IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Subway Passenger Flow Forecasting with Multi-station and External Factors

YAN DANFENG[1], WANG JING[2]

[1]State Key laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: tyshirley@bupt.edu.cn)

[2]State Key laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: yandf@bupt.edu.cn)

Corresponding author: WANG JING (e-mail: tyshirley@bupt.edu.cn).

ABSTRACT With the rapid development of urban rail transit, more and more people choose to travel by subway. Therefore, accurate passenger flow forecasting is of great significance for passengers and municipal construction. In this paper, we propose a multi-type attention-based network to forecast the subway passenger flow with multi-station and external factors. The proposed network has different types of attention mechanisms to adaptively extract relevant features including multi-station, external factors and historical data. In addition, the embedding method is applied to better combine the different kinds of data. Experiments on real subway passenger flow data in a city in China demonstrate that our method outperforms five baseline methods. Moreover, our method can visualize the impact of different stations and other factors on traffic, which plays an important role in passenger travel and subway dispatch.

INDEX TERMS Passenger flow forecasting, Attention mechanism, Recurrent neural networks

I. INTRODUCTION

In recent years, rail transit has developed rapidly. As an important part of rail transit, subway has become a major choice for people to travel with its advantages of timely and efficient. Therefore, reliable and accurate subway passenger flow forecasting is significant for passengers, transit operators, and public agencies [1], [2].

We define passenger flow as the number of passengers passing through the target station per unit time. From the time dimension, the forecasting of passenger flow can be regarded as the prediction of time series data, and there have been many studies on it. These studies focus on the prediction of a single source of time series data, and try to find the interconnection in the time dimension. However, it is not enough to consider only the information of the time dimension for subway passenger flow forecasting. Taking into account the characteristics of urban subway, we divide all the factors affecting passenger flow into three parts: the influence of subway stations on each other, external factors and historical data.

a: The influence of subway stations on each other

Urban subway can be regarded as a huge network, nodes of which are subway stations. Passengers' travel and transfer make the subway stations interact with each other. An intuitive example is the impact among sites on the same subway line. If the passenger flow of the upstream site is large, after a period of time, the passenger flow of the downstream site will increase accordingly. In addition, due to the division of urban functional areas, stations that are not on the same line would also interact with each other. For example, during the morning rush hour, passengers near a residential area flood into the central business district (CBD) to work. Similarly, during the evening rush hour, passengers return to the residential area from the commercial area. This phenomenon is shown in Fig. 1. In general, for a target station, all the other stations (common stations) have different effect at different time.

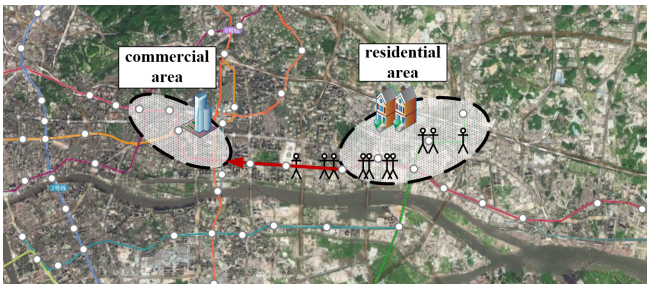

FIGURE 1. Example of passenger flow distribution during morning rush hour.

b: The influence of external factors

In addition to the subway stations, there are many factors that affect the passenger flow of the target station, including the properties of the station itself and environmental factors. For predicting time series data, the properties of the target station do not change over time, so these properties do not affect trend prediction. Environmental factors, such as weather and season, are characteristics of the time dimension and have a certain impact on passenger flow. Considering the passengers' age and occupation distributions, the traffic flow is related to the workday, i.e., passenger flow during the weekend and the working day are different. In addition, some studies have shown that holidays also have a certain impact on passenger flow. For example, during the Spring Festival, the passenger flow present a special form [3]. Moreover, different seasons and months also have some impact on passenger flow.

c: The influence of historical data

For urban subway, the daily passenger flow is basically the same, which can be considered as a time series data with a daily cycle. And for time series data, historical data contains important information.

In this paper, we propose a Multi-Type Attention-based Network to forecast the subway passenger flow with multi-station and external factors (subMTAN). The proposed network has different types of attention mechanisms for multi-station, external factors and historical data. It consists of two parts: passenger flow representation and passenger flow forecasting. In the representation part, we use relevant factors to represent the passenger flow at a certain moment. And we use different attention mechanisms to dynamically adjust the weights among different factors. In the forecasting part, we use a temporal attention mechanism to select relevant states across all the timestamps. These two parts can not only adaptively select the most relevant features, but also capture the long-term temporal dependencies of the passenger flow appropriately. Specifically, the attention vector for multi-station in the first part can be used to represent the influences of common stations on the target at different time. This plays an important role in the early warning and dispatch of subway passenger flow.

