dmTP: A Deep Meta-learning Based Framework for Mobile Traffic Prediction

Zitian Zhang, Fuyou Li, Xiaoli Chu, Yuguang Fang, and Jie Zhang

Abstract

Deep learning technologies have been widely exploited to predict mobile traffic. However, individually training deep learning models for various traffic prediction tasks is not only time consuming but also unrealistic, sometimes due to limited traffic records. In this article, we propose a novel deep meta-learning based mobile traffic prediction framework, namely, dmTP, which can adaptively learn to learn the proper prediction model for each distinct prediction task from accumulated meta-knowledge of previously learned prediction tasks. In dmTP, we regard each mobile traffic prediction task as a base-task and adopt an LSTM network with a fixed structure as the base-learner for each base-task. In order to improve the base-learner’s prediction accuracy and learning efficiency, we further employ an MLP as the meta-learner to find the optimal hyper-parameter value and initial training status for the base-learner of a new base-task according to its meta-features. Extensive experiments with real-world datasets demonstrate that while guaranteeing a similar or even better prediction accuracy, meta-learning in the proposed dmTP reduces the numbers of epochs and base-samples needed to train the base-learners by around 75 percent and 81 percent, respectively, as compared with the existing prediction models.

Introduction

In order to support the rapidly growing mobile services [1-2], key technical challenges in mobile traffic analysis, including mobile traffic prediction, anomaly detection, attack classification, website fingerprinting, and mobile traffic identification, have been widely studied in recent years [3]. Among these challenges, the accurate prediction of mobile traffic is a critical enabler of advanced management and optimization of network resources.

Estimating future traffic loads based on historical traffic records, mobile traffic prediction has been widely considered a time series forecasting problem. The existing prediction models or algorithms can be generally classified into two categories: statistical methods and machine learning based methods [4].

In statistical methods, explicit statistical models with certain parameters are used to fit the mobile traffic patterns for traffic forecasting. In [5-6], the linear autoregressive integrated moving average (ARIMA) model and seasonal ARIMA model were adopted, respectively, to capture the short-term or long-term correlation in network traffic. Li et al. [7] demonstrated that traffic loads generated by three typical mobile services follow heavy-tailed distributions and then proposed to utilize the a-stable model to predict their load fluctuations. However, since realistic mobile traffic tends to show complex irregular patterns, it is difficult for statistical models to accurately predict realistic mobile traffic.

Unlike the statistical methods, machine learning based methods package the statistical models used for prediction into opaque or semi-opaque black-boxes, which need to be trained using historical traffic records. In [8, 9], linear regression (LR) and support vector regression (SVR) were used to predict the network-level mobile traffic, respectively. Nevertheless, these shallow learning methods cannot cope with many practical prediction scenarios due to the fact that they cannot perform feature extraction on their own, but also rely on some prior knowledge of the input features. Recently, powerful deep learning tools have been leveraged for mobile traffic prediction. Nie et al. [10] employed a deep belief network based model to predict the mobile traffic loads aggregated over a city. Feng et al. [11] proposed a long short-term memory (LSTM) network based prediction model to forecast the traffic loads in the target cell. With a reduced connection complexity, a random connectivity LSTM network based traffic prediction model was proposed in [12]. A convolutional neural network based model [13] and a convolutional LSTM network based model [4] were proposed to forecast the spatial mobile traffic distribution in a city. However, in the existing works, a specific prediction model must be constructed and trained for each individual mobile traffic prediction task as the traffic patterns handled by various tasks are quite different. Separately training the prediction models for multiple tasks is not only time consuming but also unrealistic since there are not always sufficient historical traffic records available, for example, for newly built areas.

To fill in the above gaps, this article presents an early attempt to introduce deep meta-learning into mobile traffic prediction and investigate how to make the prediction model learn to learn for a specific mobile traffic prediction task according to its characteristics (meta-features). The main contributions of this work are summarized as follows:
We introduce deep meta-learning into mobile traffic analysis, including mobile traffic prediction, anomaly detection, attack classification, website fingerprinting, and mobile traffic identification, have been widely studied in recent years [3]. Among these challenges, the accurate prediction of mobile traffic is a critical enabler of advanced management and optimization of network resources.

