Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8836315, 10 pages
https://doi.org/10.1155/2020/8836315

Research Article

Machine Learning-Based Resource Allocation Strategy for Network Slicing in Vehicular Networks

Yaping Cui,[1,2,3,4] **Xinyun Huang**,[1,3,4] **Dapeng Wu**,[1,3,4] **and Hao Zheng**[1,3,4]

[1]*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[2]*School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China*
[3]*Chongqing Key Laboratory of Optical Communication and Networks, Chongqing 400065, China*
[4]*Chongqing Key Laboratory of Ubiquitous Sensing and Networking, Chongqing 400065, China*

Correspondence should be addressed to Yaping Cui; cuiyp@cqupt.edu.cn

Received 27 August 2020; Revised 2 October 2020; Accepted 26 October 2020; Published 18 November 2020

Academic Editor: Changqing Luo

The diversified service requirements in vehicular networks have stimulated the investigation to develop suitable technologies to satisfy the demands of vehicles. In this context, network slicing has been considered as one of the most promising architectural techniques to cater to the various strict service requirements. However, the unpredictability of the service traffic of each slice caused by the complex communication environments leads to a weak utilization of the allocated slicing resources. Thus, in this paper, we use Long Short-Term Memory- (LSTM-) based resource allocation to reduce the total system delay. Specially, we first formulated the radio resource allocation problem as a convex optimization problem to minimize system delay. Secondly, to further reduce delay, we design a Convolutional LSTM- (ConvLSTM-) based traffic prediction to predict traffic of complex slice services in vehicular networks, which is used in the resource allocation processing. And three types of traffic are considered, that is, SMS, phone, and web traffic. Finally, based on the predicted results, i.e., the traffic of each slice and user load distribution, we exploit the primal-dual interior-point method to explore the optimal slice weight of resources. Numerical results show that the average error rates of predicted SMS, phone, and web traffic are 25.0%, 12.4%, and 12.2%, respectively, and the total delay is significantly reduced, which verifies the accuracy of the traffic prediction and the effectiveness of the proposed strategy.

1. Introduction

Autonomous driving is one of the key scenarios in 5G networks. In order to achieve road safety of intelligent transportation systems (ITS), the ultrareliable and low-latency communications (URLLC) must be guaranteed in vehicular networks. Moreover, for vehicle-to-everything (V2X) communications, the large data transmission of diversified service requirements poses the challenges to improve the transportation efficiency of ITS [1]. Therefore, it is urgent to tailor the vehicular networks to cater to these different requirements that come from different services. Along this line of thought, multiple virtual networks are created via network slicing as a feasible way to meet its diverse needs [2].

The main function of network slicing is creating multiple logical separate networks based on public shared physic infrastructure [3]. Specially, these logical networks are independent with each other. Through network slicing, mobile network operator (MNO) can allocate network resources dynamically and flexibly to each logical network slice on demand [4, 5] so as to support extensive use cases with various performance and service needs. Network slicing can be realized by software-defined network (SDN) and network function virtualization (NFV) technologies [6], and the future network will evolve into the flexible and programmable network architecture gradually [7]. One important issue in network slicing is the scheduling policy which allocates limited resources dynamically to vehicles with various quality of service (QoS) requirements according to the traffic change and network state. Most of the existing researches about network slicing focused on core networks, and the requirements (i.e., QoS) for RAN slicing are usually assumed to be

guaranteed perfectly. However, it may be unreasonable when the radio access network (RAN) is not considered due to the resource scheduling processing [8].

Through network slicing operation, multiple network slices will share the resources of the basic network according to their specific needs [9, 10]. The traffic characteristics of each network slice were analyzed and predicted in [11, 12], and the admission control decisions for network slice requests were then studied. In [13], the problem of QoS-aware joint admission control and network slicing was studied, and a heuristic algorithm was proposed to solve this problem. A model for orchestrating network slices based on service requirements and available resources was proposed in [14]. In addition, a Markov decision process framework was proposed to formulate and determine the optimal strategy of resource allocation for the 5G networks. In [15], a novel radio resource slicing framework for 5G networks was proposed; then, radio resources were allocated to different slices based on reinforcement learning. In [16], the slicing resource allocation problem was modeled as an online winner determination problem to maximize the social welfare of auction participants. Similarly, in [17], a novel auction mechanism-based network slicing strategy is presented in which resources and revenue were jointly optimized. In [18], the authors studied a simple dynamic resource sharing policy and indicate that the slices are able to maximize their carried loads subject to performance requirements by admission control to manage users' performance.

However, due to the characteristics of the vehicular networks, such as stringent delay constraints, and complicated communication environments [19], the general network slice resource allocation strategy may be insufficient in such scenario. Fortunately, under the complex and changeable vehicular network environment, machine learning would be an effective solution to this problem. For the dynamic nature of 5G vehicular scenarios, reference [20] proposed an online learning algorithm, namely fast machine learning, to solve the problem of beam selection to achieve higher context-awareness and adaptability in millimeter-wave vehicular communications. Furthermore, in [21], the authors discussed the prospect of managing vehicular network resources by reinforcement learning, and some open issues are also highlighted. A novel resource allocation scheme based on deep learning was proposed in [22]; the proposed scheme can optimize resource allocations according to the changing demands and network dynamics in SDN-based vehicular networks.

With the accumulation of cellular traffic data and the development of machine learning and artificial intelligence [23], the idea of traffic prediction based on machine learning is becoming more and more popular in the field of communications [24]. In [25], the authors designed a hybrid deep learning model for spatiotemporal prediction, where the spatial dependence was modeled by autoencoder and the temporal dependence was captured by LSTM [26]. In [27], a deep transfer learning-based prediction architecture has been proposed to predict different service traffic more conveniently. Moreover, in order to allocate resources appropriately, literatures [28, 29] used the LSTM to predict future traffic in RAN network slicing.

Motivated by the above analysis, we use machine learning to predict the services traffic of each slice, so as to allocate radio resources to each slice to reduce the delay. In more detail, firstly, we propose a new radio resource management, namely, shared proportion fairness (SPF), to keep resource management in accordance with slicing vehicle activity, and then, we use it for resource allocation representation. Moreover, we formulated the system delay minimization problem of resource allocation as a convex optimization problem. Secondly, we use ConvLSTM, which combines CNN and LSTM, to model the temporal-spatial dependency of the slice service traffic in the vehicular communication networks. Using the ConvLSTM for traffic prediction, we can predict different service traffic to obtain the user load distribution. Finally, according to the predicted results, a primal-dual interior-point-based resource allocation strategy is used to explore the optimal slice weight.