The rest of the paper is organized as follows. We first review the related work in Section II. Then we introduce the notations used in this work and the problem we aimed to study in Section III. In Section IV, we introduce the proposed network in detail. In Section V, we collect real subway passenger flow data for forecasting, and use experimental results to demonstrate the effectiveness of our proposed method. Finally, conclusion has been presented in Section VI.

II. RELATED WORK

A. SUBWAY PASSENGER FORECASTING

Autoregression-based models (e.g., ARIMA and VAR) are widely used in subway passenger forecasting [4], [5]. They show their effectiveness on various real world applications, but they cannot model nonlinear relationships and show poor performance on long-term prediction. In recent years, there have been some traffic prediction studies that have developed the spatial-temporal prediction approaches. W. Xu [6] proposes a method that can forecast complex data with both spatial and temporal attributes. Z. Xie [7] proposes a hybrid temporal-spatio forecasting approach to obtain the passenger flow status in HRTH. The approach combines temporal forecasting based on radial basis function neural network (RBF NN) and spatio forecasting based on spatial correlation degree. Y Sun [8] further proposes a transfer passenger flow prediction model based on nonparametric regression theory. Y. Zou [9] also proposes a space-time diurnal method to predict short-term freeway travel times which considers spatial and temporal correlation and diurnal pattern of travel times. H. Yang [10] proposes a hybrid model that embraces wavelet neural network (WNN), Markov chain (MAR), and the volatility (VOA) model for short-term travel time prediction in a freeway system. These studies take into account spatial and temporal travel time information simultaneously in the prediction approach. But most of the models are not good at nonlinear modeling.

With the development of deep learning, recurrent neural networks (RNNs) [11], [12], a type of deep neural network specially designed for sequence modeling, have received a great amount of attention due to their flexibility in capturing nonlinear relationships. Traditional RNNs, however, suffer from the problem of vanishing gradients and thus have difficulty capturing long-term dependencies. Recently, long short-term memory units (LSTM) [13] and the gated recurrent units (GRU) [14] overcome this limitation and achieve great success in various applications. Toque et al. [15] uses the LSTM to forecast dynamic public transport origin-destination matrices, and suggests that future research should consider the impact of exogenous variables such as weather and traffic accidents. Polson et al. [16] uses deep learning methods to predict short-term traffic flow in Chicago and takes into account the effect of extreme weather to improve the accuracy of forecasting results. For external factors, Zheng Y et al. [17] proposes a semi-supervised learning approach based on a co-training framework that consists of two separated classifiers. One is a spatial classifier based on an artificial neural network. The other is a temporal classifier based on a linear-chain conditional random field. These models do not take into account the different degrees of influence between features.

B. ATTENTION MECHANISM

Recently, attention mechanism has become popular due to its success in general sequence-to-sequence (seq2seq) problems. Bahdanau et al. [18] first introduces a general attention model in translation task. Later, researchers develope a number of multilevel attention-based models to select the relevant features and hidden states in different applications [19], [20]. To forecast the time series data, Qin et al. [21] proposes a dual-stage attention-based recurrent neural network (DA-RNN) to

select the relevant driving series at each time interval and select relevant states across all the timestamps. These attention-based seq2seq models are widely used in general sequence-to-sequence applications, but the actual situation needs to be considered in passenger flow prediction, and the accuracy needs to be improved. In this paper, we follow these ideas, and propose an attention-based network to better forecast the subway passenger flow using spatial and temporal travel time information simultaneously.

III. PRELIMINARY

A. NOTATIONS

In this paper, we suppose there are n+1 subway stations, including one target station and n other stations. We use $\boldsymbol{x^k} = (x_1^k, x_2^k, ..., x_T^k)^T \in \mathbb{R}^T$ to represent the passenger flow of station k with length T, where $x_t^k \in \mathbb{R}$ represents the number of passengers passing through the station k at time t. Among them, we specify one station as target station for making predictions, and use $\boldsymbol{y} = (y_1, y_2, ..., y_T)^T \in \mathbb{R}^T$ to denote the target station's passenger flow. In addition, we use $\boldsymbol{z_t} = (z_t^1, z_t^2, ..., z_t^m)^T \in \mathbb{R}^m$ to represent m external factors at time t, where $z_t^m \in \mathbb{R}$ represents the value of m-th factor at time t. Note that most of these factors are categorical which cannot be fed to neural network directly, so we transform each attribute into a low-dimensional vector by feeding them into different embedding layers separately. Detailed steps are introduced in the next chapter.