Through Fast Fourier Transform (FFT), we demonstrate that the temporal variations of mobile traffic over hours, days, and weeks can be characterized by five main frequency components of the frequency spectrum. More specifically, our analysis shows that the Pearson correlation coefficient of the normalized traffic load series for two base-tasks is negatively correlated with the Euclidean distance between their frequency component vectors. In other words, the similarity of traffic load series for two base-tasks tend to have similar time domain variations if their frequency component vectors are close to each other, and vice versa.

We introduce deep meta-learning into mobile traffic prediction, where each prediction task is regarded as a base-task and is represented as a time series forecasting problem. By defining the meta-task as learning to learn the proper prediction models for different base-tasks, we propose a novel deep meta-learning based mobile traffic prediction framework (dmTP). In dmTP, we adopt an LSTM network with a fixed structure as the base-learner to forecast the traffic load for a base-task based on previous values. Using the five main frequency components as the meta-features, we employ a multi-layer perceptron (MLP) as the meta-learner to output the optimal hyper-parameter value and initial status for the base-learner of a new base-task according to the accumulated meta-knowledge and meta-features for the base-task.

The performance of dmTP is evaluated by extensive tests using real-world mobile traffic data for heterogeneous prediction tasks. Numerical results show that the meta-learning technology not only improves the prediction accuracy and learning efficiency but also makes them more adaptable to abnormal traffic patterns.

**Dataset Description and Background Knowledge of Meta-Learning Mobile Traffic Traces**

In this article, we adopt three real-world mobile traffic datasets generated in Milan (Dataset 1), Guangzhou (Dataset 2), and London (Dataset 3), respectively. Figure 1 provides the details about these datasets. Dataset 1 is publicly available [4], while Datasets 2 and 3 were purchased and are not publicly accessible.

We refer to each grid in Dataset 1 as a cell and regard the forecasting problem for each cell’s mobile traffic load as an individual prediction task. As the city area of Milan is divided into 9999 grids, there are 9999 prediction tasks for Dataset 1. For Datasets 2 and 3, we choose the forecasting problem for short massage service (SMS) traffic load generated in the whole Guangzhou city and that for Twitter traffic load generated in the whole London city as two prediction tasks, respectively. These prediction tasks from Datasets 1, 2, and 3 focus on different spatial scales and involve different kinds of mobile traffic.

**Characteristics of Mobile Traffic**

We set the time interval resolution at one hour following the settings in [4] mainly because time interval resolutions smaller than one hour will result in the situation that many cells from Dataset 1 have lots of zero values in their traffic load series, which will be too sparse for traffic prediction. The peak traffic loads among various prediction tasks from the three datasets differ by up to five orders of magnitude, and we normalize the traffic load series of each prediction task into the range of [0,1] using the max-min scaling method. Figure 2a illustrates the normalized mobile traffic loads in three prediction tasks during two weeks. We observe that although the three mobile traffic streams exhibit different temporal patterns, they all exhibit a weekly cyclic pattern, which is also observed in time series of other cells from Dataset 1.

We generate a discrete periodic signal for each mobile traffic prediction task by periodically repeating its normalized real traffic loads in the first secular week (from Monday to Sunday) and then perform FFT of it. Please note that although the constructed periodic signal only approximates the actual mobile traffic load series, it allows us to capture the main features of a prediction task with much fewer records of historical traffic loads. Figure 2b shows the amplitudes of the frequency components in the FFT results related to the three

| Dataset 1 | Location | Duration | Items |
|-----------|----------|----------|-------|
| Mobile network traffic | Milan, Italy | 11/01/2013-01/01/2014 | 319,896,289 records |

**Description**: Records provided by Telecom Italy with a temporal interval of 10 minutes. Milan city area is divided into 9,999 grids and the size of a grid is about 235 x 235m². Each traffic record has the information about its time interval, geographical grid, and data volume in bytes.

| Dataset 2 | Location | Duration | Items |
|-----------|----------|----------|-------|
| Short message service traffic | Guangzhou, China | 03/01/2019-03/31/2019 | 1,521,005 records |

**Description**: Records provided by Unicom China. Each short message service record has the information about its time-stamp and data volume.

| Dataset 3 | Location | Duration | Items |
|-----------|----------|----------|-------|
| Twitter | London, U.K. | 02/15/2016-02/28/2016 | 136,710 records |

**Description**: Twitter records purchased from the Twitter Company. Each record has the information about its time-stamp and city target.

