Bandwidth Allocation for eMBB and mMTC Slices Based on AI-Aided Traffic Prediction

Xiaoli Zhang[1], Kai Liang[1(✉)], Jen-Jee Chen[2], and Jiaxin Liu[1]

[1] Xidian University, Xi'an 710071, China
kliang@xidian.edu.cn
[2] National Yang Ming Chiao Tung University, Hsinchu 30093, Taiwan

Abstract. Network slicing has emerged as a key enabler to address the diverse requirements of the fast-growing Internet of Things (IoT), such as enhanced Mobile Broad Band (eMBB) and massive Machine Type Communication (mMTC). However, the dynamic nature of service requirements propels challenges for resource allocation of network slicing. This paper proposes a bandwidth allocation method serving eMBB and mMTC slices with the aid of deep learning-based traffic prediction. The problem is formulated as maximizing the number of users served by the mMTC slice while satisfying the rate requirement of users served by the eMBB slice. We adopt traffic prediction by deep learning methods (i.e., ConvLSTM and GRU) to facilitate dynamic allocation of the bandwidth resource, which can be obtained by solving the integer programming problem. Numerical results show that the proposed algorithm outperforms the traditional method in terms of the average number of served mMTC users.

Keywords: Network slicing · Traffic prediction · ConvLSTM · GRU · mMTC · eMBB

1 Introduction

The fast-growing Internet of Things (IoT) leads to various scenarios and applications with increasingly multifarious service requirements [1], such as massive Machine Type Communication (mMTC) with the massive access requirements (e.g., smart cities, intelligent transportation) and enhanced Mobile Broad Band (eMBB) with the high data rate requirement (e.g., immersive games, Augmented Reality) [2]. This brings challenges to serve diverse service requirements under a shared physical infrastructure. Therefore IoT needs to be sliced and tailored to form logically virtual networks, and thus network slicing emerges [3].

This work was supported in part by National Key R&D Program of China (2019YFE0113200), National Natural Science Foundation of China (61901317), Joint Education Project between China and Central-Estern European Countries (202005).

K. Rabie et al. (Eds.): IoTaaS 2022, LNICST 506, pp. 117–126, 2023.
https://doi.org/10.1007/978-3-031-37139-4_11

There has been a lot of research in the field of network slicing. The algorithm UFLOP is proposed to minimize the "over-provisioning rate" by adjusting capacity and traffic allocation with satisfying the latency constraints of users and the service level agreement of the tenant [5]. Alsenwi proposes an algorithm based on deep reinforcement learning to maximize the eMBB data rate under the reliability constraints of ultra-reliable low-latency communication [6]. However, existing researches often reserve resources to prevent busty high data transmission and thereby fail to satisfy on-demand and dynamic resource allocation.

Traffic prediction algorithms can support real-time, on-demand resource allocation [4]. Graves first proposed Long-Short Term Memory (LSTM) which is based on Recurrent Neural Network (RNN), to solve problems such as gradients in long-term memory and backpropagation [7]. The LSTM algorithm performs well on temporal correlation, but it introduces redundancy when dealing with spatial data due to the strong local characteristics of spatial data [8]. To solve this problem, Shi *et al.* proposed the Convolutional LSTM (ConvLSTM) algorithm and experimentally proved that ConvLSTM outperforms LSTM in spatial and temporal data processing [9]. Additionally, Ramakrishnan *et al.* [10] used RNN, LSTM, and Gate Recurrent Unit (GRU) models to predict traffic information, demonstrated that GRU models perform better than RNN and LSTM predictors. While these traffic prediction algorithms are well researched, they are not yet effectively combined with the implementation of real-time on-demand allocation of network resources.

In this paper, we propose a bandwidth allocation algorithm for eMBB and mMTC slices based on AI-aided traffic prediction. The contributions of this paper are summarized as follows. Firstly, we jointly consider service requirements of eMBB slices and mMTC slices, and the problem is formulated as maximizing the number of mMTC users under the eMBB data rate constraint. Then, to dynamically allocate resources, we use the AI-aided traffic prediction to forecast the real-time rate requirement of the served eMBB users and take the prediction result as the boundary to reserve resources, thereby achieving better performance than traditional resource reservation methods.