The rest of this paper is organized as follows. Section 2 describes the system model and assumptions. Using the above model, the resource allocation problem is formulated as a convex optimization problem to obtain optimal slice weight allocation in Section 3; then, an LSTM-based resource allocation is presented to minimize the system delay. In Section 4, we propose a primal-dual interior-point-based resource allocation strategy to solve the optimal slice weight problem. Simulation results are provided in Section 5 to evaluate the performance of the proposed traffic prediction and resource allocation strategy, followed by concluding remarks in Section 6.

2. System Model

We consider the cellular network consists of B Road Side Units (RSU) and V network slices. And the sets of RSUs and slices are denoted by $\mathscr{B}$ and $\mathscr{V}$, respectively. As shown in Figure 1, RSU is virtualized into three layers, namely RSU interface layer, RSU virtualization layer, and RSU virtual resource layer. Among them, the RSU interface layer provides related interface for each slice. The RSU virtualization layer implements the functions of slice management, SDN control, and slice coordination. Besides, the RSU virtual resource layer provides the virtual resources required by each slice, which are obtained from the resource sharing layer of the base station. Also, the base station controls the service traffic prediction of the entire system and schedules the resources among slices.

System states $\mathscr{U}_b^v$, $\mathscr{U}_b$, and $\mathscr{U}^v$ represent the sets of vehicles that communicate with RSU b on slice v, communicate with RSU b, and on slice v, respectively. Specially, we use n_b^v and n^vdenote the cardinalities of these sets, i.e., $|\mathscr{U}_b^v| = n_b^v$, and $|\mathscr{U}^v| = n^v$. Further, we assume that each vehicle communicates with only one RSU and connects to one slice.

In our model, the vehicle communicates with the RSU that provides it with the strongest SINR; thus, the downlink SINR can be expressed as

$$\mathrm{SINR}_{ub} = \frac{P_b G_{ub}}{\sum_{k\in\mathscr{B}} P_k G_{uk} + \sigma^2}, \tag{1}$$

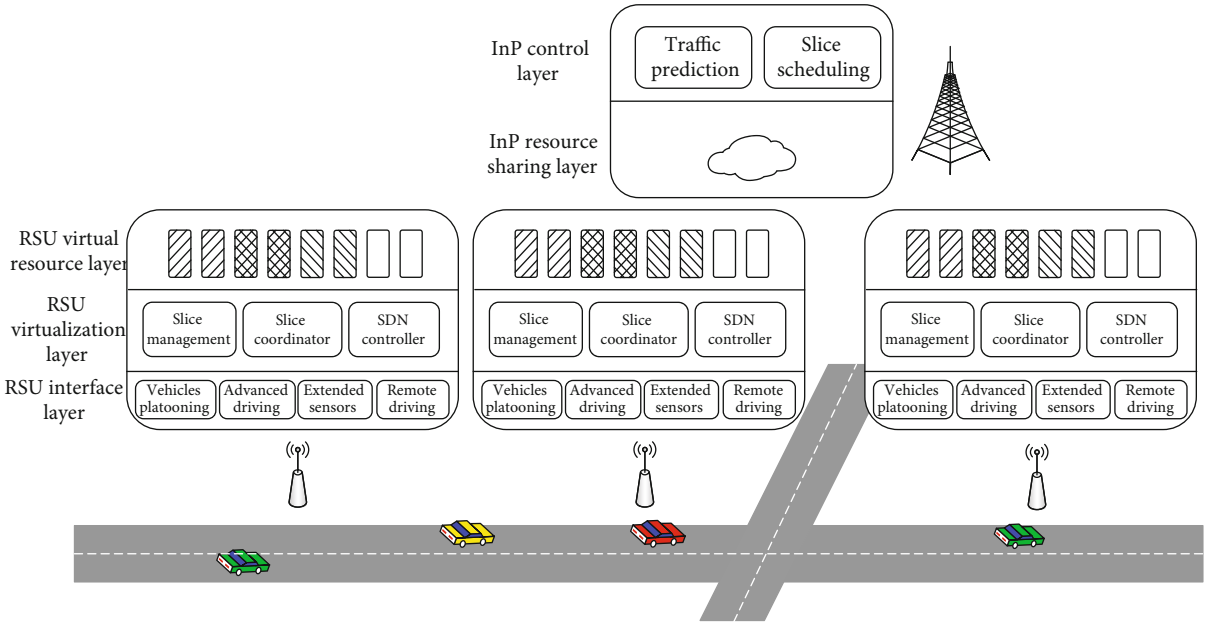

Figure 1: Network slicing framework in vehicular networks.

where the spectrum of the RSU is set to 20 MHz, and the transmit power P_b is set to 44 dBm [30]. The channel gain G_{ub} is related to path loss, shadow fading, and fast fading. The path loss is equals to $39\log_{10}(d_{ub}) + 25 + 20\log_{10}(f_c)$ [31], where d_{ub} and f_c represent the distance between the vehicle and the RSU and carrier frequency, respectively, and we set $f_c = 4$ GHz. Shadow fading follows logarithmic normal distribution with mean square deviation 4 dB. Fast fading is Rayleigh distribution depending on vehicles speed. The noise σ^2 depends on the noise spectral density $\eta = -174$ dBm/Hz and the noise figure $\gamma = 9$ dB.

According to the Shannon capacity formula, we get the spectrum efficiency of vehicle u at RSU b, which can be expressed as

$$e_{ub} = \log_2(1 + \mathrm{SINR}_{ub}). \tag{2}$$

The limited resources of each RSU are shared by all connected vehicles, so vehicle $u \in \mathcal{U}_b$ can be allocated a fraction of resources from RSU b. The allocated resources may be some resource blocks or time slots. For simplicity, we use $f_u \in [0, 1]$ that represents the allocated resources to vehicle u, which implies a proportion of total resources of RSU with $\sum_{u\in\mathcal{U}_b} f_u = 1$. Therefore, the transmission rate of RSU b to vehicle u is formulated as $r_u = f_u c_u$, where $c_u = Be_{ub}$ denotes transmission rate when all resources of RSU b are allocated to vehicle v, and B is bandwidth.