FIGURE 1 Description of the three adopted datasets. Note that our study does not breach user privacy or raise ethical or legal issues. All the datasets used have been pre-processed to ensure anonymity. In particular, all personal information has been removed before data analysis.
The Pearson correlation coefficient of two prediction tasks’ time series is negatively correlated with the Euclidean distance between their frequency component vectors. In other words, two prediction tasks’ time series tend to have similar time domain variations if their frequency component vectors are close to each other, and vice versa. This implies that the frequency component vector can be used to characterize the features of a prediction task’s traffic pattern.

FIGURE 2. Characteristics of mobile traffic in time domain and frequency domain: a) temporal patterns of three mobile traffic time series (network traffic generated in cell 1684 in Milan, SMS traffic generated in Guangzhou, and Twitter traffic generated in London); b) FFT results of those time series in a); c) CDF of base-tasks whose sum energy of the five main frequency components accounts a certain percentage of the related periodic signal’s energy; d) dependencies of time series pairs in both time domain and frequency domain.

time series in Fig. 2a. From Fig. 2b, we find that in the frequency domain, $\omega = \pi/84$, $\pi/12$, $\pi/6$, $\pi/4$, $\pi/3$ (corresponding to the periods of one week, one day, 12 hours, 8 hours, and 6 hours, respectively) are the five main frequency components for a mobile traffic prediction task. However, the amplitudes of the main frequency components vary evidently across different prediction tasks. Figure 2c shows the cumulative distribution function (CDF) of the signal energy carried by the five main frequency components. We observe that the sum energy of the five main frequency components in more than 60 percent of the considered prediction tasks exceeds 60 percent of the signal energy.

We use a frequency component vector of size 10 to record the real parts and imaginary parts of a prediction task’s main frequency components. Obviously, this vector can reflect both the amplitude and the phase of each main frequency component. Figure 2d illustrates inter-dependencies of 5000 randomly selected pairs of time series (i.e., prediction tasks) from Dataset 1 in both time domain and the corresponding frequency domain. Specifically, Fig. 2d plots the Euclidean distance between the two frequency component vectors of a pair of prediction tasks vs. the Pearson correlation coefficient between the two corresponding time series. It can be observed that the Pearson correlation coefficient of two prediction tasks’ time series is negatively correlated with the Euclidean distance between their frequency component vectors. In other words, two prediction tasks’ time series tend to have similar time domain variations if their frequency component vectors are close to each other, and vice versa. This implies that the frequency component vector can be used to characterize the features of a prediction task’s traffic pattern.

META-LEARNING

Meta-learning studies how learning systems can increase learning efficiency through experience. The goal of meta-learning is to understand how learning itself can become flexible according to the domains or tasks under study [14].

For a typical supervised learning task $\xi$, and a learner $\zeta$, each sample is denoted by labeling a number of features with an unknown target function $F_x$ and the hypothesis space of learner $\zeta, \mathcal{H}_\zeta$, is defined as the set of all the possible hypothesis functions generated by $\zeta$. The training progress of $\zeta$ can thus be seen as searching the hypothesis function $h_\zeta(x)$ that approximates $F_x$ over $\zeta$’s hypothesis space. $\zeta$ usually embeds a set of biases, which may be caused by the adopted learning algorithm, hyper-parameters, or the initial status. These biases may restrict the size of a base-learner’s hypothesis
space, and will affect how the base-learner searches the hypothesis space. Meta-learning matches the biases of a base-learner to an individual task, which is achieved by a meta-task that adaptively generates a proper set of biases for each learning task according to the learning task’s meta-features. The meta-task itself can be seen as a learning task and handled by a meta-learner. Accordingly, those original individual learning tasks are referred to as base-tasks.