B. PROBLEM STATEMENT

Given the previous passenger flow of the target station, i.e. $(y_1, y_2, ..., y_{T-1})$ with $y_t \in \mathbb{R}$, as well as the current and past values of n stations, i.e., $\boldsymbol{X_{station}} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_T}) \in \mathbb{R}^{n \times T}$ ($\boldsymbol{x_t} = (x_t^1, x_t^2, ..., x_t^n)^T \in \mathbb{R}^n$ denotes a vector of n stations' passenger flow at time t), and m external factors, i.e. $\boldsymbol{Z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_T})$, the network aims to forecast the target station's passenger flow over t' time, denoted as $\hat{\boldsymbol{y}} = (\hat{y}_{T+1}, \hat{y}_{T+2}, ..., \hat{y}_{T+t'})^T \in \mathbb{R}^{t'}$

Fig. 2 shows more vividly the problem and the features of the paper. The black circle denotes the target station, other circles of the right half denote the rest of the stations, and the plane denotes the state of these stations at a timestamp. The circles of the left half denote the external factors. As illustrated in Fig. 2, first, the passenger flow of the target station has temporal dependency on its current state and that of its previous state. Second, it is reflected by its spatial neighbors(other stations). Last, it is also reflected by some external factors. All of these factors affect the passenger flow of the target station comprehensively.

IV. MULTI-TYPE ATTENTION NETWORK

In this paper, we propose a multi-type attention-based network to forecast the subway passenger flow, and its architecture is illustrated in Fig. 3. Following the encoder-decoder architecture [22], we use the "encoder" to encode the input sequences into a feature representation and the "decoder" to decode the information and make predictions. Therefore,

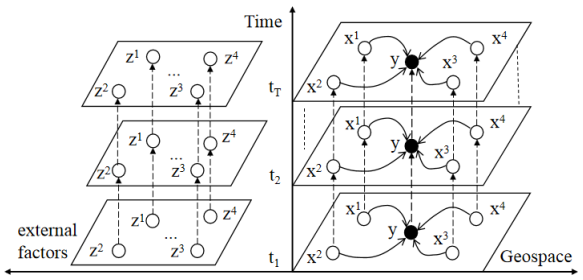

FIGURE 2. The philosophy of the proposed model.

we divide the whole model into two parts: passenger flow representation and passenger flow forecasting. Here, we use $\boldsymbol{h}_t \in \mathbb{R}^u$ to denote the hidden state of the encoder at time t. Likewise, $\boldsymbol{d}_t \in \mathbb{R}^v$ represents those of the decoder.

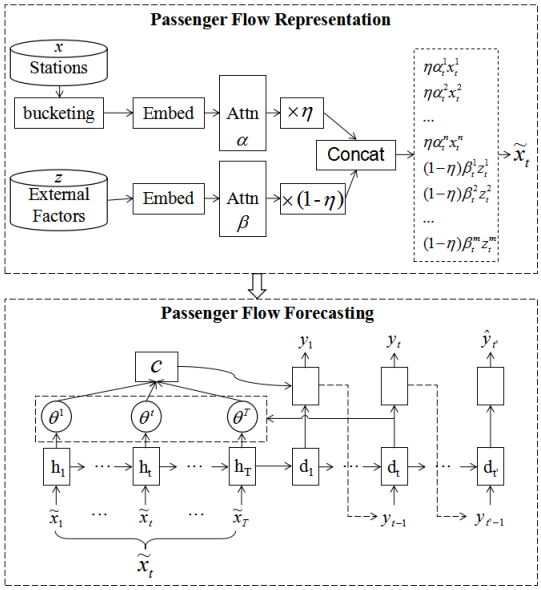

FIGURE 3. The framework of our approach. a) Passenger Flow Representation: we use stations and external factors to represent the target passenger flow, and feed the weighted input into the encoder LSTM units. Embed: embedding layer. Atten: attention layer. Concat: concatenation layer. η: a tunable hyperparameter. b) Passenger Flow Forecasting: Based on the seq2seq architecture, we use a temporal attention to calculate the weighted sum of the encoder hidden states across all the timestamps.

A. PASSENGER FLOW REPRESENTATION

In this part, we use all the relevant factors to represent the subway passenger flow. Given the input sequence $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, ..., \widetilde{\boldsymbol{x}}_T) \in \mathbb{R}^{(n+m) \times T}$, the encoder can be applied to learn a mapping from $\widetilde{\boldsymbol{x}}_t$ to $\boldsymbol{h_t}$ (at time step t) with

$$\boldsymbol{h_t} = f_1(\widetilde{\boldsymbol{x}}_t, \boldsymbol{h_{t-1}}) \tag{1}$$

where $\boldsymbol{h_t} \in \mathbb{R}^u$ is the hidden state of the encoder at time t, u is the size of hidden state, and f_1 is a non-linear activation function. In this paper, we use an LSTM [13] as the f_1.

As metioned in Section I, all factors are summarized into multi-station and external factors. In order to unify numerical and categorical data, we convert the data into vectors of

dimension d. This operation is similar to word embedding in natural language processing tasks. In this way, the categorical features have a "semantic" meaning and can be directly input into the neural network. For numerical data, in order to prevent over-fitting during training, we first bucket the data and then convert it into a d-dimensional vector. In this way, $x_t^k \in \mathbb{R}$ and $z_t^m \in \mathbb{R}$ are transformed into $\boldsymbol{x}_t^k \in \mathbb{R}^d$ and $\boldsymbol{z}_t^m \in \mathbb{R}^d$.