3. LSTM-Based Resource Allocation

In this section, we will describe the LSTM-based resource allocation. The shared proportional fairness resource allocation is first introduced, which can keep resource management in accordance with slicing vehicle activity. Then, the system delay minimization problem is formulated as a convex optimization problem. Finally, the ConvLSTM-based cellular traffic prediction is proposed at the end of this section.

3.1. Shared Proportional Fairness. For each slice in vehicular networks, we assume each slice is allocated a certain percentage of the radio resources, which is denoted by s^v, $v \in \mathcal{V}$, so we have $s^v > 0$, $\forall v \in \mathcal{V}$, and $\sum_{v\in\mathcal{V}} s^v = 1$. Next, the vehicle gets a subweight from the serving slice that depends on the number of active vehicles, i.e., for a vehicle $u \in \mathcal{U}^v$, $\forall v \in \mathcal{V}$, where $\omega_u = s^v/n^v$ means the subweight of vehicle u. Finally, The RSU allocates its resource to vehicles in proportion to their weights. Consequently, the transmission rate from RSU b to vehicle u can be obtained and written as

$$r_u = \frac{\omega_u}{\sum_{u'\in\mathcal{U}_b}\omega_{u'}} c_u = \frac{s^v/n^v}{\sum_{v'\in\mathcal{V}}\left(n_b^{v'} s^{v'}/n^{v'}\right)} c_u. \tag{3}$$

Considering there are many vehicles on slice v at RSU b so based on some further notations that are introduced in Table 1, the average transmission rate provided by RSU b to slice v is expressed as

$$r_b^v = \frac{s^v/\rho^v}{\tilde{g}_b} c_u. \tag{4}$$

According to (4), the average bit transmission delay (BTD) of the vehicle on slice v can be given by

$$\mathrm{BTD}^v = \sum_{b\in\mathcal{B}} \tilde{\rho}_b^v \mathrm{BTD}_b^v = \frac{\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}} \rangle_{\Delta_v}}{s^v}, \tag{5}$$

Table 1: Key notations.

Notation	Definition	Interpretation
ρ^v	n^v	Overall load of slice v
$\boldsymbol{\rho}^v$	$(\rho_b^v \triangleq n_b^v : b \in \mathscr{B})$	Load distribution of slice v
$\tilde{\boldsymbol{\rho}}^v$	$(\tilde{\rho}_b^v \triangleq \rho_b^v/\rho^v : b \in \mathscr{B})$	Relative load distribution of slice v
$\tilde{\boldsymbol{g}}$	$\left(\tilde{g}_b \triangleq \sum_{v\in V} s^v \tilde{\rho}_b^v : b \in \mathscr{B}\right)$	Overall weight relative load distribution
$\boldsymbol{\delta}^v$	$(\delta_b^v \triangleq \mathbb{E}[1/c_b^v]: b \in \mathscr{B})$	Mean reciprocal capacity of slice v
$\boldsymbol{\Delta}_v$	diag $(\boldsymbol{\delta}^v)$	Diagonal matrix of mean reciprocal capacity of slice v
θ^v	$(\theta_b^v \triangleq \mathbb{E}[1/(1-\sigma_b^v)]: b \in \mathscr{B})$	Waiting parameter of slice of slice v
$\boldsymbol{\theta}^v$	diag (θ^v)	Diagonal matrix of waiting parameter of slice v

where BTD^v represents the average BTD of the vehicle at RSB b on slice v. Besides, we use $\langle x_1, x_2\rangle_M \triangleq x_1^T M x_2$ and $\|x\|_M \triangleq \sqrt{x^T M x}$ that denote the weighted inner product of the vectors and the weighted norm of a vector, respectively. M denotes a diagonal matrix.

We assume the message e handling process follows the $GI/M/1/\infty$ queue model [32]. Among which, the random variable of message arrival interval obeys general distribution, $F(t), t \geq 0$, while $F(t)$ in the different time slots is independent and identically distributed. Its expectation $1/\lambda = \int_0^\infty t dF(t), \lambda > 0$, where λ is the arrival rate. The message service time is exponentially distributed, i.e., $G(t) = 1 - e^{-\mu t}, t \geq 0$, and the mean value of the message service times depends on the number of allocated resource blocks. To facilitate the analysis of the system delay, $1/\mu_0$ is used to denote the average service time when only one resource block for the message processing. Similarly, we denote the average service time as $1/\mathrm{RB}_b^v\mu_0$ when RB_b^v resource blocks for the message processing.

The average waiting delay of the vehicle $u_b^v, u_b^v \in \mathscr{U}_b^v$ is given by

$$\mathrm{WD}_b^v = \frac{1}{\mathrm{RB}_b^v}(1-\sigma_b^v), \tag{6}$$

where $\mathrm{RB}_b^v = (s^v/n^v/\sum_{v'\in\mathscr{V}}(n_b^{v'} s^{v'}/n^{v'}))c_u$, and σ_b^v can be obtained by solving the following equation:

$$\int_0^\infty e^{-\mathrm{RB}_b^v \mu^v (1-\sigma_b^v)t} \frac{r_b^v \rho_b^v e^{-\rho_b^v + \rho_b^v e^{-\rho_b^v t} - r_b^v t}}{1-e^{-\rho_b^v}} dt = \sigma_b^v. \tag{7}$$

As a result, the average waiting delay of the vehicle on slice v is

$$\mathrm{WD}^v = \sum_{b\in B} \tilde{\rho}_b^v \mathrm{WD}_b^v = \frac{\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}}\rangle_{\theta^v}}{s^v \mu^v}. \tag{8}$$

According to formula $\langle x_1, x_2\rangle_{M_1+M_2} \triangleq \langle x_1, x_2\rangle_{M_1} + \langle x_1, x_2\rangle_{M_2}$ and (5) (8), the total average delay of a vehicle on slice v can be obtained by

$$D_{\mathrm{Total}}^v = \mathrm{BTD}^v + \mathrm{WD}^v = \frac{\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}}\rangle_{\mu^v \boldsymbol{\Delta}_v + \boldsymbol{\theta}^v}}{s^v \mu^v}. \tag{9}$$

3.2. Problem Formulation. In the real implementation of network slicing, each slice would provide a guarantee of service to the vehicles, that is, the total delay on slice v does not exceed a deadline d_v. In this subsection, we will explore how to obtain the optimal solution of minimizing system delay by allocating weight to each slice.