**The Proposed dmTP**

Figure 3 shows the diagram of our proposed deep meta-learning based mobile traffic prediction framework, dmTP. In dmTP, we regard each individual mobile traffic prediction task as a base-task and present an LSTM network based prediction model with a fixed structure as the base-learner for it. We define a base-task’s frequency component vector as its meta-features. We regard the number of steps, which is a hyper-parameter defining the length of the input sequence and is denoted by $SN$, and the initial values of neural connection weights and neural thresholds as a base-learner’s set of biases. For each considered value of $SN$, the hypothesis space of a base-learner is composed of the mappings each of which transforms the $3 \times SN$-dimensional input sequence into an $1 \times SN$-dimensional output sequence. Thus, the value of $SN$ determines a base-learner’s hypothesis space and the initial parameter values determine the base-learner’s initial searching point in its hypothesis space.

Intuitively, since the meta-features of a base-task reflect the characteristics of its traffic pattern, the best set of biases of the base-learner will be influenced by this task’s meta-features. We use an MLP as the meta-learner to implicitly extract the correlation between the base-tasks’ meta-features and their best sets of biases, and output the best set of biases for a new base-task’s base-learner according to the meta-features.

**Notations in dmTP:**

$\text{SN}_{m}^\text{train}$ denotes the meta-task training set, which is equivalent to the set of base-tasks generating meta-samples for meta-task training. For base-task $m$ in $\text{SN}_{m}^\text{meta}$, we use $\text{S}_{base, m}(SN)$ to denote the large base-task training set of base-samples for training $m$’s base-learner when the number of steps equals $SN$, while we use $\text{S}_{base, m}(SN)$ to denote the set of base-samples for fine-tuning the base-learner’s performance. For base-task $m$ not in $\text{SN}_{m}^\text{meta}$, we use $\text{S}_{base, m}(SN)$ to denote the small base-task training set with few base-samples for fine-tuning the base-learner and use $\text{S}_{base, m}(SN)$ to denote the set of base-samples for testing the base-learner’s performance.

**LSTM NETWORK AS THE BASE-LEARNER**

In dmTP, we construct a multi-layer LSTM network with $Q$ layers to act as the base-learner. For a specific prediction task, this LSTM network will be continuously fed with a sequence of input vectors related to previous $SN$ time intervals and predict (as the output) normalized mobile traffic load in the next time interval. Each input vector consists of three attributes: the normalized mobile traffic load, the day of the week, and the hour of the day.

Figure 4 gives the structure of an LSTM memory block in the $q$th layer of the base-learner. An LSTM memory block is logically a recurrently connected subnet containing some functional modules called...
In practice, other structures of the meta-learner as well as the base-learners can be adopted in the proposed framework. However, finding the optimal structures for the learning models is outside the scope of this work.

The base-learner will take the given $SN_t^m$ as its number of steps and set its initial values of neural connection weights and neural thresholds according to the output of the meta-learner. Then, the base-learner is fine-tuned using the small base-task training set, $S_{\text{train, small}}$.²

**EVALUATION ON REAL-WORLD MOBILE TRAFFIC DATA**

**EXPERIMENTAL SETTINGS**

In our experiments, the base-learner is constructed as a three-layer LSTM network, where the output vectors of the first and the second LSTM memory blocks are both of size 5. According to the LSTM memory block structure, the total number of parameters to be trained for each base-learner is 428. The meta-learner has three hidden layers of size 300, 300, and 400, respectively.

We randomly select 8000 out of 9999 base-tasks from Dataset 1 to construct the meta-training set, $S_{\text{meta, train}}$. For each base-task $m$ in $S_{\text{meta, train}}$, we test $SN$ from 3 to 24. For a candidate $SN$ value, $SN_t^m$, we apply a sliding window with size $SN_t^m$ to split $m$'s normalized traffic load series and generate the base-samples by labeling each sequence of input vectors with the normalized traffic load in the next time interval. We randomly select 90 percent of these base-samples to construct $S_{\text{base, m(SN)}}$ while use the remaining ones to construct $S_{\text{base, m(SN)}}$. We examine the performance of dmTP on the remaining 1999 base-tasks of Dataset 1 that have not been selected for meta-training and the two testing base-tasks from Datasets 2 and 3. Similarly, for each testing base-task $n$, we apply a sliding window with size $SN_t^n$, which is output by the meta-learner, to generate the base-samples.