The rest of this manuscript is organized as follows. Section 2 introduces the system model and the problem formulation. Section 3 discusses our proposed algorithm. In Sect. 4, we present simulation results. Finally, we conclude our discussions in Sect. 5.

2 System Model and Problem Formulation

We consider the uplink transmission of a single cell consisting of one base station(BS) serving an eMBB slice with E users and an mMTC slice with M users, as shown in Fig. 1. The BS and each user are equipped with a single antenna. The uplink channel is shared by all users and is divided into N subchannels with bandwith f, so the total amount of uplink bandwidth is $F = f \times N$. Let $X_{e,n} = 1$ indicate that the subchannel n is allocated to eMBB user e, and $X_{e,n} = 0$, otherwise. Similarly, $X_{m,n} = 1$ denotes that the nth subchannel is allocated to

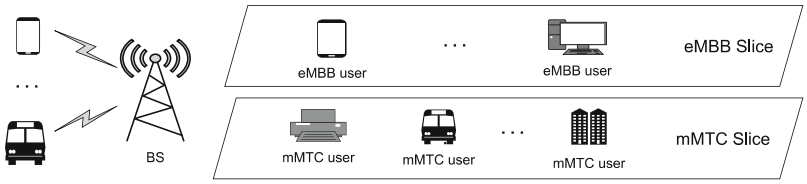


Fig. 1. System model

the mth mMTC user, and $X_{m,n} = 0$, otherwise. Because mMTC services often require massive access with a low data rate, we assume that one subchannel is enough for the data rate requirement of mMTC users. Therefore, the number of mMTC users U served by the mMTC slice is given as follows

$$U = \sum_{m}^{M} \sum_{n}^{N} X_{m,n}. \tag{1}$$

The transmit power of eMBB user e is set to P_e, so the uplink rate of eMBB user e is given by

$$R_e = \sum_{n}^{N} X_{e,n} f \log \left(1 + \frac{P_e |h_{e,n}|^2 \vartheta_{e,n}}{\sigma^2 f}\right), \tag{2}$$

where $h_{e,n} \sim CN(0,1)$ and $\vartheta_{e,n}$ denote Rayleigh fading coefficient and the path loss between the BS and eMBB user e on subchannel n, respectively. σ^2 presents the noise power spectral density.

In this paper, we focus on maximizing the number of users served by the mMTC slice with satisfying the rate requirement of eMBB users. Denote $R^e_{req}(t)$ as the rate requirement in time slot t. The optimal problem can be formulated as follows

$$\begin{aligned}
\max_{X_{e,n}, X_{m,n}} \quad & \sum_{m}^{M} \sum_{n}^{N} X_{m,n}(t) \\
\text{s.t.} \quad & C1: R_e(t) \geq R^e_{req}(t) \quad e = 1, \cdots, E \quad t = 1, \cdots, T \\
& C2: \sum_{e}^{E} X_{e,n}(t) + \sum_{m}^{M} X_{m,n}(t) \leq 1 \quad n = 1, \cdots, N \\
& C3: \sum_{n}^{N} \left[\sum_{e}^{E} X_{e,n}(t) + \sum_{m}^{M} X_{m,n}(t) \right] \leq N \\
& C4: X_{e,n}(t), X_{m,n}(t) \in \{0,1\} \\
& \quad n = 1, \cdots, N \quad m = 1, \cdots, M \quad e = 1, \cdots, E.
\end{aligned} \tag{3}$$

The constraint $C1$ stands for the rate requirements constraints of eMBB users. $C2$ means that subchannel n can be allocated to only one user. $C3$ indicates

that the number of subchannels allocated cannot exceed the total amount N, and $C4$ means that $X_{e,n}$ and $X_{m,n}$ are Boolean variables.