Considering the network with just one slice v, so we have $s^v = 1$, $\tilde{\boldsymbol{g}} = \tilde{\boldsymbol{\rho}}^v$. To satisfy the deadline d_v, from (9), we can obtain that

$$\rho^v \leq l(d_v, \tilde{\boldsymbol{\rho}}^v) \triangleq \frac{\mu^v d_v}{\|\tilde{\boldsymbol{\rho}}^v\|^2_{\mu^v \boldsymbol{\Delta}_v + \boldsymbol{\theta}^v}}, \tag{10}$$

where $l(d_v, \tilde{\boldsymbol{\rho}}^v)$ is the acceptable maximum load of slice v.

Next, considering multislice networks, each slice has its self-requirement. According to (9) (10), each slice would satisfy the following constraint to meet their requirements: $\forall v \in \mathscr{V}$

$$s^v \geq \frac{\rho^v}{l(d_v, \tilde{\boldsymbol{\rho}}^v) - \rho^v} \sum_{u\neq v} s^u \frac{\langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{\rho}}^u\rangle_{\mu^v \boldsymbol{\Delta}_v + \boldsymbol{\theta}^v}}{\|\tilde{\boldsymbol{\rho}}^v\|^2_{\mu^v \boldsymbol{\Delta}_v + \boldsymbol{\theta}^v}}. \tag{11}$$

Equation (11) can be written in a simplified form, i.e.,

$$\sum_{v\in V} s^v \boldsymbol{h}^v \geq \mathbf{0}, \tag{12}$$

where $\boldsymbol{h}^v = (h_u^v : u \in \mathscr{V})$ is share coupling vector of slice v and can be expressed by

$$h_u^v = \begin{cases} 1, & v = u \\ -\dfrac{\rho^u}{l(d_u, \tilde{\boldsymbol{\rho}}^u) - \rho^u} \dfrac{\langle \tilde{\boldsymbol{\rho}}^u, \tilde{\boldsymbol{\rho}}^v\rangle_{\mu^u \boldsymbol{\Delta}_v + \boldsymbol{\theta}^u}}{\|\tilde{\boldsymbol{\rho}}^u\|^2_{\mu^u \boldsymbol{\Delta}_v + \boldsymbol{\theta}^u}}, & v \neq u \end{cases} \tag{13}$$

Our objective is to satisfy the requirements of vehicles in each slice and minimize the system overall delay, so the objective function is the summing average delay of all slices. Consequently, the optimization problem can be formulated as

$$\begin{aligned} &\min_{s^v} \sum_{v\in V} \frac{\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}} \rangle_{\mu^v \Delta_v + \theta^v}}{s^v \mu^v} \\ &\text{s.t.} C1: \sum_{v\in V} s^v h_i^v \geq 0, i = 1, 2, \cdots, V \\ &C2: \sum_{i=1}^{V} s^i = 1 \\ &C3: s^i \geq 0, i = 1, 2, \cdots, V \end{aligned}, \tag{14}$$

where s^v is the optimization variable. Constraint $C1$ ensures the requirements of slices can be satisfied. Constraints $C2$ and $C3$ state the weight of each slice is nonnegative and is constrained by total resources. We can see that the problem (14) is an inequality constrained convex optimization problem, the method to solve it is described in Section 4.

3.3. Cellular Traffic Prediction. In practice, searching the optimal weights of each slice through minimizing the system total delay according to the current load distribution will cause some delay; we define that as arranging delay D_{ARR}. To reduce the arranging delay, traffic prediction is a feasible solution. Through traffic prediction, the system can acquire complicated traffic information in advance to calculate the optimal weighs for each slice. Hence, the arranging delay can be largely reduced.

ConvLSTM neural networks are adapted to predict cellular traffic. ConvLSTM networks can not only model the sequences information of cellular traffic accurately as same as Long Short-Term Memory (LSTM) networks [12] but also the local feature as same as convolutional neural networks (CNN). In short, it can easy to capture the spatial-temporal dependencies. Consequently, ConvLSTM networks are appropriate for predicting slice traffic in complex vehicular networks. This memory cell consists of cell states and three neural network units, i.e., input gate, forget gate, and output gate. For this specific framework, it is able to effectively store information chronically from long-term sequences.

As shown in Figure 2, the forget gate outputs a value $f_t \in [0, 1]$ to the cell according to the current input x_t and past output H_{t-1}, which determines what information would be abandon in past cell status C_{t-1} now. The calculation formula of forget gate is given by

$$f_t = \sigma\left(W_f^x * x_t + W_f^h H_{t-1} + b_f\right). \tag{15}$$

The input gate decides update when a new input comes to the ConvLSTM unit through a sigmoid function, which can

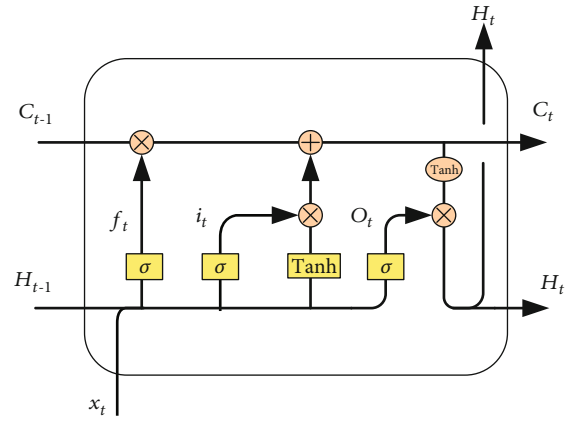

Figure 2: LSTM cell structure.

be further effects present states C_t, and expressed as

$$\begin{aligned} i_t &= \sigma\left(W_i^x * x_t + W_i^h H_{t-1} + b_i\right), \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tanh\left(W_c^x * x_t + W_c^h H_{t-1} + b_c\right). \end{aligned} \tag{16}$$