The performance of dmTP is compared with the existing time series forecasting methods including ARIMA [5], LR [8], SVR [9], and basic LSTM networks [12] for $SN = 12$, 24. For a fair comparison, we construct the basic LSTM network with the same structure of a base-learner in dmTP. We choose the mean square error (MSE) as the loss function, while the adaptive moment estimation algorithm [15] with the default learning rate is utilized to optimize the baseline LSTM networks as well as the meta-learner and base-learners in dmTP. We note that for approximately 20 percent of the base-tasks in Dataset 1, the traffic loads in more than 30 percent of time intervals are lower than 20 percent of the peak traffic load, while for approximately 84 percent of the base-tasks in Dataset 1 and the two base-tasks from Datasets 2 and 3, the traffic loads in more than 50 percent of time intervals are lower than 40 percent of the peak traffic load. Considering that the mean values of the traffic load series related to a significant portion of base-tasks are relatively low, we use the normalized root mean square error (NRMSE) to evaluate the accuracy of the considered prediction methods.

**PREDICTION PERFORMANCE**

Figures 5a–5d compare the prediction accuracy achieved by dmTP and the baseline methods. In order to test the applicability of meta-learning to other prediction methods, we evaluate the performance of ARIMA, LR, and SVR when their hyper-parameters (i.e., the number of steps for ARIMA and LR, and the kernel function for SVR) for a certain base-task are either fixed

² In practice, other structures of the meta-learner as well as the base-learners can be adopted in the proposed framework. However, finding the optimal structures for the learning models is outside the scope of this work.
require a large number of samples to train their prediction accuracy of the basic LSTM networks form better than ARIMA, LR, and SVR, as shown generated in various time intervals, and thus per-

Due to the deep learning capability, the basic LSTM networks with adequate training data are able to better performance than that of ARIMA and LR. nonlinearities in load variation, and thus achieves nonlinear prediction method, can deal with the patterns of mobile traffic loads. The SVR, which is a base-learner in dmTP and the baseline models.

First, the base-learner for each base-task in dmTP has the ability of learning and representing the complex nonlinearities in mobile traffic load variations. Second, unlike the basic LSTM networks embedded with fixed hyper-parameters and randomly selected initial parameter values, the meta-learner in dmTP will find the best SN and proper initial values of parameters for a base-learn-

As can be seen from Figs. 5a–5d, ARIMA and LR perform poorly among all the considered meth-

or selected by MLP based meta-learners from multiple candidate values (where candidate SN value ranges from 1 to 5 for ARIMA, candidate SN value ranges from 1 to 10 for LR, and candi-

dpert layer, the meta-learners for the above three baseline methods have the same structure as the meta-learner in dmTP. For each testing base-task from Dataset 1, the testing base-samples are generated during the two weeks of 12/16/2013–12/22/2013 and 12/23/2013–12/29/2013, where the traffic pattern during the second week has obvious abnormalities due to the Christmas holidays. For the two testing base-tasks from Datasets 2 and 3, the testing base-samples are generated in the last week, that is, 03/25/2019–03/31/2019 and 02/22/2016–02/28/2016, respectively. Note that for each testing base-task in Fig. 5, all the base-samples that have not been used for testing are used to fine-tune/train the base-learner in dmTP and the baseline methods.

models. The small training sets with base-samples generated in a three-week or one-week period will result in overfitting and thus poor performance. Figure 5d shows that when there are abnormalities in the testing traffic patterns, the basic LSTM networks have unsatisfactory performance as well. This is because these basic LSTM networks have high dependency on the training data and will fail to predict the abnormal traffic loads if there are no similar samples in their training sets.