The problem (3) is a zero-one programming problem and can be solved directly by the CVX toolbox [12,13]. However, since the transmission rate of each time slot of eMBB users is unknown, traditional methods will reserve enough bandwidth resources for eMBB users to fulfill busty high data rate requirements, leading to low spectral efficiency and thereby serving fewer mMTC users. To deal with this problem, we will introduce a traffic prediction-based algorithm in the next section.

3 Traffic Prediction Based Bandwidth Allocation Algorithm

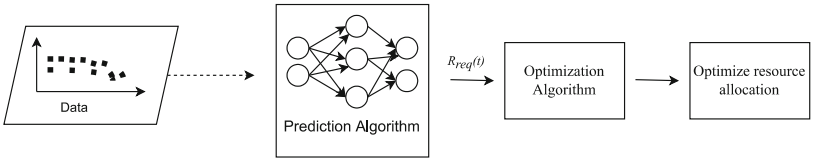


Fig. 2. Traffic Prediction based Bandwidth Allocation Algorithm.

In this section, we first predict the real-time rate of eMBB users based on ConvLSTM and GRU. After obtaining the result, we take it as the boundary of the constraint $C1$ in (3), to dynamically allocate resources, as shown in Fig. 2. Next, we first introduce two deep learning-aided traffic prediction algorithms and then use the prediction results to solve the optimization problem.

3.1 AI-Aided Traffic Predication

ConvLSTM Based Traffic Prediction. In order to improve the accuracy of the prediction, we need to normalize the rate R_t^e of the eMBB user e and turn it into a scalar r_t^e, and the expression is given as follows

$$r_t^e = \frac{R_t^e - \min(R_1^e, \cdots, R_T^e)}{\max(R_1^e, \cdots, R_T^e) - \min(R_1^e, \cdots, R_T^e)} \quad t = 1, \cdots, T, \tag{4}$$

where T is the total time. Then we take the processed data r_t^e as the input which flows to the ConvLSTM layer. The ConvLSTM units use the forget gate, input gate, and output gate to update their cell and hidden states. The forget gate decides how much of the cell state information at the previous time step is to be discarded, and the output of the forget gate is given as follows

$$f_t = \sigma(W_f * [C_{t-1}, h_{t-1}, r_t^e] + b_f), \tag{5}$$

where $*$ denotes the convolution operation and W_f represents the parameter matrix. C_{t-1} and h_{t-1} is the last moment information and b_f is the bias of every neuron. $\sigma(\cdot)$ represents the sigmoid activation function, i.e., $\sigma(x) = 1/(1+e^{-x})$.

The input gate is used to control the addition of new information, which is given as follows

$$i_t = \sigma(W_i * [C_{t-1}, h_{t-1}, r_t^e] + b_i). \tag{6}$$

Based on the previous memory cell state C_{t-1} and the forget state f_t , the cell state information updates with the formula

$$C_t = f_t \circ C_{t-1} + \tanh(W_c * [h_{t-1}, r_t^e] + b_c), \tag{7}$$

where tanh is the hyperbolic tangent function and $\circ$ is the Hadamard product.

The output gate filters the information achieved by the updated cell state, and the mathematical expression is given as follows

$$o_t = \sigma(W_o * [C_{t-1}, h_{t-1}, r_t^e] + b_o). \tag{8}$$

The hidden state is defined as follows

$$h_t = o_t \circ \tanh(C_t). \tag{9}$$

To alleviate overfitting problems, we add a dropout layer behind the first ConvLSTM layer. Then we reduce the dimension of the output of the dropout layer through flatten function. Finally, we take the output as the input of the next LSTM layer which replaces the convolution operation with matrix multiplication in ConvLSTM.

GRU Based Traffic Prediction. GRU has two gates, namely, the update gate and the reset gate. The update gate determines what information to forget and what new information needs to be added, while the reset gate is used to control how much of the previous information is forgotten. After achieving processed data r_t^e through the formula (4), the update gate z_t is given as follows

$$z_t = \sigma(W_z \times [h_{t-1}, r_t^e] + b_z), \tag{10}$$

and the reset gate s_t is formulated as

$$s_t = \sigma(W_s \times [h_{t-1}, r_t^e] + b_s). \tag{11}$$

Based on the update gate z_t and the reset gate s_t, the hidden state is updated and the formula is given as

$$\hat{h_t} = \tanh(W_h \times [s_t * h_{t-1}, r_t^e]), \tag{12}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t * \hat{h_t}. \tag{13}$$

We use the mean square error (MSE) to evaluate the performance of ConvLSTM and GRU prediction algorithms, which is given as follows

$$MSE = \frac{1}{T} \sum_{t}^{T} (\hat{r_t^e} - r_t^e)^2, \tag{14}$$

where $\hat{r_t^e}$ is the predicted result of r_t^e.