The output gate decides the output of this cell through a sigmoid function. After that, the output H_t is given by

$$\begin{aligned} o_t &= \sigma\left(W_o^x * x_t + W_o^h H_{t-1} + b_o\right), \\ H_t &= o_t \circ \tanh\left(C_t\right), \end{aligned} \tag{17}$$

where f_t, i_t, o_t, C_t, x_t, and H_t denote the output of forget gate, the output of input gate, cell status, cell input, and cell output, respectively, and W_i^x, W_f^x, W_c^x, W_o^x, W_i^h, W_f^h, W_c^h, W_o^h, b_f, b_i, b_c, and b_o are the parameters of the LSTM network. The functions $\sigma(\bullet)$ and $\tanh(\cdot)$ are sigmoid function and hyperbolic tangent function, respectively. In the above equations, the notation $*$ denotes the convolution operation, and the notation $\circ$ denotes the Hadamard product. Different from the common LSTM networks, the inputs or outputs in the ConvLSTM unit are all three-dimensional tensors. More specifically, the citywide service traffic data can be treated as a matrix or picture. Then, previous multiple service data are fed into the ConvLSTM networks to obtain future results. So, the multiply operation of common LSTM networks is replaced by convolution operation in ConvLSTM. The neural network can be accomplished via updating various parameters in each iteration, e.g., W_i^x, W_f^x, W_c^x, W_o^x, W_i^h, W_f^h, W_c^h, W_o^h, b_f, b_i, b_c, and b_o, so that the networks can minimize the error between forecasted values and ground trues.

4. A Primal-Dual Interior-Point Method-Based Resource Allocation Algorithm

In Section 3, the LSTM-based resource allocation is presented to minimize the system delay. Then, we will explore how to obtain the optimal slice weight of resources in

Input: Initial x_0, λ_0, scale factor k, residual error ϵ_{feas}, duality gap error ϵ, J times observed load distribution $\widehat{\boldsymbol{\rho}}_{t-J+1}, \widehat{\boldsymbol{\rho}}_{t-J+2}, \cdots, \widehat{\boldsymbol{\rho}}_t$
Output: Optimal solution $x^*_{t+1}, \cdots, x^*_{t+K}$
Phase 1: Predict service traffic
Training the ConvLSTM networks to obtain parameters W_i^x, W_f^x, W_c^x, W_o^x, W_i^h, W_f^h, W_c^h, W_o^h, b_f, b_i, b_c, b_o
According to J times observed load distribution predict K sequences in the future $\tilde{\boldsymbol{\rho}}_{t+1}, \cdots, \tilde{\boldsymbol{\rho}}_{t+K} = \underset{\boldsymbol{\rho}_{t+1}, \cdots, \boldsymbol{\rho}_{t+K}}{\operatorname{argmax}} p(\boldsymbol{\rho}_{t+1}, \cdots, \boldsymbol{\rho}_{t+K} \mid \tilde{\boldsymbol{\rho}}_{t-J+1}, \tilde{\boldsymbol{\rho}}_{t-J+2}, \cdots, \tilde{\boldsymbol{\rho}}_t)$
Phase 2: To obtain optimal slice weight
while True **do**
  Calculate initial value of the surrogate gap $\eta \leftarrow f(x)^T \lambda$
  if$\{\|\gamma_{\text{pri}} < \epsilon_{feas}\|\}\&\&\{\|\gamma_{\text{dual}} < \epsilon_{feas}\|\}\&\&\{\|\widehat{\eta}\| < \epsilon\}$**then**
    break
  end if
  Determine $t \leftarrow 2kV/\eta$
  Compute primal-dual search direction Δy_{pd}
  Determine initial step length $s_0 = \min\{0.99, \min\{-\lambda_i/\Delta\lambda_i \mid \Delta\lambda_i < 0\}\}$
  while $\min\{f_i(x + s\Delta x) \mid = 1, \cdots, 2V\} > 0$**do**
    Ensure satisfy the constraint condition $s \leftarrow \beta s$
  end while
  while$\|\gamma_t(x + s\Delta x, \lambda + s\Delta\lambda, v + s\Delta v) < \epsilon_{feas}\|_2 > (1 - \alpha s)\|\gamma_t(x, \lambda, v)\|_2$
do
    Determine backtracking search step length $s \leftarrow \beta s$
  end while
  Update search direction $y \leftarrow y + \Delta y_{pd}$
end while

Algorithm 1: Resource allocation algorithm based on the primal-dual interior-point method.

problem (14) and apply the LSTM in resource allocation processing.

The resource allocation strategy includes two phases: the first phase is service traffic prediction by using machine learning described in the preceding section, and the second phase is the optimizing procedure that based on the primal-dual interior-point method. Based on the predicted results in the first phase, the slice weight distribution process can be performed in advance so that the delay is decreasing obviously.

Considering there are inequality constraints in the convex optimization problem (14), so it can be solved by the primal-dual interior-point method. We will transform the inequality constrained optimization problem as an equality constrained optimization problem so that the central path of this problem can be found. Therefore, we rewrite the problem (14) as

$$\min \sum_{v \epsilon \mathscr{V}} \frac{\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}} \rangle_{\mu^v \Delta_v + \theta^v}}{s^v \mu^v} + \sum_{i=1}^{V} \left(-\frac{\log\left(\sum_{v \epsilon \mathscr{V}} s^v h_i^v\right)}{t} \right) + \sum_{i=1}^{V} - \frac{\log\left(s^i\right)}{t}$$
$$\text{s.t.} \sum_{i=1}^{V} s^i = 1 \; i = 1, 2, \cdots, V. \tag{18}$$

For simplicity, let $f_0(x) = \sum_{v \epsilon \mathscr{V}} (\rho^v \langle \tilde{\boldsymbol{\rho}}^v, \tilde{\boldsymbol{g}} \rangle_{\mu^v \Delta_{vv} + \theta^v} / s^v \mu^v)$, $\phi(x) = \sum_{i=1}^{V} (-\log(\sum_{v \epsilon \mathscr{V}} s^v h_i^v)) + \sum_{i=1}^{V} - \log(s^i)$, where $x = [s^1, s^2, \cdots, s^V, -f_0(x)]$ is the optimization variable. Considering the equivalent problem

$$\begin{aligned} &\min\ t f_0(x) + \phi(x) \\ &\text{s.t. } Ax = 1 \end{aligned}. \tag{19}$$