Since different factors have different effect on the target station at different time, we use the attention mechanisms to adaptively learn the weight of each factor at different time, and then use the weighted representation of all factors to represent the passenger flow $\widetilde{\boldsymbol{x}}_{\boldsymbol{t}}$. We develop two different attention mechanisms which capture complex correlation among stations and external factors.

a: Attention for multi-station

In the urban subway network, there is a complex correlation among stations. To address this issue, given the i-th station as our predictive target and another station j, we calculate the attention weight α (i.e., impacting weight) between them as follow.

First, we calculate the semantic similarity between stations, that is, the similarity between $\boldsymbol{x}^j$ and $\boldsymbol{h}_{t-1}$ as equation (2)

$$s_t^j = \mathbf{v}_s^T tanh(\boldsymbol{W}_s \boldsymbol{h}_{t-1} + \boldsymbol{U}_s \boldsymbol{x}^j \boldsymbol{w}_s' + \boldsymbol{b}_s) \tag{2}$$

where $\boldsymbol{v}_s, \boldsymbol{b}_s \in \mathbb{R}^T$, $\boldsymbol{w}_s' \in \mathbb{R}^d$, $\boldsymbol{W}_s \in \mathbb{R}^{T\times u}$ and $\boldsymbol{U}_s \in \mathbb{R}^{T\times T}$ are the parameters to be learned. By referring to the previous hidden state $\boldsymbol{h}_{t-1}$, this attention mechanism can adaptively select the relevant stations to make predictions. Note that the spatial factors also contribute to the correlations among different stations, we develop a matrix $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ to measure the geospatial similarity, where $D_{i,j}$ denotes the similarity between station i and j (Eq. (3)).

$$D_{i,j} = \frac{1}{dis_{i,j}^2} \tag{3}$$

where $dis_{i,j}$ is the shortest distance between station i and j. Considering the diversion effect of transfer stations (marked as Tr) on passenger flow, we redefine the distance from station i to its neighboring station j as equation (4).

$$dis_{i,j} = \begin{cases} 1 & j \notin Tr \\ p & j \in Tr \end{cases} \tag{4}$$

where p is the number of subway lines that station j connected, and Tr is the set of transfer stations.

Then, we use the semantic and geospatial similarity to get the final weight, and employ a softmax function to do the normalization.

$$\alpha_t^j = \frac{exp(\lambda s_t^j + (1-\lambda) D_{i,j})}{\sum_{l=1}^n exp(\lambda s_t^l + (1-\lambda) D_{i,l}))} \tag{5}$$

where λ is a tunable hyperparameter as a trade-off. The attention weight α_t^j represents the impact of station j on the target station at time t. It is used to visualize the interaction between stations and provide early warning.

b: Attention for external factors

Subway passenger flow also has a strong correlation with some external factors, such as weekday and holidays. Given the g-th external factor $\boldsymbol{z}^g = (\boldsymbol{z}_1^g, \boldsymbol{z}_2^g, ..., \boldsymbol{z}_T^g)^T$, we construct an attention mechanism referring to the previous hidden state $\boldsymbol{h}_{t-1}$ with

$$e_t^g = \mathbf{v}_e^T tanh(\boldsymbol{W}_e \boldsymbol{h}_{t-1} + \boldsymbol{U}_e \boldsymbol{z}^g \boldsymbol{w}_e' + \boldsymbol{b}_e) \tag{6}$$

and

$$\beta_t^g = \frac{exp(e_t^g)}{\sum_{i=1}^m exp(e_t^i)} \tag{7}$$

where $\boldsymbol{v}_e, \boldsymbol{b}_e \in \mathbb{R}^T$, $\boldsymbol{w}_e' \in \mathbb{R}^d$, $\boldsymbol{W}_e \in \mathbb{R}^{T\times u}$ and $\boldsymbol{U}_e \in \mathbb{R}^{T\times T}$ are the parameters to learn. β_t^g is the attention weight measuring the importance of the g-th input external factor at time t.

After calculating these two attention weights, we use a tunable hyperparameter η to combine the above factors. And the passenger flow can be represented as follow:

$$\widetilde{\boldsymbol{x}}_t = (\eta\alpha_t^1 \boldsymbol{x}_t^1, ..., \eta\alpha_t^n \boldsymbol{x}_t^n, (1-\eta)\beta_t^1 \boldsymbol{z}_t^1, ..., (1-\eta)\beta_t^m \boldsymbol{z}_t^m)^T \tag{8}$$

The elements in the vector of passenger flow on time t are the weighted factors that are related to the passenger flow. In this way, the passenger flow can be represented by all the relevant features and these features have their own weights respectively.