3.2 Prediction Based Optimized Algorithm

After the prediction of the above traffic prediction algorithms, we can obtain the predicted scalar $\hat{r_t^e}$ and scale it into the real-time rate requirements $\hat{R}^e_{req}(t)$ of eMBB users. However, since the results obtained by the prediction algorithm may not exactly match the actual requirements, especially in the case of high data transmission rate bursting (see Fig. 3. in Sect. 4), a correction factor η is added to prevent dissatisfying the eMBB data rate requirement $R^e_{req}(t)$. η is the median number of the difference between prediction and actual value when the prediction value is less than the actual value. Therefore, the optimization problem can be formulated as

$$\begin{aligned} \max_{x_{e,n}(t), x_{m,n}(t)} \quad & \sum_{m}^{M} \sum_{n}^{N} X_{m,n}(t) \\ \text{s.t.} \quad & C1 : R_e(t) \geq \hat{R}^e_{req}(t) + \eta \\ & C2, \quad C3, \quad C4 \end{aligned} \tag{15}$$

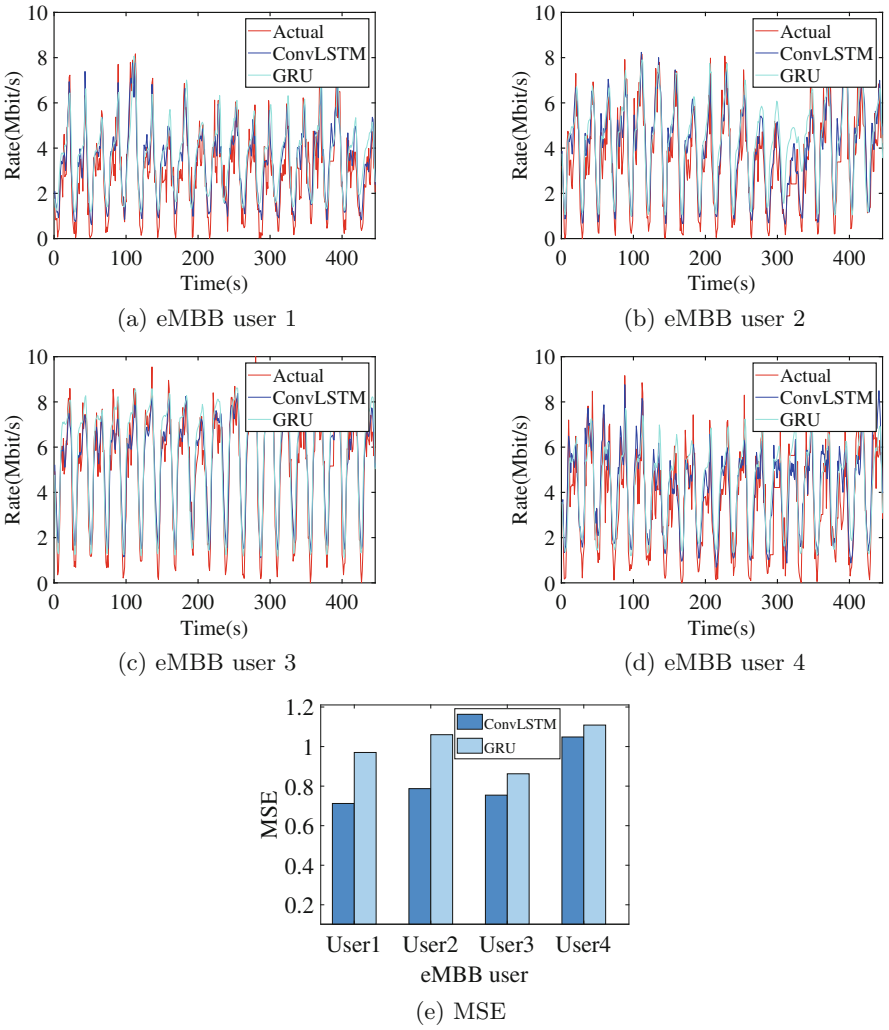
The new problem is similar to the original one, and can be solved with ease.

