Digital Communications and Networks 9 (2023) 1023–1031

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

Traffic prediction enabled dynamic access points switching for energy saving in dense networks

Yuchao Zhu, Shaowei Wang *

School of Electronic Science and Engineering, Nanjing University, Nanjing, 210000, China

ARTICLE INFO

Keywords:
Access points switching on/off
Energy-saving
Green network
Long short-term memory
Traffic prediction

ABSTRACT

To meet the ever-increasing traffic demand and enhance the coverage of cellular networks, network densification is one of the crucial paradigms of 5G and beyond mobile networks, which can improve system capacity by deploying a large number of Access Points (APs) in the service area. However, since the energy consumption of APs generally accounts for a substantial part of the communication system, how to deal with the consequent energy issue is a challenging task for a mobile network with densely deployed APs. In this paper, we propose an intelligent AP switching on/off scheme to reduce the system energy consumption with the prerequisite of guaranteeing the quality of service, where the signaling overhead is also taken into consideration to ensure the stability of the network. First, based on historical traffic data, a long short-term memory method is introduced to predict the future traffic distribution, by which we can roughly determine when the AP switching operation should be triggered; second, we present an efficient three-step AP selection strategy to determine which of the APs would be switched on or off; third, an AP switching scheme with a threshold is proposed to adjust the switching frequency so as to improve the stability of the system. Experiment results indicate that our proposed traffic forecasting method performs well in practical scenarios, where the normalized root mean square error is within 10%. Furthermore, the achieved energy-saving is more than 28% on average with a reasonable outage probability and switching frequency for an area served by 40 APs in a commercial mobile network.

1. Introduction

The vision of 5G and beyond mobile network lies in improving the system performance with better coverage and higher data rates [1]. Recently, cellular traffic still tends to keep a continuous 1000 × increase due to the popularity of mobile terminals, which poses a challenge to the access network. Network capacity upgrade is a critical factor to satisfy the explosive data rate requirement. The densely deployed network, such as the Heterogeneous Network (HetNet) and Cloud Radio Access Network (C-RAN), is defined as the network consisting of a large number of small-scale Base Stations (BSs) or Access Points (APs), and is deemed as a promising architecture to enhance the system capacity [2]. The HetNet is identified as a mixture of both macro cells and different kinds of small cells to provide seamless coverage and high-speed transmissions [3]. The C-RAN enables efficient network deployment with low complexity by separating the base band signal processing procedures and the radio transceivers with the centralized base band units and remote radio heads [4]. On the one hand, network densification is regarded as a major trend in future mobile networks for meeting the ever-growing data demand and high spectral efficiency. On the other hand, the introduction of densely deployed network also increases the energy consumption of the network significantly since APs in the active mode account for about 60%–80% of the network energy consumption [5], thereby the greenhouse gas (CO_2) emissions will be aggravated. Energy efficiency has long been a major problem of great interest in the mobile communication system. Reducing energy consumption while ensuring various Key Performance Indicators (KPIs) has become a vital design goal [6,7].

Quality of Service (QoS) is one of the most important KPIs for mobile service providers, the prerequisite of which is to ensure the transmission rates of users at any time. The commercial mobile networks are usually designed with a certain level of redundancy to guarantee the peak rate demand of the users served by the system, i.e., the number of APs deployed in the mobile network depends solely on the statistical peak traffic distribution in the service area [8]. From another viewpoint, the population migration in the service area leads to a fluctuating traffic load distribution for a given mobile network [9]. As illustrated in Fig. 1, which illustrates the 24h traffic demand in a week, people's day and night living behavior would generate periodic fluctuations in the traffic distribution.

* Corresponding author.
E-mail addresses: dz20230030@smail.nju.edu.cn (Y. Zhu), wangsw@nju.edu.cn (S. Wang).

https://doi.org/10.1016/j.dcan.2022.05.017
Received 3 March 2021; Received in revised form 17 May 2022; Accepted 20 May 2022
Available online 28 May 2022

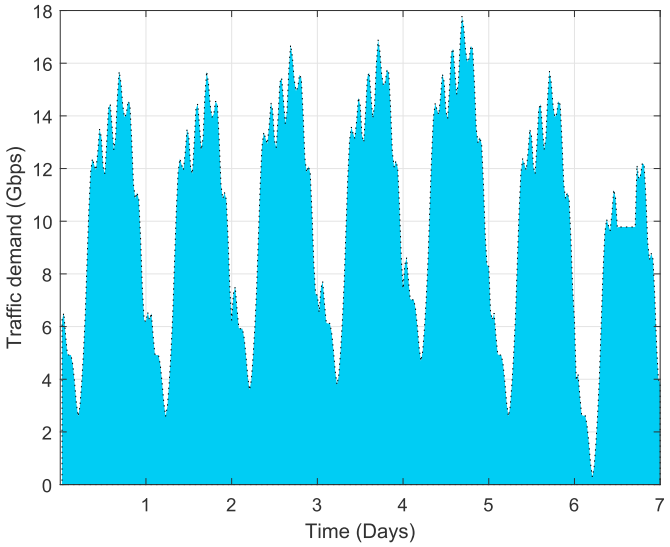


Fig. 1. An instance of the temporal distribution of cellular traffic demand during a week in a central business district.

If the APs in such areas are always active, a huge amount of energy would be wasted since the active APs still consume sizable power even though they serve few users at night. Naturally, switching off APs with a light load is a reasonable approach to save energy for both economic and environmental considerations [10].

The potential of AP switching on/off for saving energy in dense coverage areas has been corroborated in [11] based on the dataset from an operational network, where results indicate that about 17% of APs have low traffic load in 50%–99% of the time. Meanwhile, the research in [12] suggests that the proper selection of APs to be activated is also important since around 60% of APs' power consumption scales with the traffic load.

The key question then comes down to when and which APs should be switched off. The AP switching scheme should correspond to the instantaneous traffic demand, which requires the knowledge of real-time rate requirements. Once the network can capture the near future traffic patterns, it can subsequently trigger AP switching operations without deteriorating the QoS. Accurate and robust traffic prediction plays an essential role in tracing the fluctuating traffic loads of APs [13]. For example, Auto-Regressive Integrated Moving Average (ARIMA) and its derivatives are common methods for time series analysis and prediction [14]. Recently, with the development of big data and computing power, Machine Learning (ML) techniques have become feasible for network control and can be exploited in traffic prediction to achieve better performance [15]. For instance, support vector machines [16], Long Short-Term Memory (LSTM) network [17] and the convolutional neural network [18] are applied for traffic prediction.

This motivates us to investigate the dynamic AP switching on/off scheme based on traffic prediction. We aim to switch off a subset of APs at an appropriate interval to minimize energy consumption under the constraints of the network KPIs, such as the bandwidth budget, the spectral efficiency and especially, the rate requirement. It is noteworthy that the switching frequency is also an important performance indicator since frequent handovers will cause a decrease in QoS as well as yield an unbearable signaling overhead. The main contributions of this work can be summarized as follows:

- We introduce an intelligent LSTM network to deal with the APs' load prediction problem based on the raw data collected from the signal measuring instrument, so as to obtain the total traffic demand of the given service area, where the influence of prediction interval is also studied.
- We formulate a convex optimization problem of AP selection to minimize the transmission power with practical network constraints. A heuristic three-step local search algorithm is proposed to find promising solutions.
- We develop a dynamic AP switching on/off scheme to balance the QoS, the switching frequency and the energy consumption, where the AP selection procedure is triggered in consideration of the traffic trend and the current active AP set.