B. PASSENGER FLOW FORECASTING

In the forecasting part, we refer to the idea of attention-based seq2seq model [23] to produce output sequence. The network adaptively select the relevant hidden states of the encoder, i.e., model the dynamic temporal correlation between different time intervals in the target series. And is similar to that of [21]. The attention weight of each encoder hidden state at time t is calculated based upon the previous decoder hidden state $\boldsymbol{d}_{t-1}$ with

$$o_t^i = \mathbf{v}_d^T tanh(\boldsymbol{W}_d \boldsymbol{d}_{t-1} + \boldsymbol{U}_d \boldsymbol{h}_i + \boldsymbol{b}_d) \tag{9}$$

and

$$\theta_t^i = \frac{exp(o_t^i)}{\sum_{j=1}^t exp(o_t^j)} \tag{10}$$

where $\boldsymbol{v}_d, \boldsymbol{b}_d \in \mathbb{R}^v$, $\boldsymbol{W}_d \in \mathbb{R}^{v\times v}$ and $\boldsymbol{U}_d \in \mathbb{R}^{v\times u}$ are parameters to learn. The attention weight θ_t^i represents the importance of the i-th encoder hidden state for the prediction at time t. Since each encoder hidden state $\boldsymbol{h}_i$ is mapped to a temporal component of the input, the attention mechanism computes the context vector $\boldsymbol{c}_t$ as a weighted sum of all the encoder hidden states $\{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_T\}$ as equation (11).

$$\boldsymbol{c}_t = \sum_{i=1}^T \theta_t^i \boldsymbol{h}_i \tag{11}$$

Then, we update the decoder hidden state with $\boldsymbol{d}_t = f_d(\boldsymbol{d}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{c}_t)$, where f_d is an LSTM unit used in the decoder. Finally, we concatenate the context vector $\boldsymbol{c}_t$ with

the hidden state $\boldsymbol{d}_t$, which becomes the new hidden state from which we make final predictions as follows:

$$\hat{y}_T = \boldsymbol{v}_o^T(\boldsymbol{W}_w[\boldsymbol{d}_T; \boldsymbol{c}_T] + \boldsymbol{b}_w) + b_o \tag{12}$$

where $\boldsymbol{W}_w \in \mathbb{R}^{v\times(u+v)}$ and $\boldsymbol{b}_w \in \mathbb{R}^v$ map the concentration $[\boldsymbol{d}_T; \boldsymbol{c}_T] \in \mathbb{R}^{u+v}$ to the size of the decoder hidden state. And we use a linear transformation ($\boldsymbol{v}_o \in \mathbb{R}^v$ and $b_o \in \mathbb{R}$) to generate the final output.

The loss function is the mean squared error (MSE) between the predicted vector $\hat{\boldsymbol{y}}^i$ and the ground truth vector $\boldsymbol{y}^i \in \mathbb{R}^{t'}$ of station i:

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{t=1}^{t'}(y_t^i - \hat{y}_t^i)^2 \tag{13}$$

where Θ is a set of all parameters in the proposed model.

V. EXPERIMENTS

A. DATASET

We conduct our experiments on real subway passenger flow dataset of GuangZhou from 1/1/2016 to 4/1/2018, containing the information on passenger inbound and outbound. For each common stations, we sum the number of passengers entering and leaving the station to indicate the passenger flow and sample it at 10-minute intervals. As for external factors, we select 9 factors including festival, holiday, month, week, day, hour, weather, weekday and temperature. Then we divide the data into non-overlapped training set and validation set by a ratio of 9:1. i.e., we use the first two-year data as the training set, the first three months of the last year as the validation set. Table. 1 gives some statistics on dataset. Table. 2 shows an instance as an example.

TABLE 1. Simple statistics on dataset.

Information	**Value**	
total stations	159	
transfer stations	29	
external factors	9	
time spans	1/1/2016-4/1/2018	
Information	**Training set**	**Validation set**
instances	106272	11808
max passenger flow	636	528
mean passenger flow	24.4	25.2

TABLE 2. Example of an instance. (data of 08:00 1/1/2018)

Station$_1$	**...**	**Station**$_{n+1}$	**holiday**	**day**	**...**	**temperature**
23	...	109	2	1	...	20

B. EVALUATION METRICS AND HYPERPARAMETERS

a: Evaluation Metrics

We use multiple metrics to evaluate our model, including the rooted mean squared error (RMSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE), all of which are widely used in regression tasks. Assuming y_t is the target at time t and $\hat{y}_t$ is the predicted value at time t. RMSE is defined as $\boldsymbol{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_t^i - \hat{y}_t^i)^2}$, and MAE is defined as $\boldsymbol{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_t^i - \hat{y}_t^i|$. These two metrics are related to the value of dataset which cannot be compared on different datasets. MAPE expresses accuracy as a percentage, and is defined as $\boldsymbol{MAPE} = \frac{1}{N}\sum_{i=1}^{N}|\frac{y_t^i - \hat{y}_t^i}{y_t^i}| \times 100\%$.

b: Hyperparameters

Following the previous works [21] and [24], we set T=10 to predict. During the training phase, the batch size is 128 and the learning rate is 0.01. In external factors module, we embed the factors to $\mathbb{R}^4$. In multi-station module, we split a passenger flow interval into discrete bins instead of representing the passenger flow as a single continuous feature. To determine the size of the bucket, we conduct grid search over $bs \in \{1, 2, 3, 5, 10\}$. The one ($bs = 3$) that achieves the best performance over validation set is used for test. Also, we embed the bucket to $\mathbb{R}^4$. As for the λ and η, we also conduct a grid search and set $\lambda = 0.75$ and $\eta = 0.95$. For the sizes of hidden states for encoder and decoder, we set $u = v = 64$ because they achieve the best performance over the validation set.