4 Numerical Results

We consider a circular network having a radius r = 500 m. The BS is located at the center of the network area. $E = 4$ eMBB users and $M = 20$ mMTC users are uniformly distributed around it. The path loss is computed by $\vartheta(dB) = 32.6 + 36.7log_{10}(d)$ [14], where d(in km) represents the distance between eMBB users and BS and is generated randomly within 500 m. The bandwidth of the subchannel f is 30 kHz and the number of subchannels N is 24. The transmit power P_e of each user is set to be $1W$, and $\sigma^2 = -174$ dBm/Hz [15]. For rate prediction, 4 sets of periodic oscillation signals of eMBB users are taken as the input of the prediction algorithm. We adopt the MSE loss function and the Adam optimizer whose initial learning rate is 0.0001. The window size of both ConvLSTM and GRU equals to 24. The model is trained in batch mode, and the batch size is 32. The training set and testing set contain 70% and 30% of the total dataset, respectively.

Figure 3 shows the prediction results obtained by ConvLSTM and GRU algorithms for each eMBB user. Except for some particularly high data transmission rates required, ConvLSTM and GRU algorithms both profoundly capture the dynamic fluctuations of the oscillating signal, and from Fig. 3 (e) we find that ConvLSTM outperforms GRU, because ConvLSTM introduces convolution operations and has a better grasp of the data features than GRU, e.g., for eMBB user 2, the MSE of ConvLSTM achieves 0.78, 25.7% lower than that of GRU.

We compared the performance of the proposed algorithm with the traditional method which adopts the way of reserving resources to prevent bursty rate requirements of eMBB users. The perfect prediction knows the rate requirements of eMBB users and is considered as a reference standard. To facilitate



(a) eMBB user 1 (b) eMBB user 2

(c) eMBB user 3 (d) eMBB user 4

(e) MSE

Fig. 3. Prediction results

the analysis of the optimal results, we average the optimized number of served mMTC users over a period T, which is formulated as follows

$$\overline{U} = \frac{1}{T} \sum_{t}^{T} U(t), \tag{16}$$

where $U(t)$ represents the number of served mMTC users at time slot t through the optimized algorithm. Let $\overline{U}_{conv}$ denote the average optimized result of ConvLSTM based optimization algorithm, and $\overline{U}_{gru}$ denotes the counterpart of GRU.

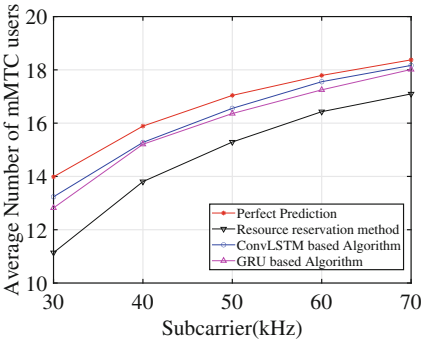


Fig. 4. Subchannel bandwidth versus the number of served mMTC users.

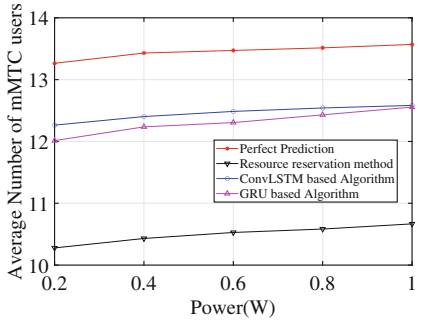


Fig. 5. Transmission power versurs the number of served mMTC users.