The rest of this paper is organized as follows. Related work is presented and discussed in Section 2. In Section 3, the network model along with the formulated optimization problem for AP selection are elaborated. The introduced LSTM network for traffic forecasting and the AP switching on/off algorithm are presented in Section 4. Experiment results are given with discussions in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

AP switching on/off, also known as BS sleeping, has been extensively investigated in the literature. In the broad sense, it is a special type of cell zooming, which adjusts the cell size to adapt to the cell load, service requirement and channel conditions, where zooming in the cell coverage to zero is equivalent to switching off the AP. In [19], both centralized and distributed cell zooming algorithms are investigated to achieve a trade-off between the energy consumption and the outage blocking probability. In [20,21], relay nodes are introduced to expand the coverage range of the BS so as to ensure the communication requirements of the users at the edge of the cell with a high energy efficiency. In [22], a distance-aware AP switching algorithm is presented based on the intuition that the transmission power increases with the distance between the user and the APs. In [23], a distributed AP on/off algorithm with the concern of traffic offloading is proposed. The AP that yields the minimal effect to the network would be switched off by introducing a network-impact parameter to maintain network performance. In [24], an effective local search AP selection algorithm is proposed to minimize the power consumption while satisfying practical network constraints in the C-RAN. In [25], an AP switching scheme with user association is designed to reduce the power consumption as well as balance the load among APs, where a trade-off between energy saving and network stability is achieved. In [26], a joint AP clustering and switching strategy based on the stochastic optimization theory is presented. By using the queue length information to capture the mismatch between the achievable data rate and the traffic demand, it can achieve up to 80% energy reduction while guaranteeing the network capacity.

It is worth noting that real-time traffic demand is also a key issue for AP switching on/off in practice. Studies [27–29] have shown the predictability of cellular network traffic, which endows the system to selectively turn off APs during the off-peak hours. In [27], the traffic is decomposed into regularity and randomness components, where the predictability of the regularity part is verified. In [28], the large periodic components and the small random components of traffic are handled by Fourier analysis and the LSTM, respectively. And Gaussian process regression is adopted to recover the residual components. In [29], the authors jointly consider the spatio-temporal modeling and prediction of cellular traffic, which can achieve satisfying accuracy. Generally, the design of AP switching on/off scheme involves precisely prefiguring the traffic load in the service area at a certain time. In [30], the authors propose a linear regression and naive forecasting combined with the hybrid traffic prediction model to make the forecasting as accurate as possible. The APs are also sorted according to their coverage to make the turning-off order more reasonable. Results show that up to 48% of APs can be switched off without impacting the QoS. Different from [30], which just considers statistical models due to the concern of complexity, neural networks for traffic forecasting are investigated in [31], based on

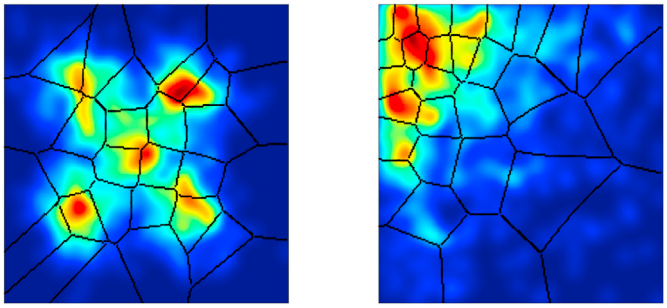


Fig. 2. Illustration of TDA: two instances of load balancing partition to find TDAs with approximately equal traffic demand.

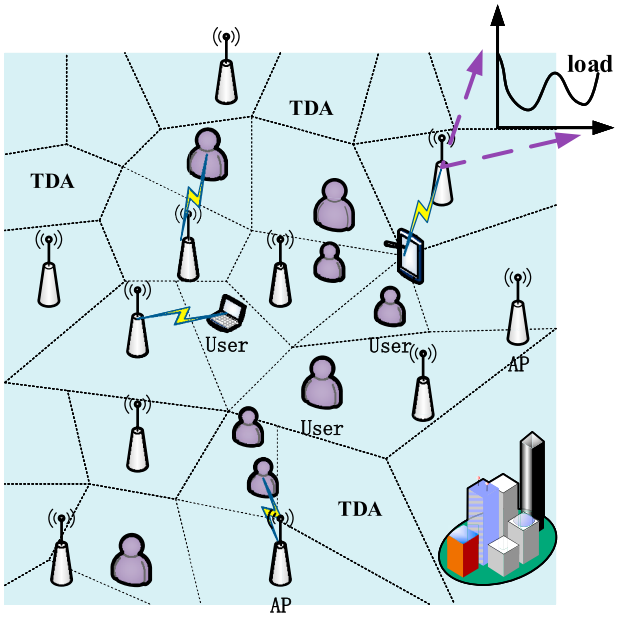


Fig. 3. The network model.

which a small cell AP would be switched off if its predicted load is under a threshold, and its load would be migrated to a macro cell AP. It shows that the energy consumption can be reduced while not causing obvious coverage loss in HetNets. In [32], a *K*-means clustering algorithm is used to divide the APs into different categories so that models only need to be trained for different clusters. Subsequently a classification method is applied to predict the states of APs according to the traffic load. The proposed mechanism can forecast the idle periods in advance so as to switch off some APs while guaranteeing the QoS. However, all these studies pay little attention to the switching frequency. It is still of significance to match the AP switching scheme with the fluctuating traffic under an acceptable signaling overhead. In this paper, we study the joint traffic prediction and dynamic AP switching on/off problem to achieve a trade-off between energy saving and network stability, including switching frequency and network KPIs.

3. Network model and problem formulation

3.1. Traffic model

We consider a square region $\mathcal{R} \in \mathbb{R}^2$ served by the densely deployed network, including N APs and a large number of user nodes. To simplify the formulation later, Traffic Demand Areas (TDAs) are introduced to abstractly represent the traffic distribution. It is considered that each TDA contains multiple user nodes with different rate requirements, and the total traffic demand of TDAs can be seen as approximately equal by applying a load-balancing partition method proposed in [33], as illustrated in Fig. 2. Then our considered service area, where the traffic distribution is inhomogeneous, can be discretized into M TDAs with an equal traffic demand, as depicted in Fig. 3.

Denote $\mathcal{N} = \{1, 2, \ldots, N\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$ as the set of APs and TDAs, respectively. And the traffic load distributions of different APs are denoted by $\mathcal{L} = \{L_1, \ldots, L_N\}$, where L_n is the time-varying load of AP $n \in \mathcal{N}$.