Due to the number of slices is three in the simulation setting, so $x = [s^1, s^2, s^3, -f_0(x)]$ and $A = [1, 1, 1, 0]$. According to reference [33], the Newton step Δ_y is given by the modified KKT equations

$$\begin{bmatrix} \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla^2 f_i(x) & Df(x)^T & A^T \\ -\mathbf{diag}(\lambda) Df(x) & -\mathbf{diag}(f(x)) & 0 \\ A & 0 & 0 \end{bmatrix}$$
$$\begin{bmatrix} \Delta x \\ \Delta\lambda \\ \Delta v \end{bmatrix} = \begin{bmatrix} \nabla f_0(x) + Df(x)^T \lambda + A^T v \\ -\mathbf{diag} f(x) - (1/t)\mathbf{1} \\ Ax - \mathbf{1} \end{bmatrix} = \begin{bmatrix} \gamma_{\text{dual}} \\ \gamma_{\text{cent}} \\ \gamma_{\text{pri}} \end{bmatrix}, \tag{20}$$

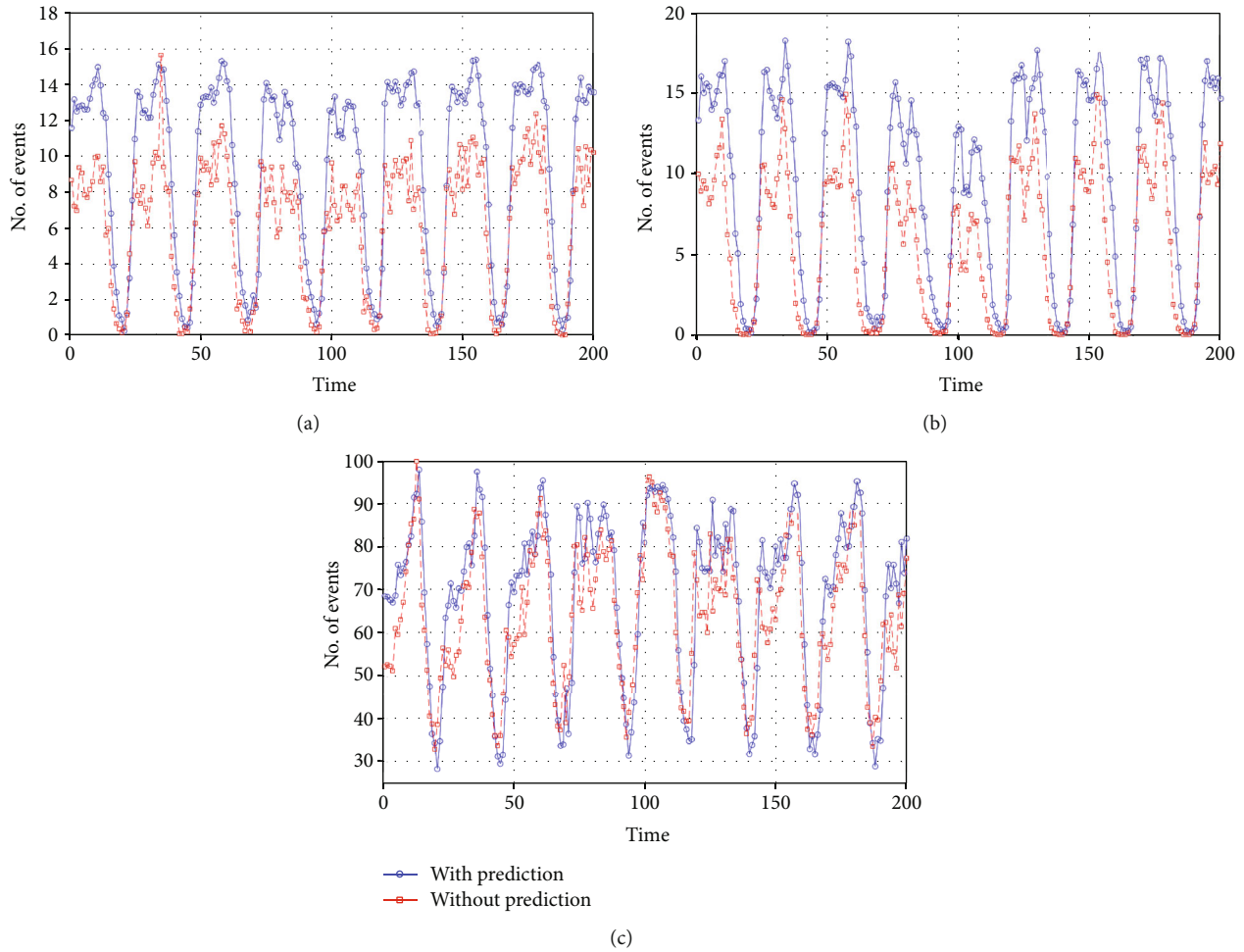

FIGURE 3: The comparison of traffic with and without prediction for three types of service. (a) SMS. (b) Phone. (c) Web.

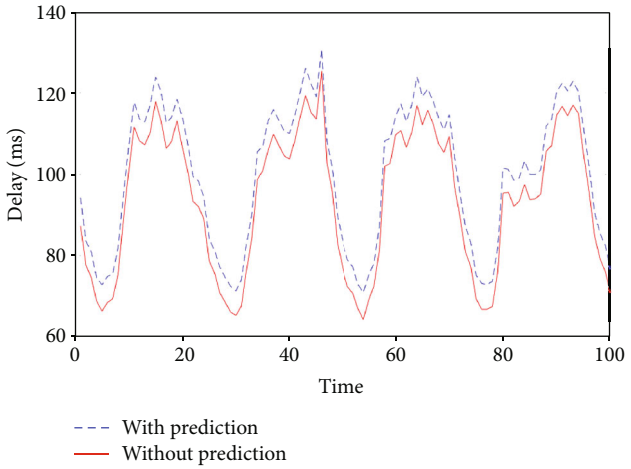

FIGURE 4: Network slicing delay with and without prediction.

where $\lambda_i = -(1/tf_i(x))\boxtimes i = 1, \cdots, 2$,

$$(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_{2V}(x) \end{bmatrix} \tag{21}$$

is the inequality constrained function of the original problem,

$$Df(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_2(x)^T \end{bmatrix} \tag{22}$$

is its derivative matrix. γ_{dual} and γ_{pri} are dual residual and primal residual, respectively, and these residuals are used for the termination condition in the primal-dual interior-point method. The solution of (20) is the current primal-dual search direction $\Delta y_{\text{pd}} = (\Delta x_{\text{pd}}, \Delta\lambda_{\text{pd}}, \Delta v_{\text{pd}})$.

