IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 12, 2020, accepted January 23, 2020, date of publication February 12, 2020, date of current version February 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973406

Metro Passenger Flow Prediction Model Using Attention-Based Neural Network

JUN YANG[1], XUCHEN DONG[2], AND SHANGTAI JIN[2]

[1]Big Data and Internet of Things Research Center, China University of Mining and Technology, Beijing 221116, China
[2]Advanced Control Systems Laboratory, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Shangtai Jin (shtjin@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61573054.

ABSTRACT Metro passenger flow prediction plays an essential role in metro operation system. Due to characteristics of metro operation system, the station operation state is difficult to be described by the passenger flow at a single station. Thus, a novel attention mechanism based end-to-end neural network is presented to predict the inbound and outbound passenger flow to improve predictive effect. The novel model explores the latent dependency between flow of forecast target station and historical flows from surrounding stations by attention mechanism. The relation between variable length flow lists with respect to target station is represented as a fix length vector by the attention mechanism. Furthermore, a deep and wide structure is presented to deal with the inherent information of each station, which are discretized into high dimensional categorical features. Experiments on Beijing Subway line 5 with 1.8 million samples demonstrate the effectiveness of presented approach, which shown the performance on capturing latent dependency.

INDEX TERMS Attention mechanism, attention based neural network, deep and wide structure, metro passenger flow prediction.

I. INTRODUCTION

In metro operation system, system management such as train operation plan is formulated by history flow data and predicted passenger flow data. Thus, the flow prediction model that offers a robust and accurate result plays a vital role in the metro operation system. In the past decades, passenger flow prediction has been a hot issue in metro operation system, and many results were obtained.

As a pioneer work, historical average [1], smoothing techniques [2], and autoregressive integrated moving average (ARIMA) [3], [4] have been widely applied to forecast in transportation system. Among these parametric prediction methods, ARIMA has become one of the most common forecasting approaches in regression problems and time series problems such as speed, occupancy, passenger flow prediction [5], [6]. Although some improved methods are applied to obtain better performance [7], [28], the results are limited in practice due to the nonlinear traffic data.

In order to solve the problem mentioned above, various machine-learning approaches are applied in forecasting passenger flow to handle non-linear problems. Such as using tree method mining relationship between short-term subway ridership and its influential factors [8], [9] utilizing integrated bayesian approach to forecast complete and incomplete passenger flow data, and also various support vector machine (SVM) methods are proposed for short-term traffic flow forecasting [10]. Furthermore, various end-to-end neural network architecture have been proposed to handle passenger flow predicting problem. Dependence of manual feature engineering can be reduced by using deep neural network architectures. For example, A special structure of deep neural network (DNN) is proposed which can deeply and abstractly extract the nonlinear features embedded in the input without any labels [11]. The SAE model was also improved to adapt different flow distribution (daytime and nighttime) [12]. Considering the temporal characteristics of passenger flow, an LSTM (Long short-term memory) architecture [13] is used to capture the dependences of time series for travel time prediction. In addition, the modified LSTM is designed to forecast traffic flow [14]. The performance of results is greatly influenced by traffic incident, furthermore, the unusual flow caused by traffic incident will affect from one station to stations around. Thus, spatial characteristics

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

are considered in neural network architecture. Polson and Sokolov proposed a deep learning architecture can capture these nonlinear spatial-temporal effects [15]. To take advantage of external environmental features, temporal dependencies, and spatial characteristics a Deep Passenger Flow (Deep PF) model [21] is built, which embed external environmental features by fully connected layers and handle time series data by LSTM cells. Ann and Cluster-Based LSTM are used to increase the effect of passenger predict [25], [26], [29].

However, different metro station has different pattern of flow trend because of the location of metro station, surrounding stations, structure of station and so on. It is necessary to illustrate characteristics of metro station by all available information [27]. Particularly, the latent relationship with surround stations is difficult to capture by LSTM neural network and is hard to deal with time series for Convolutional Neural Network (CNN). Attention Mechanism [20] is utilized to solve this problem. Attention mechanism originates from Neural Machine Translation (NMT) field [16]. NMT takes a weighted sum of all the annotations to get an expected annotation and focuses only on information relevant to the generation of next target word. We attempt to utilize this property to mining characteristics of metro station.