Let $x \in \mathcal{R}$ be a location in the service region, the density of rate requirement is represented by function $\Psi(x)$. R_{total} and R_m represent the total traffic requirement in the area $\mathcal{R}$ and TDA $\mathcal{R}_m$, respectively. Then we have

$$\begin{aligned} \iint_{\mathcal{R}} \Psi(x)\mathrm{d}\sigma &= R_{total} \\ \iint_{\mathcal{R}_m} \Psi(x)\mathrm{d}\sigma &= R_m \approx R_{total} \Big/ M \end{aligned} \tag{1}$$

Denote $h_{m,n}$ as the channel gain between AP n and TDA m. $b_{m,n}$ and $p_{m,n}$ are represented as the bandwidth and the power that AP n allocated to TDA m, respectively. The transmission rate between AP n and TDA m can be calculated as follows:

$$r_{m,n} = b_{m,n}\log_2\left(1 + \frac{p_{m,n}|h_{m,n}|^2}{b_{m,n}(N_0 + I_{m,n})}\right) \tag{2}$$

where N_0 is the noise power, and $I_{m,n}$ represents the interference introduced by the active APs with unit bandwidth. And $I_{m,n}$ is set as 0 for simplicity since the inter-cell interference can be eliminated by signal processing methods.

Assume that $\mathcal{M}_n$ is the set of TDAs served by AP n, and the total rate requirement of $\mathcal{R}$ at time t can be calculated as

$$R^t_{total} = \sum_{m\in\mathcal{M}} R_m = \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}_n} r_{m,n} = \sum_{n\in\mathcal{N}} l_{n,t} \tag{3}$$

where $l_{n,t} \in L_n$ is denoted as the traffic load of AP n at time t.

3.2. Power model

The total power consumption of the dense network can be written as follows:

$$P_{total} = \sum_{n\in\mathcal{N}} P^n_{fix} + \sum_{n\in\mathcal{N}} P^n_{trans} \tag{4}$$

where P^n_{fix} and P^n_{trans} are the fixed power and the transmission power of AP n, respectively. And the fixed part P^n_{fix} influenced by the state of AP can be expressed as

$$P^n_{fix} = \begin{cases} P^a_f, & \text{AP } n \text{ is active} \\ P^s_f, & \text{AP } n \text{ is inactive} \end{cases} \tag{5}$$

Note that the AP still consumes some power to guarantee the wake up from the inactive mode, although it is much smaller compared with the power when the AP is active [10]. The saved power consumption when AP n is switched off can be written as $P_{save} = P^a_f - P^s_f$.

The on/off state of AP $n \in \mathcal{N}$ is defined by a binary variable z_n:

$$z_n = \begin{cases} 1, & \text{AP } n \text{ is active} \\ 0, & \text{AP } n \text{ is inactive} \end{cases} \tag{6}$$

Then the total power consumption can be transformed into

$$\begin{aligned} P_{total} &= \sum_{n\in\mathcal{N}}\left(P_f^a z_n + P_f^s(1-z_n)\right) + \sum_{n\in\mathcal{N}}\frac{z_n}{\eta_n}\sum_{m\in\mathcal{M}} p_{m,n} \\ &= \sum_{n\in\mathcal{N}}(P_f^a - P_f^s)z_n + \sum_{n\in\mathcal{N}}\frac{z_n}{\eta_n}\sum_{m\in\mathcal{M}} p_{m,n} + \sum_{n\in\mathcal{N}} P_f^s \\ &= \sum_{n\in\mathcal{N}} P_{save} z_n + \sum_{n\in\mathcal{N}}\frac{z_n}{\eta_n}\sum_{m\in\mathcal{M}} p_{m,n} + \sum_{n\in\mathcal{N}} P_f^s \end{aligned} \tag{7}$$

where η_n is the power amplifier efficiency factor, and the last term is a constant for variable z_n. Also note that, according to (2), $p_{m,n}$ can be written as

$$p_{m,n} = \frac{N_0 b_{m,n}}{|h_{m,n}|^2}\left(2^{r_{m,n}/b_{m,n}} - 1\right) \tag{8}$$

3.3. Problem formulation

Our goal is to dynamically select a set of APs to minimize the total power consumption of the network while guaranteeing the practical constraints. Define p_n^{max} and b_n^{max} as the maximum transmission power and the available bandwidth for AP $n \in \mathcal{N}$, respectively. To simplify the notations, we collect the variables z_n's, $b_{m,n}$'s and $p_{m,n}$'s into vectors $\vec{z}$, $\vec{b}$, and $\vec{p}$, respectively. And we define $\mathbf{Z} = \{\vec{z} \mid z_n \in \{0,1\}\}$. The problem can be mathematically formulated as follows:

$$\begin{aligned} \underset{\vec{z},\,\vec{b},\,\vec{p}}{\text{minimize}} \quad & z_n\left(\sum_{n\in\mathcal{N}} P_{save} + \sum_{n\in\mathcal{N}}\frac{1}{\eta_n}\sum_{m\in\mathcal{M}} p_{m,n}\right) \\ s.t. \quad & C_1: \sum_{m\in\mathcal{M}} p_{m,n} \le z_n p_n^{max}, \forall n \in \mathcal{N} \\ & C_2: \sum_{m\in\mathcal{M}} b_{m,n} \le z_n b_n^{max}, \forall n \in \mathcal{N} \\ & C_3: \sum_{n\in\mathcal{N}} r_{m,n} \ge R_m, \forall m \in \mathcal{M} \\ & C_4: p_{m,n} \ge \triangle_{m,n} b_{m,n}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \\ & C_5: \vec{z} \in \mathbf{Z}, \quad \vec{b} \in \mathbb{R}_+^{MN}, \quad \vec{p} \in \mathbb{R}_+^{MN} \end{aligned} \tag{9}$$

In (9), C_1 and C_2 represent the maximum transmission power constrains and the available bandwidth budgets for the selected APs. C_3 is one of the network performance guarantee, which ensures the rate requirement of each TDA. C_5 is intuitive. C_4 is another performance guarantee, which represents that the selected APs should meet the minimum spectral efficiency requirement of TDAs. To be specific, the spectral efficiency requirement of TDA m is defined as $S_{m,n} = \frac{r_{m,n}}{b_{m,n}} \ge S_m^{min}$, and we can rewrite it as follows according to (2):

$$\log_2\left(1 + \frac{p_{m,n}|h_{m,n}|^2}{b_{m,n}N_0}\right) \ge S_m^{min} \tag{10}$$

Let $\triangle_{m,n} = (2^{S_m^{min}} - 1)N_0/|h_{m,n}|^2$, we can get C_4.

As users keep moving between different areas, the distribution of traffic demand in the service area is changing. Solving the optimization problem defined by (9) requires knowledge of the future traffic demand on each TDA (i.e., R_m), which is the part that needs to be forecasted. According to (3), the total traffic demand of the region at time t is the aggregation of the load of all the APs, so that the problem of spatial and temporal flow distribution prediction can be transformed into a time series prediction problem for each single AP. We can use time series prediction methods, such as the ARIMA and the LSTM, to handle the problem after obtaining the historical loads of APs.