Figure 4 shows the trend of the average number of served mMTC users $\overline{U}$ with different subchannel bandwidths $f \in \{30, 40, 50, 60, 70\}$ kHz. We find that both $\overline{U}_{conv}$ and $\overline{U}_{gru}$ are ascended as the subchannel bandwidth increases. When the subchannel bandwidth $f = 50$ kHz, the number of served mMTC users based on ConvLSTM and GRU prediction algorithm increases by 8.3% and 6.9% compared to the traditional method respectively. The number of served mMTC users is almost equal to the result based on perfect prediction when $f = 70$ kHz.

Figure 5 shows the average number of served mMTC users $\overline{U}$ with different values of transmission power $P_e \in \{0.2, 0.4, 0.6, 0.8, 1\}W$ of eMBB users. We find that the proposed algorithms outperform the traditional method in terms of the number of served mMTC users. The algorithm we proposed exhibits an average increase of 18.87%, with a maximum increase of 19.92%.

5 Conclusions

Network slicing technology promotes the diverse requirements of the fast-growing IoT, such as eMBB and mMTC. However, the traditional resource reservation method of network slicing cannot satisfy the dynamic nature of business

demands. In this paper, we propose a bandwidth allocation algorithm based on deep learning traffic prediction for eMBB and mMTC slices. Firstly, we maximize the number of served mMTC users with the traditional method. Then, based on the result of the traffic prediction algorithm (i.e., ConvLSTM and GRU), we dynamically allocate bandwidth resources to eMBB and mMTC users by solving an integer programming problem. Finally, the numerical results demonstrate that the proposed method outperforms the traditional method in terms of increasing the number of served mMTC users.

References

1. Shafique, K., Khawaja, B.A., Sabir, F., Qazi, S., Mustaqim, M.: Internet of things (IoT) for next-generation smart systems: a review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. IEEE Access **8**, 23022–23040 (2020)
2. Wijethilaka, S., Liyanage, M.: Survey on network slicing for Internet of Things realization in 5G networks. IEEE Commun. Surv. Tutor. **23**(2), 957–994 (2021)
3. Cao, J., et al.: A survey on security aspects for 3GPP 5G networks. IEEE Commun. Surv. Tutor. **22**(1), 170–195 (2019)
4. Zhang, J., Zheng, Y., Sun, J., Qi, D.: Flow prediction in spatio-temporal networks based on multitask deep learning. IEEE Trans. Knowl. Data Eng. **32**(3), 468–478 (2019)
5. Chien, H.T., Lin, Y.D., Lai, C.L., Wang, C.T.: End-to-end slicing with optimized communication and computing resource allocation in multi-tenant 5G systems. IEEE Trans. Veh. Technol. **69**(2), 2079–2091 (2019)
6. Alsenwi, M., Tran, N.H., Bennis, M., Pandey, S.R., Bairagi, A.K., Hong, C.S.: Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: a deep reinforcement learning based approach. IEEE Trans. Wireless Commun. **20**(7), 4585–4600 (2021)
7. Graves, A.: Long Short-Term Memory. Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45 (2012)
8. Essien, A., Giannetti, C.: A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders. IEEE Trans. Industr. Inf. **16**(9), 6069–6078 (2020)
9. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
10. Ramakrishnan, N., Soni, T.: Network traffic prediction using recurrent neural networks. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 187–193. IEEE (2018)
11. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
12. Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. https://cvxr.com/cvx (2013)
13. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control (a tribute to M. Vidyasagar), pp. 95–110, Lecture Notes in Control and Information Sciences, Springer, London (2008). https://doi.org/10.1007/978-1-84800-155-8_7, https://stanford.edu/~boyd/graph_dcp.html

14. Jiang, T., Cheng, H.V., Yu, W.: Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation. IEEE J. Sel. Areas Commun. **39**(7), 1931–1945 (2021)
15. Liang, K., Zhao, L., Yang, K., Chu, X.: Online power and time allocation in MIMO uplink transmissions powered by RF wireless energy transfer. IEEE Trans. Veh. Technol. **66**(8), 6819–6830 (2017)