In this paper, we propose an architecture of attention based neural network(DNN-Attention) and apply it to predict the passenger flow for Beijing metro stations. We hope the precise of predicting passenger could be improved. We make three major contributions: First, a Deep FM method based on the discretized high-dimensional discretization metro data is proposed for passenger flow prediction. In order to describe the flow variation characteristics for every different metro station, we not only use the encoded station ID, but also discretize inbound and outbound passenger flow to obtain more smooth data. Second, the dependency of spatial and temporal characteristics formulated by attention mechanism instead of traditional LSTM method. Latent relationship between two different stations is captured from both traffic state of target station and the states list of surrounding stations when predicting outbound flow and inbound flow. Third, the experimental result turns out that the flow variation characteristics of metro station can be formulated in a high accuracy using out architecture of attention based neural network.

The remainder of the paper is organized as follows. Section 2 defines the problems, which this paper is going to solve. Section 3 prerequisites of deep learning models are described. Section 4 presents the details of Architecture of attention based neural network. Section 5 presents the experiential results of prediction model. Finally, this paper is concluded in section 6.

II. PROBLEM DEFINITION

In this paper, the passenger flow of metro stations is forecasted by various characteristics (e.g. inbound/outbound flow of surrounding stations, time information and station encoding information) within 5 min interval. The passengers flow in *i-th* station in $t+1$-*th* time interval can be denoted as $X_i(t+1)$. Moreover, the history flow of *i-th* station can be described as $H_{i,t+1} = \{X_i(t), X_i(t-1)\ldots\}$, and surrounding stations history flow of *i-th* station can be described as $S_{i,t+1} = \{H_{i,t+1,1}, H_{i,t+1,2}, \ldots\}$, where $H_{i,t+1,1}$ denotes the first station adjacent to *i-th* in $t+1$-*th* time interval. Thus, the issue in this paper is using history information of *i-th* station to predict $X_i(t+1)$, which can be formulated as:

$$X_i(t+1) = F[S_{i,t+1}, D_t + 1, O_t + 1]. \tag{1}$$

where $D_t + 1$ contains date characteristics like timestamp, date of the year, weekend or holiday. And $O_t + 1$ includes other related information like up or down order, and station status.

III. PREREQUISITE OF DEEP LEARNING MODELS

In this section, an attention based neural network is employed to deal with the sparsity features and formulate flow dependency among metro stations. We will present a brief review of DeepFM architecture and attention mechanism in neural network.

A. DEEP FM

Deep FM (Factorization-Machine based Deep&Wide) [17] is a powerful end-to-end learning model emphasizes both low and hight order feature interactions. The DeepFM model combines the power of factorization machines and the ability of feature learning in a neural network, and thus has a great hlpreformence in handling sparsity inputs.

As shown in FIGURE 1, there are two important components in DeepFM, FM component and Deep component, which shared same input and same dense embedding layer. For *j-th* discrete field in inputs, the latent matrix w_j is used to weigh its order-1 importance, and the latent matrix V_j is used to weigh its order-2 dependency with other fields. In the process of the implementation, discrete inputs coded in one-hot form. Thus each value in field can be embedded into independent matrix. All the embedding parameters are trained in DeepFM model. The forecast result can be formulated as:

$$X_i = MLP(y_{FM} + y_{deep}). \tag{2}$$

where the y_{FM} and y_{deep} denote the result of FM part and deep part, respectively. In FM (Factorization Machines) component [18], order-1 feature is directly modeled by latent embedding vector, and pairwise order-2 feature interaction is modeled by inner product of respective latent embedding vectors. The output of FM component can be denoted as:

$$y_{FM}(x) = \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < V_i, V_j > x_i x_j. \tag{3}$$

where x_i denotes the *i-th* field in inputs like: $S_{i,t}$,D_t,O_t, and$< \cdot, \cdot >$ denotes the dot product of two matrixs of size k:

$$< V_i, V_j >= \sum_{f=1}^{k} v_{if} v_{jf}. \tag{4}$$

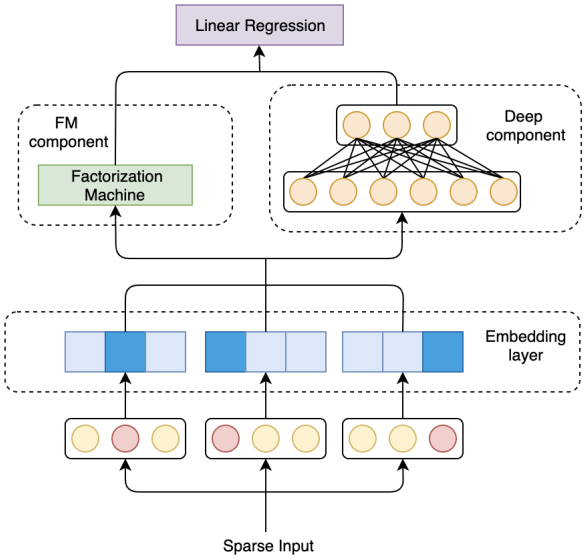


FIGURE 1. The wide & deep structure of Deep FM.

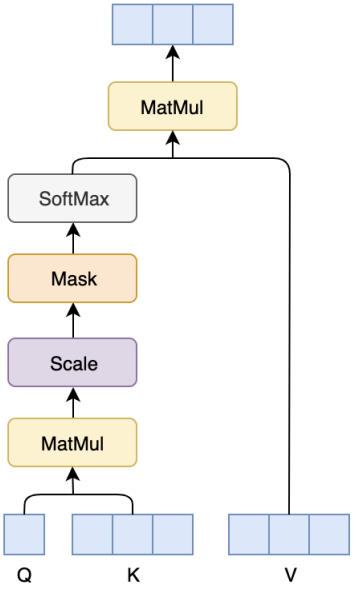


FIGURE 2. Scaled dot-product attention.

$k \in N_0^+$ is the hyper-parameter to restrict the dimension of latent matrix. In deep component, a naive DNN (Deep Neural Network) is used to capture unseen feature interactions through low-dimensional embeddings. The $(l+1)$-*th* hidden layer output a^{l+1} denotes as:

$$a^{l+1} = f\left(W^l a^l + b^l\right). \tag{5}$$

where f is the activation function, ReLU is used in this paper. W^l and b^l are the model weights and bias for l-*th* layer. Finally, the outputs of FM and deep components will be simply concatenated like formula (2).

B. ATTENTION MECHANISM

Attention mechanism has a great performance in many sequence-based tasks [19]. Two major benefits of attention mechanism are taken advantage of in this paper. The first is that the attention mechanism allows for dealing with variable sized inputs. The second is that the location of items in sequence are ignored when an attention mechanism is used to compute a representation of a single sequence.

An attention mechanism is to find the relationship between query and key-value pairs, and to compute the output by a weighted sum of values [20]. The weight assigned to each value is computed by a Softmax function using query with the corresponding key. The structure of a normal attention mechanism is shown in FIGURE 2. The Q, K, V represent the input query, keys and values, respectively. The output of attention mechanism is given as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{6}$$

where the dot product of Q and K reveals the similarity between query with each key. $\sqrt{d_k}$ is the scaling factor to limit the value of dot product, which is the Scale part in FIGURE 2. In Mask part, each padding elements will be multiplied by 0. In this paper, padding elements are missing time point in data.

IV. ARCHITECTURE OF ATTENTION BASED NEURAL NETWORK

In this paper, an end-to-end neural network is presented to explore interactions among all characteristics described in section 2. In addition, the dependency between passenger flow of prediction target station and flow lists of surrounding stations are explored as flow state representation vector by attention mechanism.

A. FIELD REPRESENTATION

Data described in section 2 mostly are with categorical form (e.g. station id, timestamp, up or down order), which can be transformed into high-dimensional sparse binary features via one-hot encoding. Encoded category field with k individual values can be formulated as $p \in R^k$ vector. Thus, $X_i(t+1)$ in section 2 is formulated as:

$$X_i(t+1) = F[S_{i,t+1}, p_D^T, p_O^T] \tag{7}$$

The values in one-hot encoding vector are binary. For example, if the weekday of random sample is Friday, the weekday field of this sample will be encoded as $\{0, 0, 0, 0, 1, 0, 0\}$.