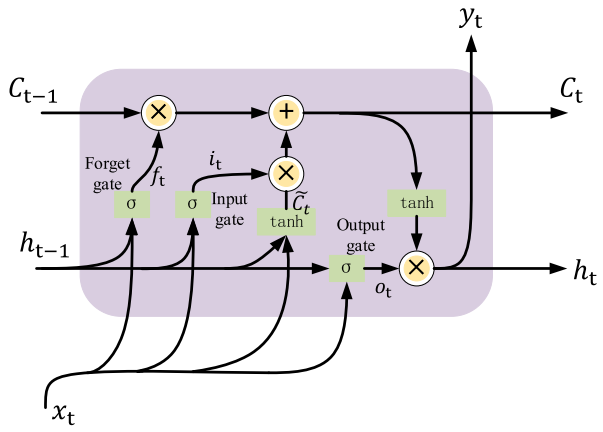


Fig. 4. Standard LSTM hidden unit.

As aforementioned, the collected data of APs throughout time interval T (in hour) is $\mathcal{L}$, where $L_n = \{l_{n,1}, l_{n,2}, \ldots, l_{n,S}\}$ stores the time-varying load of AP n. Let δ denote the granularity of the data. It is worth noting that the APs can not be switched at every timeslot δ when δ is relatively small (i.e., less than 20 min) because of the non-negligible operation time and signaling overhead [34]. Therefore, instead of predicting the traffic of the next timeslot, traffic values of the next x timeslots are predicted each time, then a sufficient interval $T_p = x\delta$ can be taken to be the switching interval as well as the desired prediction interval. And the total traffic demand of the region in duration $[t, t + T_p]$ is given by

$$R_{total} = \sum_{n\in\mathcal{N}} \frac{\sum_{j=t+1}^{j=t+x} l_{n,j}}{x} \tag{11}$$

where the average of predicted values is taken as the traffic demand of the service area at the next interval to achieve a trade-off between the switch frequency and the real-time traffic demand.

4. Our proposed algorithm

Eq. (9) is a mixed integer programming problem since it involves both binary variables z_n's and real variables $p_{m,n}$'s, $b_{m,n}$'s, which are NP-hard in general. In this paper, the LSTM network is introduced to handle the traffic prediction problem and offer a reference for choosing an appropriate time interval for AP switching. Then a local search algorithm for AP selection is proposed to deal with (9). Finally, a dynamic AP scheme with a judgment of load tendency is designed to determine when to trigger the AP selection procedure.

4.1. Cellular traffic prediction

Recurrent Neural Networks (RNNs) are commonly used for time series prediction. RNNs have chain-like structures to allow information to persist, thus forecasting future information with the knowledge of previous data. The LSTM network is a special type of RNN, where cell states and various gates are introduced to avoid the vanishing-gradient problem in RNNs.

As illustrated in Fig. 4, a standard LSTM unit has three gates interacting with each other to learn the long-term dependencies, namely, the input gate, the forget gate, and the output gate. The output of these gates can be written as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \otimes tanh(C_t) \end{aligned} \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function, $\otimes$ stands for the Hadamard product[1], W and b are the weight matrix and the bias of the gates, respectively.

Algorithm 1 Traffic prediction algorithm by LSTM

1: Initialization: $\mathcal{N}$, $\mathcal{L}$, δ, T, Num, x ($0<x<Num/2, x \in \mathbb{Z}$), $\mathcal{P} = \{P_1, P_2, ...P_N\}$, $P_i = 0$ ($i \in \{1, 2, ...N\}$).
2: Set the parameters of LSTM network;
3: **for** $n = 1 : N$; **do**
4: $T_x \leftarrow \{l_{n,1}, ..., l_{n,Num-x}\}$, $T_y \leftarrow \{l_{n,2}, ..., l_{n,Num-x+1}\}$;
5: $V_x \leftarrow \{l_{n,Num-x+1}, ..., l_{n,Num-1}\}$, $V_y \leftarrow \{l_{n,Num-x+2}, ...l_{n,Num}\}$;
6: INPUT: Training set $\mathcal{T}_n \leftarrow (T_x, T_y)$, Validation set $\mathcal{V}_n \leftarrow (V_x, V_y)$.
7: OUTPUT: $P_n = \{l_{n,t+1}, l_{n,t+2}, ..., l_{n,t+x}\}$, Calculate $NRMSE$ using (13).
8: **end for**
9: **return** R_{total} using (11).

As mentioned before, we obtained the T-hour traffic data of AP $n \in \mathcal{N}$ in a given service area $\mathcal{D}$ from the signal measuring instrument. It is denoted as L_n, whose granularity is δ (in hour, $0 < \delta < 1$). The number of obtained historical data is $Num = T/\delta$ for each AP $n \in \mathcal{N}$, and the data of all APs is stored in a dataset $\mathcal{L}$. For each AP n, we can divide the dataset into training sets and validation sets to train the LSTM network, and the prediction procedure is given in Algorithm 1. We employ the Normalized Root Mean Square Error (NRMSE) as the metric of accuracy, which is defined as

$$NRMSE = \frac{1}{y_{\max} - y_{\min}} \sqrt{\frac{1}{x} \sum_{i=1}^{x} y_i - y_i^{'2}} \tag{13}$$

where x is the total number of predicted points, y_i and $y_i^{'}$ are the observation value and prediction value at time i, respectively. $y_{\max}$ and $y_{\min}$ are the maximum and minimum of y, respectively.

4.2. Local search algorithm for minimizing power consumption

Recall (9), we can decompose this problem to two subproblems as follows for a given **Z**:

- Feasibility: Given a subset of APs, is it possible to meet the traffic demand of all the TDAs under the constraints of power and bandwidth budgets?
- Optimality: If possible, is this subset of APs the one that minimizes the total power consumption under practical network constraints?

These two questions can be summarized as *bandwidth and power allocation*.

First, we deal with the feasibility problem of *bandwidth and power allocation*, where a capacity margin is reserved for the prediction error and emergency situations. That is, under a given subset of APs, the traffic demand of TDA m, which is defined as $R_m^{'}$ (e.g., $R_m^{'} = 1.2R_m$), is supposed to be satisfied. It can be mathematically formulated as the following problem, where $\mathcal{N}_s = \{n \mid z_n = 1\}$ represents the given subset of APs:

$$\begin{aligned} & \text{find} \quad \overrightarrow{b}, \overrightarrow{p} \\ s.t. \quad & C_1: \sum_{m \in \mathcal{M}} p_{m,n} \le p_n^{max}, \forall n \in \mathcal{N}_s \\ & C_2: \sum_{m \in \mathcal{M}} b_{m,n} \le b_n^{max}, \forall n \in \mathcal{N}_s \\ & C_3: \sum_{n \in \mathcal{N}} r_{m,n} = R_m^{'}, \forall m \in \mathcal{M} \\ & C_4: \overrightarrow{b} \in \mathbb{R}_+^{MN}, \quad \overrightarrow{p} \in \mathbb{R}_+^{MN} \end{aligned} \tag{14}$$

It is equivalent to the following minimization problem:

$$\begin{aligned} & \underset{\overrightarrow{b}, \overrightarrow{p}, \xi}{\text{minimize}} \quad \xi^2 \\ s.t. \quad & C_1: \sum_{m \in \mathcal{M}} p_{m,n} \le p_n^{max}, \forall n \in \mathcal{N}_s \\ & C_2: \sum_{m \in \mathcal{M}} b_{m,n} \le b_n^{max} + \xi, \forall n \in \mathcal{N}_s \\ & C_3: \sum_{n \in \mathcal{N}_s} r_{m,n} = R_m^{'}, \forall m \in \mathcal{M} \\ & C_4: \overrightarrow{b} \in \mathbb{R}_+^{MN_s}, \quad \overrightarrow{p} \in \mathbb{R}_+^{MN_s} \end{aligned} \tag{15}$$

Eq. (15) is a standard convex optimization problem with a non-singular Hessian matrix, which can be solved by a fast barrier method [24]. The solution to (14) exists only if the optimal value of (15) is 0. If a feasible solution to (14) exists, we claim that the selected AP set $\mathcal{N}_s$ could meet TDAs' future rate requirements with given power and bandwidth budgets.