The proposed dmTP always obtains the best prediction accuracy, attributing to two aspects. First, the base-learner for each base-task in dmTP has the ability of learning and representing the complex nonlinearities in mobile traffic load variations. Second, unlike the basic LSTM networks embedded with fixed hyper-parameters and randomly selected initial parameter values, the meta-learner in dmTP will find the best SN and proper initial values of parameters for a base-learner, which leads to a higher prediction accuracy and a stronger adaptability to the abnormal traffic pattern. Compared with ARIMA, LR, and SVR, dmTP reduces the NRMSE by about 25-43 percent, 73–85 percent, and 20–39 percent, respectively, for the testing base-tasks from Datasets 1, 2, and 3. As shown in Fig. 5a, when there is ade-

Train (s)

FIGURE 5. Performance of dmTP and the baseline methods: a) average NRMSE achieved over testing base-tasks from Dataset 1 (nor-

mal traffic); b) NRMSE for base-task related to Dataset 2; c) NRMSE for base-task related to Dataset 3; d) average NRMSE achieved over testing base-tasks from Dataset 1 (abnormal traffic); e) Prediction results of the dmTP and the basic LSTM networks for a testing base-task in Dataset 1 (cell 1684); f) ATT and TCMTM needed by dmTP and the baseline methods.
prediction values than the basic LSTM networks when the traffic pattern has anomalies or sudden changes. From Fig. 5, we also find that the three baseline methods in conjunction with meta-learning technology perform better than their counterparts without meta-learning technology, and the meta-learners can improve the prediction accuracy by about 9 percent, 17 percent, and 8 percent, for ARIMA, LR, and SVR, respectively. These results confirm the applicability of the meta-learning technology to ARIMA, LR, and SVR.

Figure 5f shows the average training time (ATT) needed by the base-learners in dmTP and the baseline methods for the testing base-tasks as well as the time needed to construct the meta-task training set and train the meta-learner (TCMTM). Compared with the baseline methods, although dmTP consumes much more TCMTM, after the initial parameter values have been properly set, the base-learner of a new base-task will converge faster and need shorter ATT than the basic LSTM networks. Note that since the meta-samples can be obtained off-line and the meta-learner only needs to be trained once, the extra off-line complexity caused by meta-learning in dmTP can be justified by the improved on-line convergence speed and reduced training time of base-learners.

**Learning Efficiency Improvement for Base-Learner**

In Fig. 6, we use the testing base-tasks from Data-set 1 to further examine how the meta-learner can help the base-learners improve their learning efficiency in terms of convergence speed and the number of base-samples needed. For each testing base-task, the testing base-samples are generated during the two weeks of 12/16/2013–12/22/2013 and 12/23/2013–12/29/2013, while different from Fig. 5, only a portion of the remaining base-samples are randomly selected to fine-tune/train the base-learner in dmTP and the baseline models.

Figure 6a shows the average NRMSE achieved by the proposed dmTP, basic LSTM networks with random initial parameters, and the LSTM network with transferred biases from a certain base-task in $e_{meta}^{train}$ (cell 6395 with the best SN of 6) vs. the number of epochs in training, where 840 base-samples are used to fine-tune/train the base-learner in dmTP and the baseline models for each testing base-task. We can see that for predicting normal mobile traffic, the NRMSE for basic LSTM networks decreases as the number of epochs increases and remains at a good accuracy level after 40 epochs. Due to the diversity of mobile traffic patterns among various base-tasks, the LSTM network with transferred biases has slower convergence speed and higher NRMSE than dmTP, but the transferred initial parameters result in lower initial NRMSE value and higher convergence speed than the basic LSTM networks. The dmTP’s performance becomes stable after 10 epochs, leading to a 75 percent reduction in the number of epochs needed than the basic LSTM networks. Thanks to the chosen optimal SN value, dmTP reduces the NRMSE by about 12 percent as compared with the basic LSTM networks. We can also see that for predicting abnormal mobile traffic, dmTP has much faster convergence speed than both the basic LSTM networks and the LSTM network with transferred biases. Moreover, the accuracy improvement for dmTP over the baselines after they all become stable is much larger than that for predicting normal mobile traffic. This demonstrates that the meta-learner can not only elevate the base-learners’ learning efficiency but also make them more adaptable to abnormal mobile traffic because the accumulated meta-knowledge will help each base-learner in dmTP handle unknown traffic patterns.