Furthermore, as shown in FIGURE 3, the randomness of short-term (5-min interval) passenger flow causes a problem for list of surrounding flow to formulate similarity with forecast target flow. In order to relieve this effect, we discrete the flows in surrounding lists into categorical features. Furthermore, discretization of continuous features also means increasing the dimension of input data by using embedding vectors representing categorical features. A dynamic discretization strategy is used to weaken the effect of randomness and assign more available training samples for each

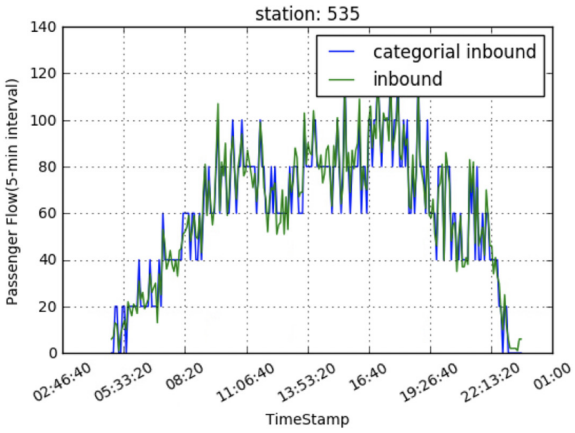
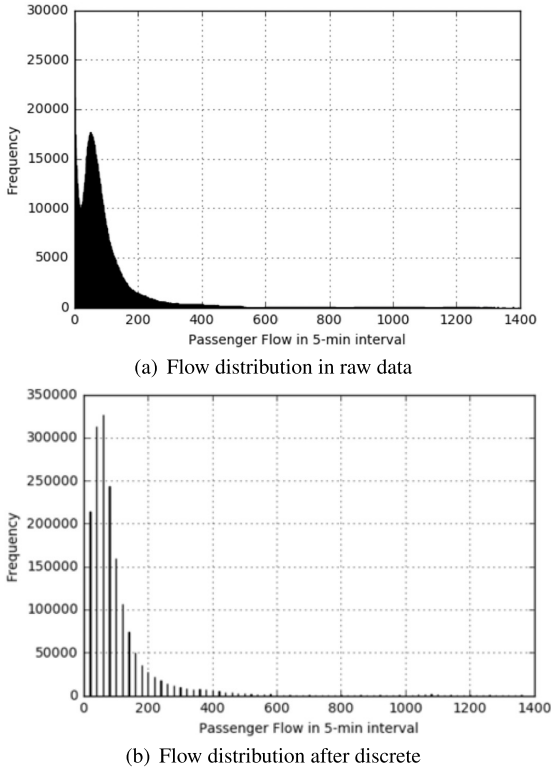


FIGURE 3. Passenger flow of a random station and a random day.



(a) Flow distribution in raw data

(b) Flow distribution after discrete

FIGURE 4. Compare distribution of passenger flow occurrences.

category. In this way, the passenger flows in surrounding flow lists will be treated as category features, in which the list length is the hyper-parameter. The distribution of all samples before and after discretization are shown in FIGURE 4 (a) and FIGURE 4 (b). The horizontal axis represents the passenger flow in 5-min interval, and the vertical axis means the frequency of flow appearance. Notes that there are no manual combination features here, all dependency among features is explored by end-to-end neural network.

B. FORMULATE DEPENDENCY FOR EACH STATION

Traditional LSTM is widely used in existing works to capture the dependency among time sequence [5], but the flow