Then, if (14) is feasible, we try to optimize the *bandwidth and power allocation*, the problem of minimizing total power consumption can be further expressed as minimizing the transmission power under the selected subset of APs:

$$\begin{aligned} & \underset{\overrightarrow{b}, \overrightarrow{p}}{\text{minimize}} \quad \sum_{n \in \mathcal{N}} \frac{1}{\eta_n} \sum_{m \in \mathcal{M}} p_{m,n} \\ s.t. \quad & C_1, \ C_2 \text{ in } (14) \\ & C_3: \sum_{n \in \mathcal{N}_s} r_{m,n} \ge R_m, \forall m \in \mathcal{M} \\ & C_4: p_{m,n} \ge \triangle_{m,n} b_{m,n}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}_s \\ & C_5: \overrightarrow{b} \in \mathbb{R}_+^{MN}, \quad \overrightarrow{p} \in \mathbb{R}_+^{MN} \end{aligned} \tag{16}$$

Eq. (16) has a similar form with (15), which can also be solved by the fast barrier method. Denoting $P(\mathcal{N}_s)$ as the optimal solution to (16), we propose a heuristic three-step local search algorithm to perform the AP selection procedure. Starting with a feasible solution, such as $\mathcal{N}_s = \mathcal{N}$, we search the optimal subset of APs as follows:

Open: Switch on AP $n \notin \mathcal{N}_s$, update the power consumption, and if $P(\mathcal{N}_s \cup \{n\}) < P(\mathcal{N}_s)$, add AP n to $\mathcal{N}_s$: $\mathcal{N}_s \leftarrow \mathcal{N}_s \cup \{n\}$.

Close: Switch off AP $n \in \mathcal{N}_s$, update the power consumption, and if (14) is feasible and $P(\mathcal{N}_s \setminus \{n\}) < P(\mathcal{N}_s)$, remove AP n from $\mathcal{N}_s$: $\mathcal{N}_s \leftarrow \mathcal{N}_s \setminus \{n\}$.

Exchange: Activate AP $n^{'} \notin \mathcal{N}_s$, turn off AP $n \in \mathcal{N}_s$, update the power consumption, and if (14) is feasible and $P(\mathcal{N}_s \setminus \{n\} \cup \{n^{'}\}) < P(\mathcal{N}_s)$, make the exchange: $\mathcal{N}_s \leftarrow \mathcal{N}_s \setminus \{n\} \cup \{n^{'}\}$.

4.3. Dynamic AP switching on/off scheme

A dynamic AP switching on/off scheme is needed to find the optimal selection of APs at every switching interval while keeping the system signaling overhead as low as possible. Different from our prior work [35], where the AP selection procedure is triggered at each switching interval

[1] Element-wise multiplication for matrix.

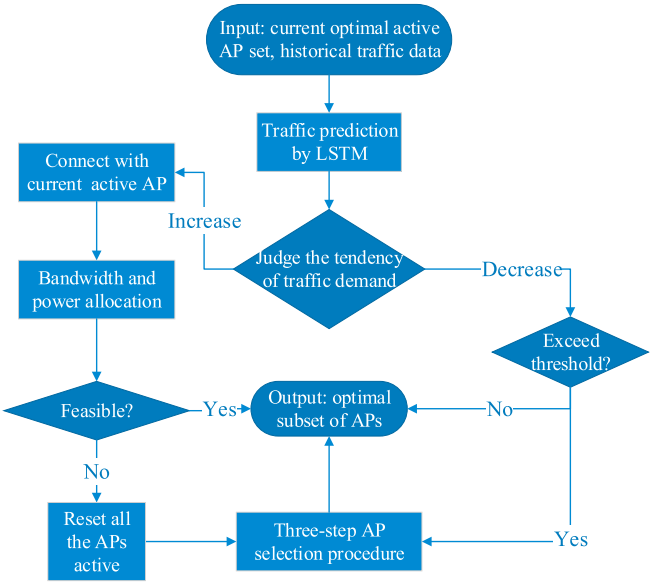


Fig. 5. Flowchart of traffic prediction enabled AP switching on/off mechanism.

even if current active APs could satisfy the demand, an improved AP switching scheme with a threshold is proposed to balance the switching frequency and power saving, the system stability of which is better than that in [35].

As depicted in Fig. 5, a judgment of the traffic tendency is introduced based on the predicted traffic distribution obtained by the LSTM. The AP selection procedure will be triggered when the decrease of traffic exceeds a given threshold γ or the bandwidth and power requirements can not be satisfied by current active APs. The decrease indicator is defined as $(R_{total}^{t-1} - R_{total}^{t})/R_{total}^{t-1}$, the range of which is [0, 1].

By taking the current set of active APs and the traffic tendency into consideration, the proposed switching scheme can avoid unnecessary switch operations, which makes the system more stable.

5. Experiment results

Experiments are conducted to evaluate the performance of our proposed scheme. Consider a commercial mobile network with 40 APs, whose sites are distributed uniformly in the service region. The number of divided TDAs is set as $K = 100$, and the spectral efficiency requirement of each TDA is 0.1 bit/Hz. For each AP, the maximum transmission power and available bandwidth are 1 W and 100 MHz, respectively. The efficiency of radio frequency power amplifier η_n is 25%. The power consumption of APs is 3.85 W and 0.75 W in the active and the inactive mode, respectively. The path loss (in dB) between AP and TDA can be calculated as $140.7 + 36.7\log_{10}(d)$, where d (in km) is the distance between the AP and the center of TDA. The noise PSD is -184 dBm/Hz, and the standard deviation of lognormal shadowing is 10 dB.

The chosen hyperparameters for the LSTM are given in Table 1, where the number of LSTM hidden layers is set as 2 with the concern of complexity and computational cost [36]. The learning rate is shrunk after 150 epochs by multiplying by a factor of 0.2 to avoid divergence. The Adam optimization is applied to update the network weights [37]. We use the collected 24-h mobile traffic data from the APs in the service area to forecast the traffic distribution in future timeslots. The granularity of the original historical data is 5 min. We later divide the training sets and the test sets according to the required prediction interval.