The step length can be obtained by backtracking line search which is based on the norm of the residual. Through continuous iteration, the best solution will be returned when

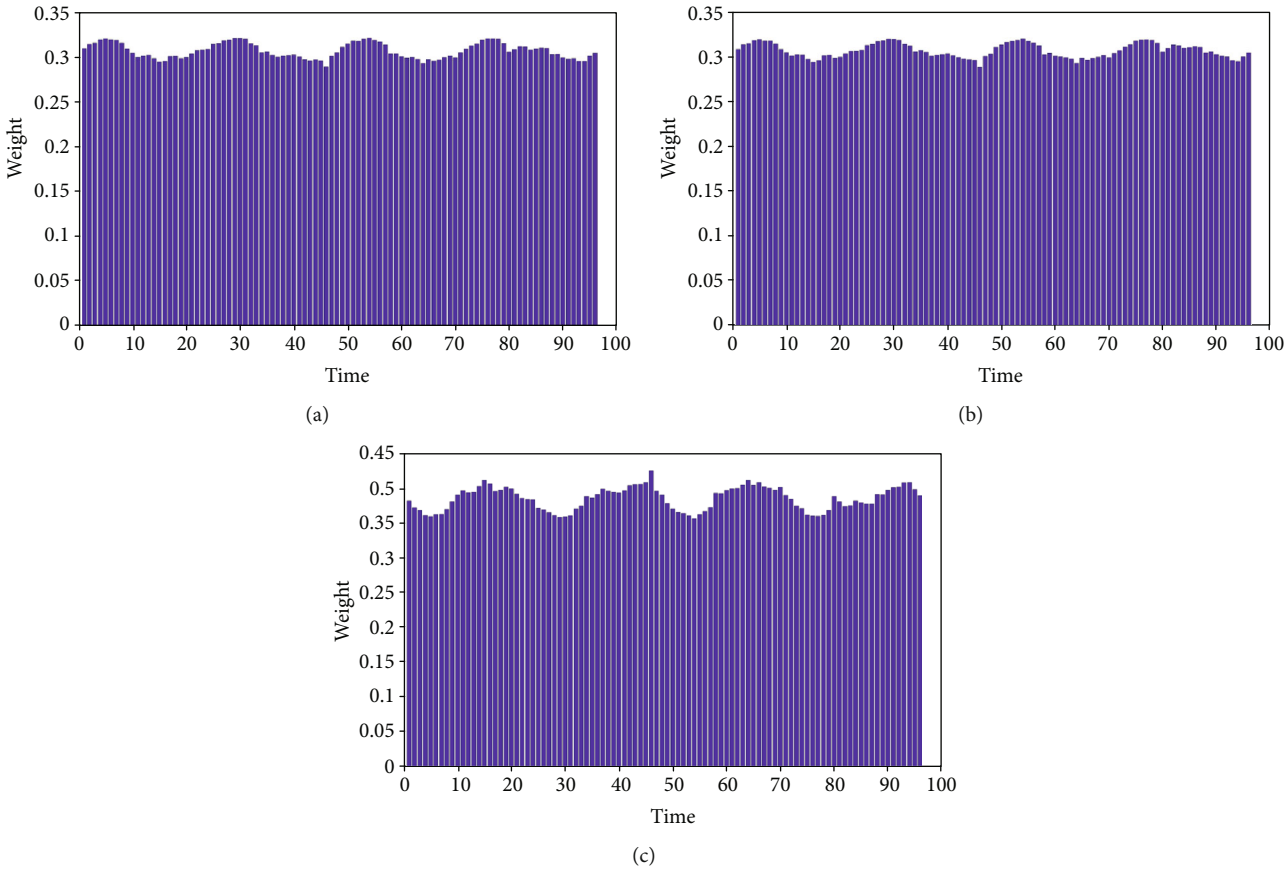

Figure 5: Optimal slice weight allocation at different times of three types of slice. (a) SMS. (b) Phone. (c) Web.

the present state satisfies the termination condition; the details are elaborated in Algorithm 1.

5. Performance Evaluation

In this section, we present numerical experiments to illustrate the performance of the proposed traffic prediction and resource allocation strategy. We use the cellular traffic dataset [34] of Milan, Italy, to train our neural networks; this dataset contains three categories of service traffic, i.e., SMS, phone, and web traffic, and we can deem it as three slices. The area of Milan is divided into a grid overlay of 100×100 squares; the dataset records 1000 samples of each square with a temporal interval of one hour. We choose the 800 samples from this dataset as the training set, the rest of the samples as the testing set. In the simulation setting of ConvLSTM, the neural network is set to three layers, and each layer has three ConvLSTM cells. In the training stage, we set the learning rate as 0.01, batch size as 32. After 100 times iteration, the trained model is used to predict the three types of service traffic. Then, the primal-dual interior-point resource allocation strategy is used to solve the optimal slice weight allocation according to the predication. In simulation parameters, we set scale factor $k = 2$, residual error $\epsilon_{\text{feas}} = 10^{-6}$, and duality gap error $\epsilon = 10^{-8}$.

Figure 3 shows the difference between predicted traffic and real traffic in a certain cell. The vertical axis of Figure 3 is the number of user access for service, and the horizontal axis of Figure 3 is the temporal dimension. It can be seen from Figure 3 that the temporal activities of three services show a strong daily pattern and follow weekly-periodic properties. Besides, the traffic volume expresses a slight difference among the three services. Specifically, the number of web access is always larger than that of the other two services. It can be seen that the service traffic shows a strong daily pattern. The average error rates of SMS, phone, and web traffic are 25.0%, 12.4%, and 12.2%, respectively.

Figure 4 shows the network slicing delay with and without prediction when slice weight is fixed, and the default of arranging delay D_{ARR} was set to 20 ms. The system delay also exhibits daily characteristics. Further, we can get that the more the number of user access, the greater the total average delay of the system. It can be seen that the periodicity of service traffic is mapped dynamically to system delay completely. Using the service traffic prediction, the slice weight can be distributed on average 15.33 ms in advance.