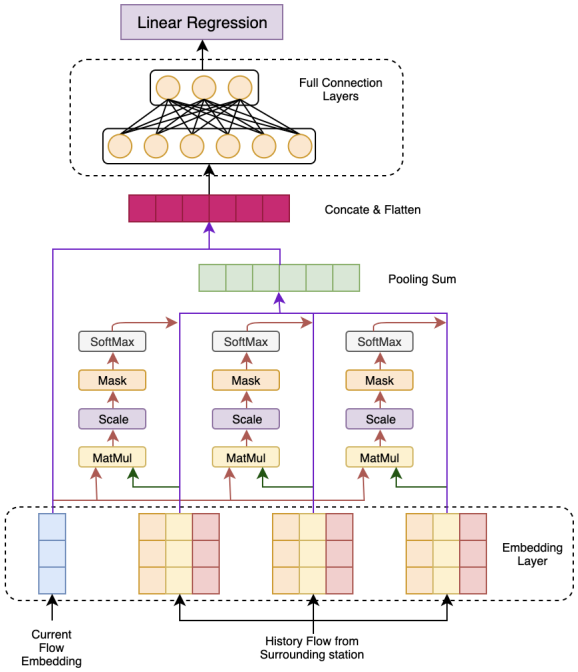


FIGURE 5. Using attention mechanism to capture the correlation between forecast target flows and surround station flow.

correlation between forecast target station and surround stations is not suitable for LSTM. Because different from single time sequence, there is not directly related to position in sequence. For example, the outbound passenger flow at t-th time interval may have stronger correlation with first up order station's flow at t-1 th time interval or second up order station's flow at t-2 th time interval, because of the minimum operating interval. Notes that the regular station's inbound passenger flow is considered as an individual value, which means the flow list of surrounding stations is empty. In FIGURE 5, an attention unit is designed to calculate a weight to perform the relation of a flow from surround station. According to attention mechanism in section 3, the attention unit is formulated as:

$$\begin{aligned} U_{i,t+1}(\bar{X}) &= Attention(\bar{X}_{i,t+1}, \bar{S}_{i,t+1}) \\ &= Attention(\bar{X}_{i,t+1}, \bar{H}_{i,t+1,1}\bar{H}_{i,t+1,2,\ldots,}) \\ &= \sum_{t} +1 \sum_{j=1}^{k} a(\frac{\bar{X}_{i,t+1}, \bar{H}_{i,t+1,j}}{\sqrt{d}})\bar{H}_{i,t+1,j}. \quad (8) \end{aligned}$$

where the $\bar{X}$ and $\bar{H}$ are the embedding vectors of forecast target station flow and flow from surrounding stations. Function a here is a feed-forward network with generating the weight. Notes that the order in sequence will be ignored, and incident from surrounding station will be found by attention mechanism. At last, full connection layer will improve the ability of generalization.

C. FM COMPONENT

Besides the dependency for each station, time features and station's characteristics are also fed into prediction model.

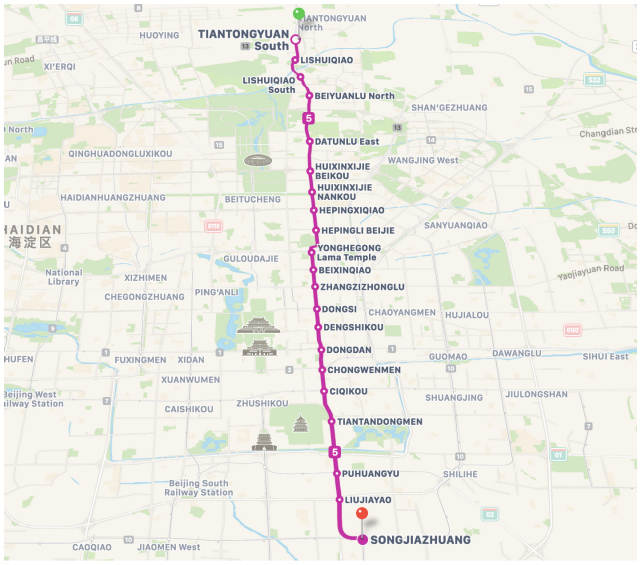


FIGURE 6. Distribution of metro stations.

These categorical features could directly determine station's characteristics. When the embedding vector of flow category are not fully trained, we need the model could predict by station's characteristics. Thus, the FM component are used to capture the missing information from attention mechanism, like predicting normal station's inbound flow.

In summary, the overall framework is shown in FIGURE 7, the output of the attention part is a weighted sum of all input features. Furthermore, all deep component's em-bedding result will be concatenated directly; FM component will deal with the other features like time stamp, date of year, up and down order. At last, FM part and deep component output will be also concatenated to feed into last predicting layer.

V. EXPERIMENTS AND RESULT

In this section, experimental setup details and analyzing results on Beijing Metro System dataset will be presented. Furthermore, the result of our network will be compared with other three baseline models.