C. MODEL COMPARISON AND PREDICTION RESULTS

To demonstrate the effectiveness of the proposed model, we compare our model with five baselines as follows:

- **ARIMA** [25]: A well-known model for forecasting future values in a time series. The basic idea of the ARIMA model is to treat the data sequence formed by the predicted object over time as a random sequence, and use a mathematical model to approximate the sequence. In this paper, we do stationary processing and test the order of the model, and finally use ARIMA(0,1,1).
- **GBRT** [26]: An ensemble method for the regression tasks and widely used in practice. Gradient Boosting Regression Tree (GBRT) models a data set based on relevant features and predicts the time series using the tree model. In this paper, we use common stations and external factors as relevant features. With grid search, we set the max depth to 10 and the estimator (number of regression trees) to 200.
- **DNN** [27]: A deep neural network(DNN)-based prediction model for spatio-temporal data. Same as GBRT, the DNN use all relevant features as input. In this paper, we set two dense layers, and each layer has 256 hidden units.
- **LSTM** [13]: A classic recurrent neural network for time series data prediction. Using a unique cellular structure, the LSTM cell can capture the long-term temporal dependencies. In this paper, we use two layers of LSTM with 128 hidden units.

- **DA-RNN** [21]: A dual-staged attention model for time series prediction, which shows the state-of-the-art performance in time series prediction. In the first stage, the model use an input attention mechanism to adaptively extract relevant driving series at each timestamp. In the second stage, a temporal attention mechanism is used to select relevant encoder hidden states across all the timestamps. In this paper, we use the same hyperparameters of the DA-RNN and use other stations' passenger flow as driving series.

The results are shown in Table. 3

TABLE 3. Performance comparison among different methods.

Method	RMSE	MAE	MAPE($\times 10^{-2}\%$)
ARIMA	9.029	3.682	13.16
GBRT	7.512	2.213	9.06
DNN	7.563	2.234	9.15
LSTM	7.221	1.613	6.61
DA-RNN	6.586	1.234	5.05
subMTAN	**5.782**	**1.024**	**4.19**

Table. 3 illustrates that our proposed method clearly outperforms all the baselines on all metrics, which verifies the advantage of our multi-type attention mechanisms. Notice that the ARIMA method is generally worse than RNN-based methods, because it only considers the historical data of the target and ignores other factors. This suggests that other factors can provide more positive information. The GBRT and DNN methods include the external factors but they just fit the nonlinearity relationship on the features, and cannot capture the long-term temporal dependencies of the passenger flow. LSTM method solves this problem but it cannot adaptively select the relevant features. It treats all features as the same. Our proposed model is 19.9% better than this method which indicates the reasonable choice among different features is important to make accurate predictions.

Specifically, DA-RNN can improve these shortcomings, but it use one single attention mechanism to select features. Our proposed model shows 17.0% and 12.2% improvements beyond the DA-RNN on MAE and RMSE respectively which demonstrates the advantages of the proposed multi-type attention mechanisms.

For visual comparison, we show the prediction results of these methods in Fig. 4. We choose a best method (DA-RNN) through these baselines, and observe that subMTAN generally fits the ground truth better than DA-RNN.

For subway management agencies, passenger flow at different time periods should be treated differently. So we divide prediction results into three categories including peak hour, off-peak hour in workday and weekend. We define the peak hour from 7:00 am to 9:00 am and 5:00 pm to 8:00 pm, and the rest of time as off-peak hour. Similarly, we choose DA-RNN to compare with our proposed model. The results are shown in Table. 4.

From Table. 4 we can see that the two models perform better during off-peak hour than the rest of the time. This is

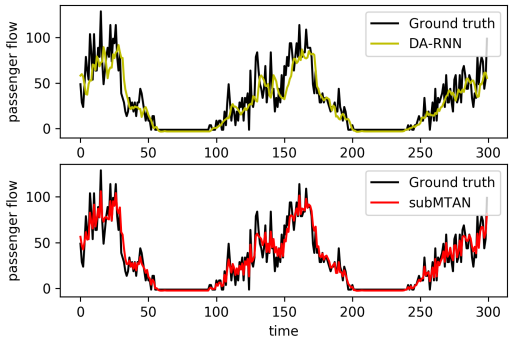

FIGURE 4. Passenger flow vs. Time. subMATN DA-RNN prediction result.