Figure 6b displays the average NRMSE achieved by dmTP and the baselines vs. the number of training base-samples selected to fine-tune/train the predicting models for each testing base-task with 100 training epochs. We can see that for predicting normal mobile traffic, the basic LSTM networks and the LSTM network with transferred biases obtain a high prediction accuracy if the base-task training sets are large enough (e.g., more than 500 training base-samples). The prediction accuracy of dmTP with 150 training base-samples exceeds that of the basic LSTM networks with 800 training base-samples. Compared with the basic LSTM networks, the meta-learner in dmTP helps the base-learners reduce the training base-samples needed by about 81 percent. Since each base-learner for a testing
base-task in dmTP is given a proper set of biases by the meta-learner, only a limited number of base-samples are required to fine-tune a base-learner to achieve high accuracy. For predicting abnormal traffic, the NRMSE for dmTP is significantly lower than those for the basic LSTM networks and the LSTM network with transferred biases. This is because without meta-knowledge, the LSTM networks will fail to accurately predict unknown traffic patterns if there are no similar base-samples in their training sets.

**Conclusion**

In this article, we introduce deep meta-learning into mobile traffic prediction and propose a novel deep meta-learning based mobile traffic prediction framework, dmTP. The prediction accuracy of dmTP has been tested on multiple real-world mobile traffic datasets. Experimental results demonstrate that dmTP is able to forecast the mobile traffic loads for prediction tasks with different spacial scales and application types. With meta-learning, the proposed framework can achieve a higher prediction accuracy and a higher adaptability in coping with abnormal traffic patterns than the existing prediction methods. Moreover, the meta-learner reduces the numbers of epochs and base-samples needed to fine-tune the base-learners by about 75 percent and 81 percent, respectively, as compared with LSTM network based prediction models.

Building on this work, it will be interesting to mathematically model and analyze the correlation between a base-task’s frequency component vector and the best set of biases for the task’s base-learner, as well as investigate whether the meta-learning technology can be used to optimize the structure of the base-learner for a prediction task. Additionally, it is worth studying if some other characteristics of mobile traffic patterns can be used as a base-task’s meta-features.

**Acknowledgment**

This work was supported in part by the EUs H2020 Research and Innovation Programme under Grant Agreement no. 778305 and 843401.

**References**

[1] Y. Huang, J. Tan, and Y.-C. Liang, “Wireless Big Data: Transforming Heterogeneous Networks to Smart Networks,” Commun. Inf. Netw., vol. 2, no. 1, Mar. 2017, pp. 19–32.

[2]Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016–2021, White Paper, 2017.

[3] G. Aceto et al., “Toward Effective Mobile Encrypted Traffic Classification through Deep Learning,” Neurocomputing, vol. 409, Oct. 2020, pp. 306–15.

[4] C. Zhang et al., “Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data,” IEEE JAC, vol. 37, no. 6, June 2019, pp. 1389–1401.

[5] B. Zhou, D. He, and Z. Sun, “Traffic Modeling and Prediction Using ARIMA/GARCH Model,” Proc. 2nd Conf. Next Gen. Internet Design Eng., 2006.

[6] Y. Shu et al., “Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models,” Proc. IEEE ICC, May 2003, pp. 1675–1679.

[7] R. Li et al., “The Learning and Prediction of Application-level Traffic Data in Cellular Networks,” IEEE Trans. Wireless Commun., vol. 16, no. 6, June 2017, pp. 3899–3912.

[8] H. Sun et al., “Short-Term Traffic Forecasting Using the Local Linear Regression Model,” Proc. 82nd Annu. Meet. Transp. Res. Board, Jan. 2003, pp. 1–10.

[9] N. I. Saprycky and R. Sankar, “Time Series Prediction Using Support Vector Machines: A Survey,” IEEE Comput. Intell. Mag., vol. 4, no. 2, May 2009, pp. 24–38.

[10] L. Nie et al., “Network Traffic Prediction Based on Deep Belief Network in Wireless Mesh Backbone Networks,” Proc. IEEE WCNC, Mar. 2017, pp. 1–5.

