \) of the corresponding action from the environment. The weight \( \theta \) of AEN is trained in a way to minimize the squared loss between \( e_t \) and \( \hat{e}_t \):
\[
L(\theta) = (e_t(s_t, a_t) - \hat{e}_t(s_t, a_t, \theta))^2.
\]

1However, if the active BSs cannot serve the users, we impose a strong penalty (e.g. \( r_t = -1000 \)) to the corresponding action.
C. DRL Training

Using the reduced action space $A'$, the agent determines the active/sleep mode of BSs using DQN (see Fig. 3). In this process, we use the $\epsilon$-greedy rule to choose an action to balance the exploration and exploitation. As a result of the action performed, the agent receives the immediate reward $r_t$, the elimination indicator $e_t$, and the next state $s_{t+1}$ from the UDN environment. In each time slot, the transition tuple $(s_t, a_t, r_t, e_t, s_{t+1})$ observed by the agent is stored to the replay memory. Using this transition data, the weights of DQN and AEN are updated in a direction to minimize the loss function (7) and (8).

D. Transmission Power Allocation

When the active/sleep modes $\{\alpha_m\}_{m=1}^M$ of all BSs are determined by AE-DRL, DU allocates the transmit power of the active BSs. While minimizing the transmit power $\{\gamma_{m,k}\}$, we also need to satisfy the rate constraint and the transmit power constraint. Let $A_m$ be the set of active BSs. Then, the transmit power allocation problem can be formulated as

$$
P_2 : \min_{\{\gamma_{m,k}\}} \sum_{m \in A_m} P_{m,\text{trans}} \quad \text{s.t.} \quad R_k \geq R_{k,\text{min}}, \quad \forall k = 1, \ldots, K$$

$$
P_{m,\text{trans}} \leq P_{m,\text{max}}, \quad \forall m = 1, \ldots, M. \quad \text{(9c)}$$

Using the rate expression in (1) and the power consumption model in (2), $P_2$ can be re-expressed as

$$
P_2' : \min_{\{\gamma_{m,k}\}} \sum_{m \in A_m} \frac{K}{\eta} \gamma_{m,k} \quad \text{s.t.} \quad \sum_{m \in A_m} \beta_{m,k} \gamma_{m,k} \geq \sigma_{sh}^2 (2R_{k,\text{min}} - 1) \sum_{j \neq k \in A_m} \beta_{m,k} \gamma_{m,j} \eta \geq \sigma_{sh}^2 (2R_{k,\text{min}} - 1), \quad k = 1, \ldots, K$$

$$
\alpha_m \sum_{k=1}^K \gamma_{m,k} \leq P_{m,\text{max}}, \quad m = 1, \ldots, M. \quad \text{(10c)}$$

The modified problem $P_2'$ is a linear program (LP) and thus can be solved by the convex optimization tool (e.g., CVX).

V. SIMULATION RESULTS

A. Simulation Setup

In this section, we present numerical results to evaluate the energy efficiency of the proposed DRL-based UDN. In our simulations, we consider the UDN environment where one macro BS and 9 micro BSs serve the users in a rectangular area of size $0.5 \times 0.5$ km$^2$ [11] and 4 users move freely at a constant speed ($3$ km/h). Note that the duration of time slot in the proposed scheme is set to 6 (secs) considering transition latency (deactivation + reactivation latency) of BS [12]. Both DQN and AEN consist of 5 fully connected layers and the width of a hidden layer is set to 256. The large-scale fading coefficient $\beta_{m,k}$ (Hata-COST231 model) is given by $\beta_{m,k} = PL_{m,k} \cdot 10^{-\frac{m,z_{m,k}}{10}}$, where $PL_{m,k}$ represents the path loss and $10^{-\frac{m,z_{m,k}}{10}}$ represents the shadow fading with $z_{m,k} \sim N(0,1)$. Specifically, $PL_{m,k}$ is given by

$$
PL_{m,k} = \begin{cases} 
-L - 35 \log_{10}(d_{mk}), & \text{if } d_{mk} > d_1 \\
-L - 15 \log_{10}(d_1) - 20 \log_{10}(d_{mk}), & \text{if } d_0 < d_{mk} \leq d_1 \\
-L - 15 \log_{10}(d_1) - 20 \log_{10}(d_0), & \text{if } d_{mk} \leq d_0
\end{cases}
$$

where

$$
L = 46.3 - 33.9 \log_{10} f - 13.82 \log_{10} h_b - (1.1 \log_{10} f - 0.7)h_u - (1.56 \log_{10} f - 0.8).
$$

We compare the proposed AE-DRL scheme with three baseline methods: 1) full association where all BSs are in
active mode, 2) conventional greedy algorithm which turns off the BS having the minimum impact on the energy consumption until it reaches to point where the user’s rate requirement is violated [13], and 3) vanilla DQN where the active/sleep mode of BSs is determined by using DQN [14].

B. Simulation Result

In Fig. 4, we plot the cumulative reward as a function of the number of iterations in the training phase. We observe that the proposed AE-DRL scheme has far better cumulative reward than the vanilla DQN-based scheme on average. These results demonstrate that the proposed AE-DRL scheme can make the energy-efficient active/sleep mode decision.

In Fig. 5, we evaluate the average power consumption as a function of user rate demand. We observe that the proposed AE-DRL scheme saves 61% power over the conventional greedy algorithm and 21% power over the vanilla DQN-based scheme on average. The reason behind this results is that the proposed scheme can make the optimal active/sleep mode decision from a long-term perspective by accounting for the mode transition power as well as active/sleep mode power. In particular, the proposed AE-DRL scheme is more energy-efficient than vanilla DQN when the user demand is high. This is because AE-DRL scheme successfully removes the irrelevant actions that cannot satisfy the user rate demand.

VI. CONCLUSION

In this paper, we proposed the DRL-based energy-efficient ultra-dense network framework. Specifically, we decided the active/sleep mode of the BSs using DQN equipped with the action elimination network. Then, we allocated the optimal transmission power of the active BSs using convex optimization method. By using the proposed AE-DRL scheme, DRL-based DU can effectively make a sequential decision minimizing long-term energy consumption. From numerical evaluation, we demonstrated that the proposed scheme provides significant energy saving in UDN.

REFERENCES