TABLE 4. Performance comparison among different time periods. The left side of "/" indicates the result of DA-RNN, and the right side indicates the result of subMATN

Time periods	RMSE	MAE	MAPE($\times 10^{-2}\%$)
peak hour	9.884 / 8.383	1.591 / 1.318	6.52 / 5.68
off-peak hour	4.291 / 4.220	0.842 / 0.840	3.46 / 3.44
weekend	7.573 / 6.649	1.289 / 1.075	5.28 / 4.44

because the factors affecting passenger flow during peak hour and weekend are more complicated. At the same time, we can see that in the off-peak hour, the metrics of the two models are relatively close. But during peak hours and weekend, DA-RNN's evaluation index increased more than our proposed model. This shows that our model can also better predict the peak traffic.

D. HYPERPARAMETERS SELECTION

When selecting the hyperparameters λ and η, we conduct grid search from 0 to 1. Notice that when λ is 1, it means that we remove the geospatial similarity. And when η is 1, it means that external factors are not considered. Fig. 5 shows the grid search' result of different hyperparameters.

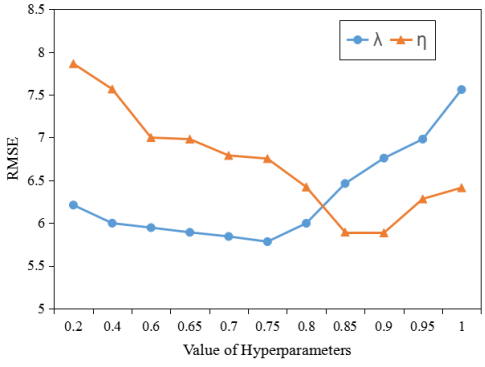

FIGURE 5. RMSE performance of λ and η among different values.

The resulting cruves of the two hyperparameters are all U-shape. For λ, the resulting curve shows that if we remove the geographical similarity factor (λ=1) or simply rely on

it (λ=0) for attention weight calculations, the effect will be worse. RMSE will be best only if we weight geographic similarity and semantic similarity with a weight of 0.75. Likewise, if we simple rely on external factors (η=0) or multi-station (η=1), the results are not good.

As for other hyperparameters, such as bs, u and v, we also conduct grid search to select the best value.

E. VISUALIZATION

The attention weight α_t^j indicates common stations' relationship with the target. Fig. 6(a) presents the relevance of common stations for the target (a station in the downtown business erea) during a day. The X-axis of the figure denotes the different stations, and Y-axis denotes the different time during a day. In this way, we can find where the network is focused on most at each time step.

We find that each station has a different impact on the target at different time. We represent the weights as different colors and draw them on the city map, as shown in Fig. 6(b). It is found that stations that have a strong influence in the morning are almost all in residential areas. This is consistent with our previous analysis.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a multi-type attention-based network for forecasting the subway passenger flow with the multi-station and external factors. The model contains three different attention mechanisms to adaptively select the relevant spatial and temporal features for the target passenger flow. With this weighted representation, we use encoder-decoder architecture to predict the passenger flow. In order to include more different kinds of data, we bucket the numerical data and add an embedding layer to unify categorical and numerical data. The experiments show that our proposed model achieves the best performance against five baselines in terms of three metrics (RMSE, MAE and MAPE) simultaneously. Moreover, we visualize the attention weights to show the interpretation of all stations in urban subway.

In the future, we will expand more relevant features, including some text or image information. We believe that more information will help with the prediction. Moreover, we will explore more efficient encoder unit and better model structures.

REFERENCES

[1] M.-C. Chen and Y. Wei, "Exploring time variants for short-term passenger flow," Journal of Transport Geography, vol. 19, no. 4, pp. 488–498, 2011.

[2] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal patterns of urban human mobility," Journal of Statistical Physics, vol. 151, no. 1-2, pp. 304–318, 2013.

[3] W. Xu, Y. Qin, and H. Huang, "A new method of railway passenger flow forecasting based on spatio-temporal data mining," in The International IEEE Conference on Intelligent Transportation Systems, 2004. Proceedings, pp. 402–405, 2004.

[4] S. Feng and G. Cai, "Passenger flow forecast of metro station based on the arima model," in Proceedings of the 2015 International Conference on Electrical and Information Technologies for Rail Transportation, pp. 463–470, Springer, 2016.

[5] Z. Q. L. D.-w. WANG Ying, HAN Bao-ming, "Forecasting of entering passenger flow volume in beijing subway based on sarima model," Journal of Transportation Systems Engineering and Information Technology, vol. 15, no. 6, p. 205, 2015.

[6] W. Xu, H. K. Huang, and Y. Qin, "A spatio-temporal forecasting method of railway passenger flow," in International Conference on Machine Learning and Cybernetics, pp. 1550–1554 vol.3, 2004.

[7] Z. Xie, L. Jia, Y. Qin, and L. Wang, "A hybrid temporal-spatio forecasting approach for passenger flow status in chinese high-speed railway transport hub," Discrete Dynamics in Nature and Society,2013,(2013-11-24), vol. 2013, no. 5, pp. 248–259, 2013.