[11] L. Feng et al., “DeepTP: An End-to-End Neural Network for Mobile Cellular Traffic Prediction,” IEEE Network, vol. 32, no. 6, Nov. 2018, pp. 108–15.

[12] Y. Hua et al., “Traffic Prediction Based on Random Connectivity in Deep Learning with Long Short-Term Memory,” Proc. IEEE VTC-Fall, Aug. 2018, pp. 1–6.

[13] C. Zhang et al., “Citywide Cellular Traffic Prediction Based on Densely Connected Convolutional Neural Networks,” IEEE Commun. Lett., vol. 22, no. 8, Aug. 2018, pp. 1656–59.

[14] D. Li et al., “Learning to Generalize: Meta-Learning for Domain Generalization,” Proc. AAAI Conf. Art. Intell., Apr. 2018, pp. 3490–97.

[15] Y. LeCun et al., “Deep Learning,” Nature, vol. 521, May 2015, pp. 436–44.

**Biographies**

**Zitian Zhang** (zitian.zhang@ranplanwireless.com) is a Mary Curie research fellow at Ranplan Wireless Network Design Ltd. He received his B.S. and Ph.D. degrees in 2010 and 2016, respectively, from Shanghai Jiao Tong University. After his Ph.D. program, he worked as a research engineer at the China Aeronautical Radio Electronics Research Institute from 2016 to 2017. From 2018 to 2020, he was with East China University of Science and Technology. His research interests include deep learning, big data analytics, and device-to-device communications.

**Fuyou Li** (f17@sheffield.ac.uk) received the M.Sc. degree in communication and signal processing from Imperial College London in 2014. He is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, University of Sheffield. His research interests include wireless and mobile network systems, data analytics, and machine learning.

**Xiaoli Chu** (S03, M06, SM15) (x.chu@sheffield.ac.uk) is a lecturer at the University of Sheffield. She received a Ph.D. degree from Hong Kong University of Science and Technology in 2005. From 2005 to 2012, she was with Kings College London. She has published more than 80 peer-reviewed journal and conference papers. She is the lead editor/authors of the book Heterogeneous Cellular Networks, CUP, 2013. She was a Co-Chair of the Wireless Communications Symposium for IEEE ICC 2015, and a Guest Editor for IEEE Transactions on Vehicular Technology and ACM/Springer journal on Mobile Networks & Applications.

**Yuguang Fang** (Y08) (fang@eece.ufl.edu) received an M.S. degree from Qufu Normal University, Shandong, China in 1997, a Ph.D. degree from Case Western Reserve University in 1994, and a Ph.D. degree from Boston University in 1997. He joined the Department of Electrical and Computer Engineering at the University of Florida in 2000 as an assistant professor, then was promoted to an associate professor in 2003 and a full professor in 2005, and has been a distinguished professor since 2019. He received the 2018 IEEE Vehicular Technology Outstanding Service Award, the Best Paper Award from IEEE ICNP (2006), and the 2010-2011 UF Doctoral Dissertation Advisor/Mentoring Award. He was the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2013-2017), and the Editor-in-Chief of IEEE Wireless Communications (2009-2012). He is a Fellow of IEEE and AAAS.

**Jie Zhang** (M06, SM16) (jie.zhang@sheffield.ac.uk) has held the Chair in Wireless Systems at the Department of Electronic and Electrical Engineering, University of Sheffield (www.sheffield.ac.uk) since January 2011. He is also the Founder, Board Director and Chief Scientific Officer (CSO) of Ranplan Wireless (www.ranplanwireless.com), a public company listed on Nasdaq OMX. Along with his students and colleagues, he has pioneered research in small cell and heterogeneous networks (HetNets) and published some of the landmark papers and books on these topics, widely used by both academia and industry. He has published some of the earliest papers on using social media network data for proactive network optimization. Since 2010, he and his team have also developed ground breaking work in modelling and designing smart built environments considering both wireless and energy efficiency. His Google Scholar citations are in excess of 7000 with an H-index of 35. Prior to his current appointments, he studied and worked at Imperial College London, Oxford University, the University of Bedfordshire, and East China University of Science and Technology, reaching the rank of lecturer, reader and professor in 2002, 2005, and 2006, respectively.