The predicted results express system load distribution in the future; the optimal slice weight can be solved accordingly. In Figure 5, we present the optimal slice weight under various times. Especially, at the 12th, 36th, 60th, and 84th time slots, the optimal slice weights of SMS are 0.3011, 0.3028, 0.3001, and 0.3076, respectively; the slice weights of phone and web traffic are 0.3032, 0.3059, 0.3018, and 0.3015; and 0.3957, 0.3913, 0.3982, and 0.3819. The results clearly show that the

optimal resource allocation problem is resolved under different times; the optimal slice weight is dynamic and varies with the service traffic pattern.

From the above results, it can be seen that the traffic prediction can be used to predict the service traffic in the future, so the user load distribution can be obtained. Therefore, the slice weight can be distributed in advance to save arranging delay. Furthermore, the primal-dual interior-point-based resource allocation strategy can calculate the current optimal slice weight distribution.

6. Conclusions

For slicing resource allocation problem in vehicular networks, this paper proposes an LSTM-based resource allocation, which contains two phases, i.e., traffic prediction phase and resource allocation phase. Moreover, we use ConvLSTM to capture the spatial-temporal dependencies of service traffic of each slice. Therefore, the user load distribution is obtained through traffic prediction. Then, based on the predicted user load distribution, we propose a primal-dual interior-point-based resource allocation strategy to solve the optimal slice weight problem for minimizing the system total delay. The resource allocation strategy proposed in this paper can make the slice weight be allocated in advance, thus greatly saving the delay. Moreover, with the development of AI, the research field of wireless communication will be impacted deeply. In the future, we will investigate the implementation of resource allocation with machine learning methods in vehicular network slicing.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

An earlier version of this article was presented at the 2020 IEEE/CIC International Conference on Communications in China. This work was supported in part by the National Natural Science Foundation of China (grant numbers 61801065, 61771082, 61871062, and 61901070) and the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202000603).

References

[1] S. Seng, C. Luo, X. Li, H. Zhang, and H. Ji, "User matching on blockchain for computation offloading in ultra-dense wireless networks," *IEEE Transactions on Network Science and Engineering*, 2020.

[2] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Communications Magazine*, vol. 54, no. 4, pp. 84–91, 2016.

[3] NGMN, *Description of Network Slicing Concept Version 1.0*, NGMN, Frankfurt, Germany, 2016.

[4] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 881–895, 2019.

[5] H. Halabian, "Distributed resource allocation optimization in 5G virtualized networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 627–642, 2019.

[6] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: mobility, resource management, and challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[7] R. Li, Z. Zhao, X. Zhou et al., "Intelligent 5G: when cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, 2017.

[8] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: enabling customization in multi-tenant networks," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, 2017.

[9] *System Architecture for the 5G System; Stage 2 (Release 15), document 3GPP TS 23.501 v1.0.0*, 2018.

[10] W. Paper, "5G Radio Access Capabilities and Technologies, Ericsson, White Paper Uen 284 23-3204 Rev C," 2016, April 2019, https://www.ericsson.com/assets/local/publications/white-papers/wp5g.pdf.

[11] V. Sciancalepore, K. Samdanis, X. Costa-Perez et al., "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, 2017.

[12] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5G wireless communications: a deep learning approach," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 227–236, 2020.

[13] H. M. Soliman and A. Leon-Garcia, "QoS-aware frequency-space network slicing and admission control for virtual wireless networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Washington, DC, USA, 2016.

[14] D. T. Hoang, D. Niyato, P. Wang, A. De Domenico, and E. C. Strinati, "Optimal cross slice orchestration for 5G mobile services," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Chicago, IL, USA, 2018.

[15] A. Aijaz, "Hap-SliceR: a radio resource slicing framework for 5G networks with haptic communications," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2285–2296, 2018.

[16] L. Liang, Y. Wu, G. Feng, X. Jian, and Y. Jia, "Online auction-based resource allocation for service-oriented network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8063–8074, 2019.

[17] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing in 5G: an auction-based model," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, May 2017.

[18] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2528–2541, 2018.

[19] H. Peng, L. Liang, X. Shen, and G. Y. Li, "Vehicular communications: a network layer perspective," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1064–1078, 2019.

[20] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2487–2500, 2018.

[21] L. Liang, H. Ye, and G. Y. Li, "Toward intelligent vehicular networks: a machine learning framework," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 124–135, 2019.

[22] S. Khan, H. A. Khattak, A. Almogren et al., "5G vehicular network resource management for improving radio access through machine learning," *IEEE Access*, vol. 8, pp. 6792–6800, 2020.

[23] J. Wang, B. Gong, H. Liu, and S. Li, "Multidisciplinary approaches to artificial swarm intelligence for heterogeneous computing and cloud scheduling," *Applied Intelligence*, vol. 43, no. 3, pp. 662–675, 2015.

[24] L. Nie, D. Jiang, S. Yu, and H. Song, "Network traffic prediction based on deep belief network in wireless mesh backbone networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–5, San Francisco, CA, USA, 2017.

[25] J. Wang, J. Tang, Z. Xu et al., "Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, 2017.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[27] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, 2019.

[28] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y. C. Liang, "Intelligent resource scheduling for 5G radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691–7703, 2019.

[29] R. Li, C. Wang, Z. Zhao, R. Guo, and H. Zhang, "The LSTM-based advantage actor-critic learning for resource management in network slicing with user mobility," *IEEE Communications Letters*, vol. 24, no. 9, pp. 2005–2009, 2020.

[30] 3GPP, "TR 38.802 v14.2.0. Technical Specification Group Radio Access Network; Study on new radio access technology physical layer aspects," in 3GPP, 2017.

[31] Y. J. Bultitude and T. Rautiainen, "IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II Channel Models," EBITG, TUI, UOULU, CU/CRC, NOKIA, Tech. Rep, 2007.

[32] L. Kleinrock, *Queueing Systems. Volume I: Theory*, Wiley, New York, NY, USA, 1972.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, Cambridge, UK, 2004.

[34] G. Barlacchi, M. de Nadai, R. Larcher et al., "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Scientific Data*, vol. 2, no. 1, article 150055, 2015.