[8] Y. Sun, G. Zhang, and H. Yin, "Passenger flow prediction of subway transfer stations based on nonparametric regression model," Discrete Dynamics in Nature & Society, vol. 2014, no. 3, pp. 1–8, 2014.

[9] Y. Zou, X. Zhu, Y. Zhang, and X. Zeng, "A spaceâĂŞtime diurnal method for short-term freeway travel time prediction," Transportation Research Part C: Emerging Technologies, vol. 43, pp. 33 – 49, 2014. Special Issue on Short-term Traffic Flow Forecasting.

[10] H. Yang, Y. Zou, Z. Wang, and B. Wu, "A hybrid method for short-term freeway travel time prediction based on wavelet neural network and markov chain," Canadian Journal of Civil Engineering, vol. 45, 09 2017.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors.," Parallel Distributed Processing Explorations in the Microstructure of Cognition, vol. 323, no. 6088, pp. 399–421, 1988.

[12] P. J. Werbos, "Backpropagation through time: what it does and how to do it," Proc IEEE, vol. 78, no. 10, pp. 1550–1560, 1990.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," Computer Science, 2014.

[15] F. ToquÃl', E. CÃt'me, M. K. E. Mahrsi, and L. Oukhellou, "Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks," in IEEE International Conference on Intelligent Transportation Systems, 2016.

[16] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," Transportation Research Part C Emerging Technologies, vol. 79, pp. 1–17, 2017.

[17] Y. Zheng, F. Liu, and H. P. Hsieh, "U-air:when urban air quality inference meets big data," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1436–1444, 2013.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Computer Science, 2014.

[19] L. Wang, Z. Cao, G. D. Melo, and Z. Liu, "Relation classification via multi-level attention cnns," in Meeting of the Association for Computational Linguistics, pp. 1298–1307, 2016.

[20] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in Computer Vision and Pattern Recognition, pp. 4187–4195, 2017.

[21] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," pp. 2627–2633, 2017.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," vol. 4, pp. 3104–3112, 2014.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Computer Science, 2014.

[24] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267–2276, 2015.

[25] G. E. P. Box and D. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," Publications of the American Statistical Association, vol. 65, no. 332, pp. 1509–1526, 1970.

[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[27] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "Dnn-based prediction model for spatio-temporal data," in ACM Sigspatial International Conference on Advances in Geographic Information Systems, p. 92, 2016.

[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Ad-

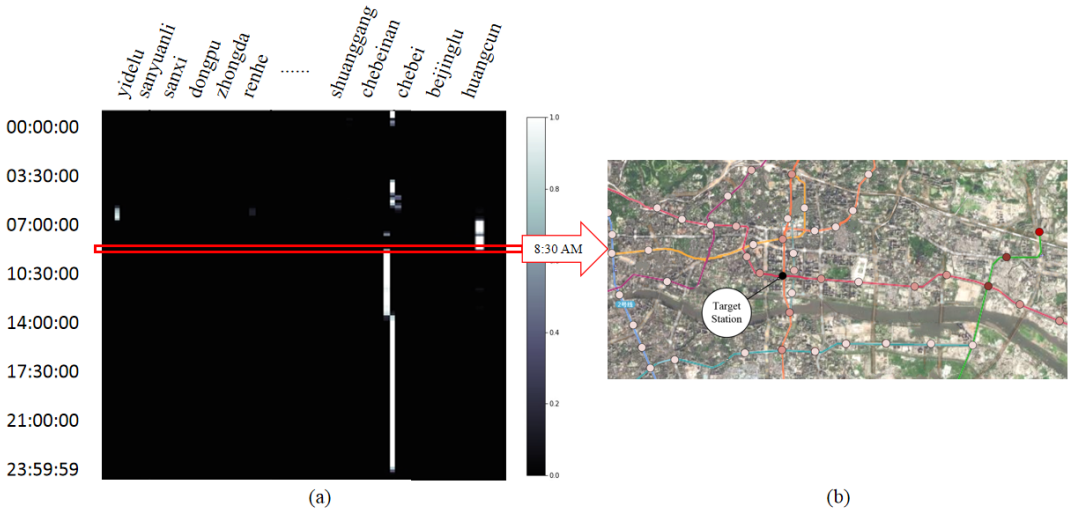

FIGURE 6. (a)The attention weights of different stations for the target during a day. (b)The attention weights for the target station in 8:30 AM. The darker the color, the greater the correlation.

vances in Neural Information Processing Systems, vol. 26, pp. 3111–3119, 2013.

YAN DANFENG Professor in State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications(BUPT). She obtained a doctor degree of computer science in BUPT. Her current research interests include Data Mining, Big Data and Analytics. Email: yandf@bupt.edu.cn

WANG JING was born in 1994. She received the bachelor degree in automation from the Beijing University of Posts and Telecommunications, China, in 2016. She started her studies with computer technology in State Key laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. She is currently purchasing the master degree. Email: tyshirley@bupt.edu.cn