A. EXPERIMENTAL SETUP

The distribution of metro stations used in this paper is shown in FIGURE 6. The data set collected from Beijing Metro System is used to verify the performance of architecture of attention based neural network. The data recoded the inbound/outbound passenger flow for Metro Line 5 from January to December 2017. After drop the samples not in operating time, the whole data set consists 1854994 samples. We use all data before December 1-*st* as training data set, and data in December as testing data set. In practice, 5-min interval and 5-min interval will be both tests. For our neural network, non-category features will be normalized at the same time for training and testing data set. For all the FC (fully connection) lays in our architecture,

TABLE 1. Performance of Beijing metro system passenger flow data.

Model	RMSE		MAE	
	inbound	outbound	inbound	outbound
DNN	45.46	61.05	21.03	31.07
XGBoost	36.50	43.84	19.05	24.30
DeepFM	32.05	40.13	17.85	23.29
DNN-Attention	27.34	37.50	16.63	23.38

BN (batch normalization) and dropout technic will be both tests to prevent over-fitting. For all test models and our structure, we use Adam as the optimizer and mini-batch size is set to be 32. In this section, performance will be compared with three different baseline models:

DNN (Deep Neural Network): naive feed-forward neural network without any additional technique.

XGBoost [22]: a scalable tree boost system, which is an improvement in GBDT (Gradient Boost Decision Tree) [23], which is widely used in prediction and classification tasks. Because of the characteristic of tree method, XGBoost have a great performance in task with continuous features. Thus, this method will be fed by samples without discretization.

DeepFM(Factorization-Machine based Deep&Wide): combine the factorization machine and DNN. The deep and wide structure and embedding ideas [24] are designed for high dimension categorical input.

In passenger flow prediction field, RMSE (Root Mean Square Error), MAE (Mean Absolute Error) are widely used measures to evaluate model. We adapt these measures in our experiments, which can be formulated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{9}$$

$$MAE = \frac{1}{N}\sum_{i}^{N}\left|\hat{y}_i - y_i\right|. \tag{10}$$

where the y_i is the observed passenger flow, and $\hat{y}_i$ is the forecast value of prediction models for *i-th* sample.

B. RESULTS

Table 1 shows the inbound and outbound results of data set collected from Beijing Metro System, which predict by various prediction methods. From the table, we can ob-serve the RMSE of inbound passenger flow in validation data are 45.46, 36.50, 32.05 and 27.34 respectively, and RMSE of outbound flow are 61.05, 43.84, 40.13 and 37.50 respectively. The same results shown in MAE results. Making com-parisons among models, we report several conclusions. First, the results of DNN-Attention method achieved best performance in both inbound and outbound passenger flow compared with all the DNN, XGBoost and DeepFM. DNN-Attention method achieves 27.34RMSE, 16.63 MAE for predicting inbound flow and 37.50 RMSE, 23.38 MAE for outbound flow. It validates DNN-Attention architecture can describe the station flow trend better then compared