First, we investigate the performance of the LSTM network with different prediction intervals T_p. The prediction interval is chosen from [30, 60, 90, 120] min, i.e., the training sets are the same and the number of desired prediction timeslots is $x = [6,12,18,24]$, respectively. Fig. 6

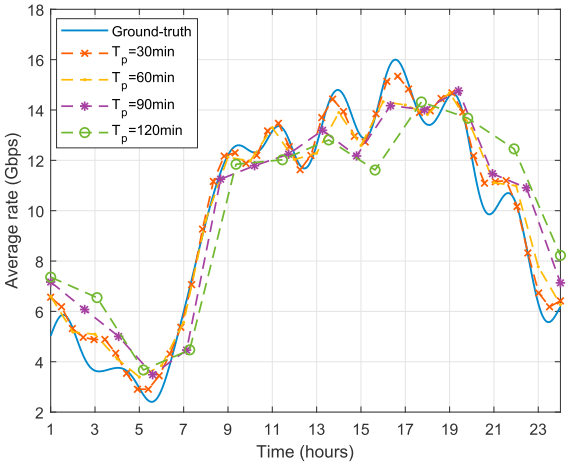


Fig. 6. Traffic prediction throughout a day with different T_p.

Table 1
Training hyperparameters for LSTM.

Parameter	Value
Maximum Num. of Iterations	200
Initial Learning Rate	0.005
Learning Rate Drop Factor	0.2
LSTM Hidden Units	128
LSTM Hidden Layers	2
Optimization Algorithm	Adam
Loss Function	NRMSE

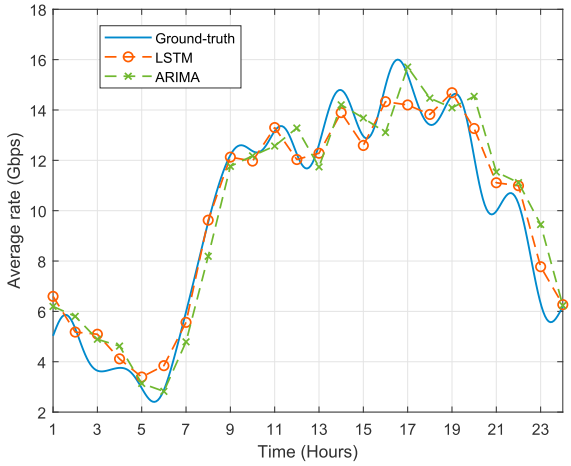
Table 2
The average NRMSE with different T_p. ($\delta = 5$ min).

T_p (min)	Average NRMSE
30	0.0185
60	0.0576
90	0.1154
120	0.1704

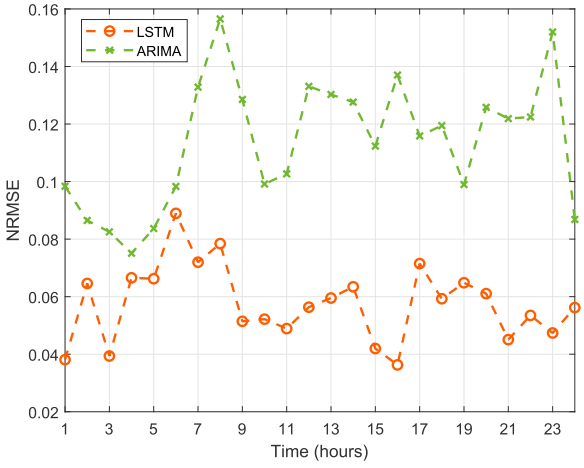
shows that the longer the prediction interval is, the lower the prediction accuracy will be. On the one hand, the precision will decrease due to the accumulative error, which is related to the inherent characteristic of the LSTM network [17]. On the other hand, since the average of the predicted values is used to represent the traffic demand in the future period, the data will be too sparse to depict the trend of traffic when the prediction interval is too large.

Table 2 shows more clearly that the accuracy is inversely proportional to the prediction interval. Although the accuracy of $T_p = 30$ min is the best, when considering the signaling overhead, we can conclude that the appropriate prediction interval is $T_p = 60$ min since it could balance the prediction error and the switching frequency.

Then, the performances of the ARIMA and the LSTM in the case of $T_p =$ 60 min are compared in Fig. 7. The ARIMA model is a widely used time series prediction method based on statistic models. It is defined by three parameters (p, d, q), which denote the auto-regressive term, differentiation term and moving average term, respectively. Here a (3, 1, 3) model is used. The performances of these two methods are roughly similar since the errors are weakened by the operation of taking the average of predicted values as the hourly traffic demand. Fig. 7(b) further shows the NRMSE of the two methods in detail, which suggests that the LSTM outperforms the ARIMA. The NRMSE of the LSTM is generally half of the ARIMA's. In addition, since the ARIMA model requires that the data is difference-stationary, the LSTM will be more suitable for cellular traffic prediction.



(a) Prediction result throughout a day



(b) Hourly NRMSE

Fig. 7. Comparison of ARIMA and LSTM, $T_p = 60$ min.

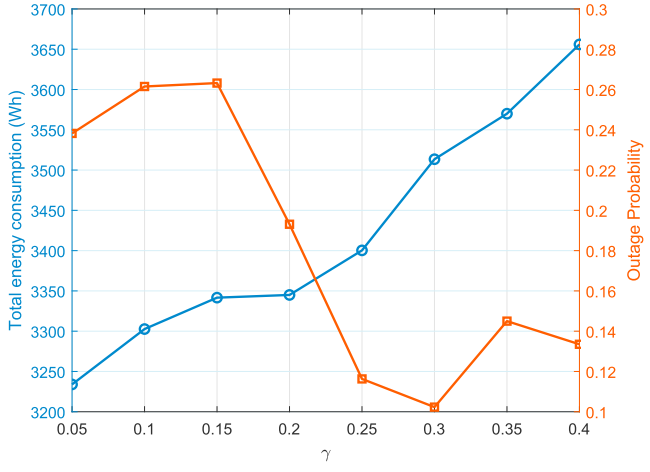


Fig. 8. The energy consumption and outage probability with different thresholds, $T_p = 60$ min.