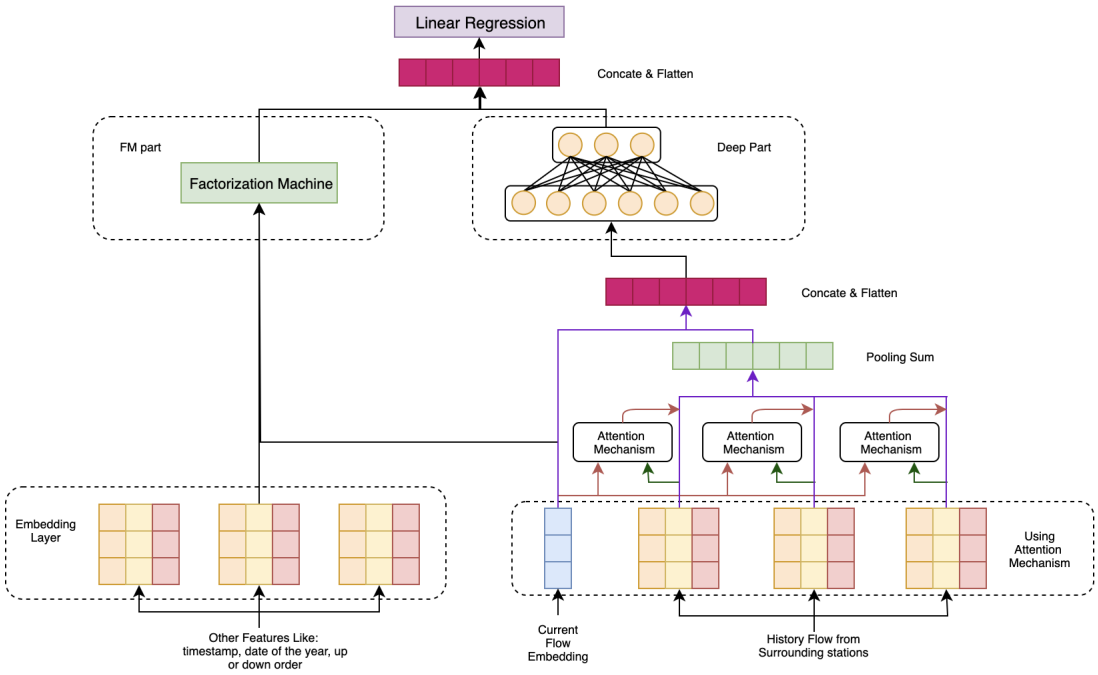


FIGURE 7. Framework of DNN-Attention.

TABLE 2. Performance of outbound passenger flow.

Model	Regular Station		Transfers Station		Weekdays		Weekends	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
DNN	72.84	36.76	44.18	25.65	69.90	35.97	30.71	19.31
XGBoost	53.40	29.72	32.10	20.77	49.00	26.91	27.74	18.06
DeepFM	49.00	28.72	30.25	20.76	44.35	25.46	27.46	18.10
DNN-Attention	48.02	27.56	29.09	19.32	40.86	25.12	26.83	17.20
Increase percentage(%)	2	4.03	3.83	6.93	7.86	1.33	2.29	4.97

TABLE 3. Performance of inbound passenger flow.

Model	Regular Station		Transfers Station		Weekdays		Weekends	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
DNN	54.11	25.05	43.43	18.43	52.32	24.36	21.37	13.96
XGBoost	42.90	22.04	29.18	16.94	41.42	21.31	20.25	13.62
DeepFM	37.25	20.40	26.57	16.21	35.95	19.66	19.72	13.50
DNN-Attention	33.35	19.86	26.36	16.67	30.07	18.07	19.28	13.17
Increase percentage(%)	10.46	2.64	0.79	-2.83	16.35	8.08	2.23	2.44

methods. Second, regarding forecasting outbound passenger flow, DNN-Attention result remarkably improved, which proved attention mechanism, can formulate the latent dependence among surrounding stations. Otherwise, the external information is poor to predict inbound passenger flow, thus model performance will be attenuated to normal DeepFM. We considering the reason of inbound flow's RMSE lower then outbound flow is the training data only contains one subway line. Thus, the affect will be brought by other subway lines.

Metro operation has various distinctive characteristics, for example, running interval, transfers stations and change of OB matrix. The model ability to handle different distribution passenger flow are shown by Table 2 and Table 3. First, the reason of insignificant improvement is similar to the above analysis. Flow data from one subway line cause the lack of information from surrounding station. However, from the temporal characteristics, our architecture can also obtain basic performance. Second, comparing results of weekdays and weekends, the significant improvement of performance is caused by attention mechanism. Rapidly fluctuation from surrounding stations in weekends will be reflected by attention mechanism.

In summary, feeding discretizing data to deep neural network and utilizing attention mechanism to mining latent dependency of spatial and temporal characteristics can accurately describe station characteristics and timely respond to sudden flow.

VI. CONCLUSION

In this paper, we focus on the task of metro passenger flow prediction modeling by data set from Beijing Metro System. Using traditional time series methods has become a bottleneck for capturing description of flow trend. To obtain better performance of forecasting method, an attention based deep neural network architecture is used to capture the latent relationship of spatial and temporal characteristics and to describe with discrete data, which can be easily generalized to the entire subway network. The result shows that our architecture obtain better performance: Inbound flow increased average 14.41% RMSE, 6.83% MAE, outbound flow increased average 6.55% RMSE.