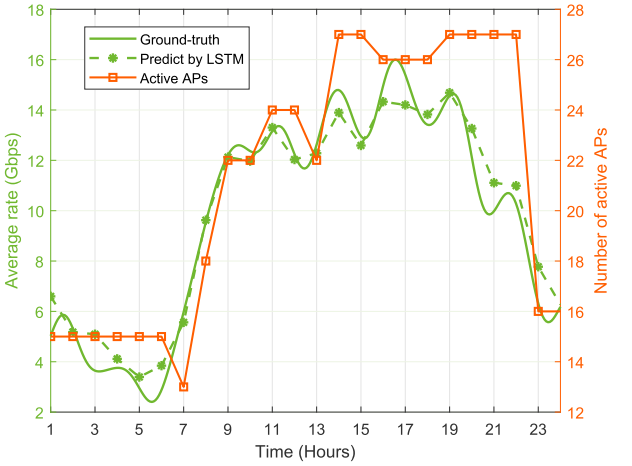
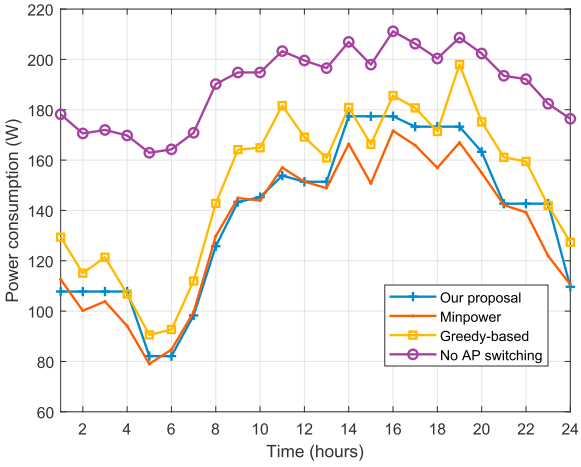
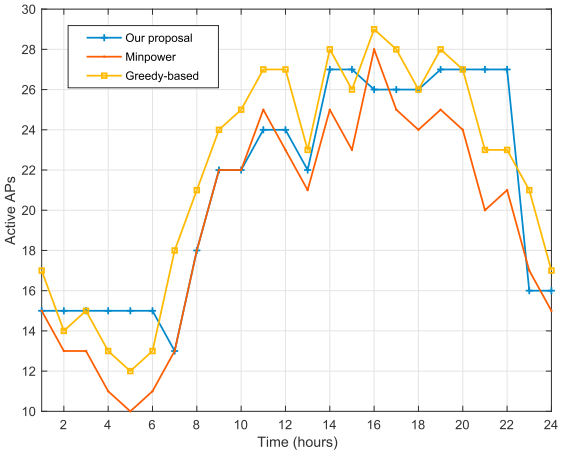


Fig. 9. The traffic distribution and number of active APs throughout a day, $\gamma = 0.25$, $T_p = 60$ min.



(a) Total power consumption.



(b) Number of active APs.

Fig. 10. Total power consumption and active number of APs throughout a day, $\gamma = 0.25$, $T_p = 60$ min.

Finally, the performance of our proposed dynamic traffic based AP switching scheme is investigated. The rough switching interval is set as 1 h as discussed above, i.e., we trigger the AP switching scheme at the beginning of every hour. The traffic demand margin is set as 50% to compensate for the prediction error. We compare our proposal with the following ones: no AP switching, greedy-based AP switching [38] and minimum power based AP switching (Minpower) [24]. For no AP switching, which is used as a benchmark, we only do bandwidth and power allocation to minimize the transmission power according to the real-time traffic demand in every hour. For the greedy-based method, all the APs are active at first, then the AP that yields the largest power saving would be switched off in each local search based on the traffic prediction result. For the Minpower, the AP selection procedure is similar to our proposal, while it does not consider the trend of traffic and the current active sets of APs. Outage probability is introduced to measure the system stability. Once the active APs can not satisfy the demand of TDAs, we regard it as an outage.

Fig. 8 shows the influence of different thresholds γ on our proposal. It can be seen that the energy consumption is proportional to γ and the outage probability is roughly inversely proportional to γ. Since the larger the threshold is, the lower the probability of performing the AP selection procedure is. We can conclude from Fig. 8 that it is reasonable and necessary to choose an appropriate threshold to meet the system requirements. γ is set as 0.25 to achieve a trade-off between the energy consumption and the outage probability.

Fig. 9 depicts the number of active APs by using our proposed algorithm along with the changing traffic demand throughout a day. The trend of number of APs required to be switched on approximately coincides with the traffic loads, and the selected AP set will not change when the fluctuation of traffic is within a narrow range. The switching frequency is 41.67%.

We also investigate the saved power throughout a day. As seen from Fig. 10(a), the power saving is significant compared with no AP switching, especially when the load is relatively low (e.g., 1:00–7:00), which corresponds with the expectation. Up to 50% power can be saved at 7:00 and at least 10% power saving can be achieved at 16:00. The number of active APs throughout a day is given in Fig. 10(b), which has the same trend with the power consumption. It can be observed from Fig. 10 that the performance of our proposal is close to that of Minpower and generally outperforms the greedy-based method. The result is reasonable since our proposal tends to maintain the status when the traffic changes insignificantly while the Minpower method still keeps searching for the optimal set of active APs all the time. And the greedy-based method does not find the optimal or minimum subset of APs each time, which still has redundancy and consumes more power. Table 3 gives the switching frequency of each scheme, the Minpower and the greedy-based algorithm will change the selected AP set at almost every switching time, while the switching frequency of our proposal is less than half of them.

In Fig. 11, the outage probabilities and the total energy consumption of different methods are shown. The energy consumption decreases 28% on average by our proposal compared with the no AP switching. Our proposal can achieve a much lower outage probability at the cost of 8.9% more energy consumption as compared to the Minpower. And 29.7% more energy can be saved as compared to the greedy-based method with a similar outage probability.

In conclusion, although the Minpower can find the optimal subset of APs with the minimum transmission power to satisfy the traffic demand, the switching frequency and outage probability of our proposal are much lower. And compared with the greedy-based algorithm, our proposal can obtain higher energy saving with a lower switching frequency and a similar outage probability.

Table 3
Switching frequency of different methods.

Method	Percentage (%)
Greedy-based	95.83
Minpower	100
Our proposal ($\gamma = 0.25$)	41.67

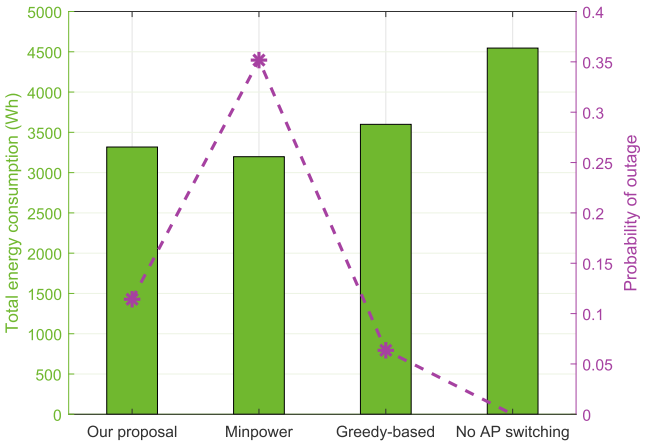


Fig. 11. Total energy consumption and average outage probability, $\gamma = 0.25$, $T_p = 60$ min.

6. Conclusions

In this paper, we studied the traffic load based AP switching problem for energy saving in dense networks, where we took the prediction interval, the switching frequency and the outage probability into account. Firstly, we transformed the complex spatial-temporal traffic prediction into time series prediction for APs. The problem was solved by an effective LSTM prediction network. Then, an "open/close/exchange" local search algorithm was introduced to implement the AP selection procedure, where bandwidth and power allocation was considered. Finally, an improved dynamic AP switching scheme based on a threshold was proposed to adjust the switching frequency and outage probability. Experiment results reveal that the LSTM network can give accurate forecasts for the traffic demand, based on which we can further implement AP switching with a trade-off between signaling overhead and real-time traffic demand. Numerical results validate that the proposed dynamic AP switching scheme can achieve a balance among energy consumption, outage probability and switching frequency. In general, our proposed scheme can save up to 50% energy when the traffic is relatively low throughout a day under our scenario, which is promising for building a green network.

Declaration of competing interest

No conflict of interest exits in the submission of this manuscript.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grants 61801208, 61931023, and U1936202.

References

[1] W. Saad, M. Bennis, M. Chen, A vision of 6G wireless systems: applications, trends, technologies, and open research problems, IEEE Netw 34 (3) (2020) 134–142.
[2] M. Kamel, W. Hamouda, A. Youssef, Ultra-dense networks: a survey, IEEE Commun. Surv. Tutorials 18 (4) (2016) 2522–2545.
[3] K.K. Mensah, et al., Energy efficiency based joint cell selection and power allocation scheme for HetNets, Digit. Commun. Netw. 2 (4) (2016) 184–190.
[4] B. Mahapatra, et al., CLB: a multilevel co-operative load balancing algorithm for C-RAN architecture, Digit. Commun. Netw. 5 (4) (2019) 308–316.
[5] C. Han, et al., Green radio: radio techniques to enable energy-efficient wireless networks, IEEE Commun. Mag. 49 (6) (2011) 46–54.
[6] S. Buzzi, et al., A survey of energy-efficient techniques for 5G networks and challenges ahead, IEEE J. Sel. Areas in Commun. 34 (4) (2016) 697–709.
[7] Q.D. La, et al., Enabling intelligence in fog computing to achieve energy and latency reduction, Digit. Commun. Netw. 5 (1) (2019) 3–9.
[8] M. Feng, S. Mao, T. Jiang, Base station on-off switching in 5G wireless networks: approaches and challenges, IEEE Wirel. Commun. 24 (4) (2017) 46–54.
[9] F. Xu, et al., Understanding mobile traffic patterns of large scale cellular towers in urban environment, IEEE/ACM Trans. Netw. 25 (2) (2017) 1147–1161.
[10] C. Liu, B. Natarajan, H. Xia, Small cell base station sleep strategies for energy efficiency, IEEE Trans. Veh. Technol. 65 (3) (2016) 1652–1661.
[11] Energy savings in mobile broadband network based on load predictions: opportunities and potentials, in: 2012 IEEE 75th Vehicular Technology Conference, VTC Spring, Yokohama, Japan, 2012, pp. 1–5.
[12] J. Wu, et al., Energy-efficient base-stations sleep-mode techniques in green cellular networks: a survey, IEEE Commun. Surv. Tutorials 17 (2) (2015) 803–826.
[13] X. Zhou, et al., The predictability of cellular networks traffic, in: 2012 International Symposium on Communications and Information Technologies (ISCIT), Gold Coast, QLD, 2012, pp. 973–978.
[14] J. Contreras, et al., ARIMA models to predict next-day electricity prices, IEEE Trans. Power Syst. 18 (3) (2003) 1014–1020.
[15] M. Elsayed, M. Erol-Kantarci, AI-enabled future wireless networks: challenges, opportunities, and open issues, IEEE Veh. Technol. Mag. 14 (3) (2019) 70–77.
[16] N.I. Sapankevych, R. Sankar, Time series prediction using support vector machines: a survey, IEEE Comput. Intell. Mag. 4 (2) (2009) 24–38.
[17] H.D. Trinh, L. Giupponi, P. Dini, Mobile traffic prediction from raw data using LSTM networks, in: 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Bologna, Italy, 2018, pp. 1827–1832.
[18] C. Zhang, et al., Citywide cellular traffic prediction based on densely connected convolutional neural networks, IEEE Commun. Lett. 22 (8) (2018) 1656–1659.
[19] Z. Niu, et al., Cell zooming for cost-efficient green cellular networks, IEEE Commun. Mag. 48 (11) (2010) 74–79.
[20] Y. Li, et al., Energy efficiency maximization by jointly optimizing the positions and serving range of relay stations in cellular networks, IEEE Trans. Veh. Technol. 64 (6) (2015) 2551–2560.
[21] Y. Li, et al., Energy-efficient optimal relay selection in cooperative cellular networks based on double auction, IEEE Trans. Wirel. Commun. 14 (8) (2015) 4093–4104.
[22] A. Bousia, et al., 'Green' distance-aware base station sleeping algorithm in LTE-Advanced, in: IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 2012, pp. 1347–1351.
[23] E. Oh, K. Son, B. Krishnamachari, Dynamic base station switching-on/off strategies for green cellular networks, IEEE Trans. Wirel. Commun. 12 (5) (2013) 2126–2136.
[24] W. Zhao, S. Wang, Traffic density-based RRH selection for power saving in C-RAN, IEEE J. Sel. Areas Commun. 34 (12) (2016) 3157–3167.
[25] X. Lin, S. Wang, Efficient remote radio head switching scheme in cloud radio access network: a load balancing perspective, in: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, GA, USA, 2017, pp. 1–9.
[26] J. Kim, H.W. Lee, S. Chong, Traffic-aware energy-saving base station sleeping and clustering in cooperative networks, IEEE Trans Wirel. Commun. 17 (2) (2018) 1173–1186.
[27] F. Xu, et al., Big data driven mobile traffic understanding and forecasting: a time series approach, IEEE Trans. Serv. Comput. 9 (5) (2016) 796–805.
[28] W. Wang, et al., Cellular traffic load prediction with LSTM and Gaussian process regression, in: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 2020, pp. 1–6.
[29] N. Zhao, et al., Spatial-temporal attention-convolution network for citywide cellular traffic prediction, IEEE Commun. Lett. 24 (11) (2020) 2532–2536.
[30] S. Dawoud, et al., Optimizing the power consumption of mobile networks based on traffic prediction, in: IEEE 38th Annual Computer Software and Applications Conference, Vasteras, Sweden, 2014, pp. 279–288.
[31] I. Donevski, G. Vallero, M.A. Marsan, Neural networks for cellular base station switching, in: EEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Paris, France, 2019, pp. 738–743.
[32] D. Sesto-Castilla, et al., Use of machine learning for energy efficiency in present and future mobile networks, in: IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 2019, pp. 1–6.
[33] C. Ran, S. Wang, C. Wang, Cellular networks planning: a workload balancing perspective, Comput. Netw. 84 (2015) 64–75.
[34] G. Rizzo, B. Rengarajan, M. Ajmone Marsan, The value of BS flexibility for QoS-aware sleep modes in cellular access networks, in: IEEE International Conference on Communications Workshops (ICC), Sydney, NSW, Australia, 2014, pp. 883–888.
[35] Y. Zhu, S. Wang, Joint Traffic Prediction and Base Station Sleeping for Energy Saving in Cellular Networks, in: ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1–6. Montreal, Canada.
[36] G. Vallero, et al., Greener RAN operation through machine learning, IEEE Trans. Netw. Service Manag. 16 (3) (2019) 896–908.
[37] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[38] Y. Shi, J. Zhang, K.B. Letaief, Group sparse beamforming for green Cloud-RAN, IEEE Trans. Wirel. Commun. 13 (5) (2014) 2809–2823.