This paper is an initial research to describe passenger flow by end-to-end neural network. To better capture the associate of spatial and temporal characteristics, it is necessary to mine more additional affect factors in future work.

REFERENCES

[1] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *J. Transp. Eng.*, vol. 123, no. 4, pp. 261–266, Jul. 1997.

[2] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.

[3] J. V. Hansen, J. B. McDonald, and R. D. Nelson, "Time series prediction with Genetic-Algorithm designed neural networks: An empirical comparison with modern statistical models," *Comput. Int.*, vol. 15, no. 3, pp. 171–184, 1999.

[4] S. Lee, Y. Lee, and B. Cho, "Short-term travel speed prediction models in car navigation systems," *J. Adv. Transp.*, vol. 40, no. 2, pp. 122–139, 2006.

[5] M. M. Hamed, H. R. Al-Masaeid, and Z. M. Said, "Short-term prediction of traffic volume in urban arterials," *J. Transp. Eng.*, vol. 121, no. 3, pp. 249–254, 1995.

[6] S. Lee and D. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transp. Res. Rec.*, vol. 1678, no. 1, pp. 179–188, 1999.

[7] Y. Jia, P. He, S. Liu, and L. Cao, "A combined forecasting model for passenger flow based on GM and ARMA," *Int. J. Hybrid Inf. Technol.*, vol. 9, no. 2, pp. 215–226, Feb. 2016.

[8] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustainability*, vol. 8, no. 11, p. 1100, Oct. 2016.

[9] Z. Zhu, B. Peng, C. Xiong, and L. Zhang, "Short-term traffic flow prediction with linear conditional Gaussian Bayesian network," *J. Adv. Transp.*, vol. 50, no. 6, pp. 1111–1123, Oct. 2016.

[10] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, Oct. 2015.

[11] L. Liu and R.-C. Chen, "A novel passenger flow prediction model using deep learning methods," *Transp. Res. C, Emerg. Technol.*, vol. 84, pp. 74–91, Nov. 2017.

[12] Y. Duan, Y. Lv, and F.-Y. Wang, "Performance evaluation of the deep learning approach for traffic flow prediction at different times," in *Proc. IEEE Int. Conf. Service Oper. Logistics, Informat. (SOLI)*, Jul. 2016, pp. 223–227.

[13] Y. Duan, Y. Lv, and F. Y. Wang, "Travel time prediction with LSTM neural network," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1053–1058.

[14] Y. Tian and L. Pan, "Predicting short-term traffic flow by long short-term memory recurrent neural network," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 153–158.

[15] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[17] H. Guo and R. Tang, "DeepFM: A factorization-machine based neural network for CTR prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell*, 2017, pp. 1725–1731.

[18] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2010, pp. 995–1000.

[19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] Y. Liu, Z. Liu, and R. Jia, "DeepPF: A deep learning based architecture for metro passenger flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 101, pp. 18–34, Apr. 2019.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[24] H. T. Cheng, L. Koc, and J. Harmsen, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, pp. 7–10, 2016.

[25] M. Gallo, G. De Luca, L. D'Acierno, and M. Botte, "Artificial neural networks for forecasting passenger flows on metro lines," *Sensors*, vol. 19, no. 15, p. 3424, Aug. 2019.

[26] J. Zhang, F. Chen, and Q. Shen, "Cluster-based LSTM network for short-term passenger flow forecasting in urban rail transit," *IEEE Access*, vol. 7, pp. 147653–147671, 2019.

[27] J. Wang, X. Kong, W. Zhao, A. Tolba, Z. Al-Makhadmeh, and F. Xia, "STLoyal: A spatio-temporal loyalty-based model for subway passenger flow prediction," *IEEE Access*, vol. 6, pp. 47461–47471, 2018.

[28] N. O. Alsrehin, A. F. Klaib, and A. Magableh, "Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study," *IEEE Access*, vol. 7, pp. 49830–49857, 2019.

[29] J. Guo, Z. Xie, Y. Qin, L. Jia, and Y. Wang, "Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM," *IEEE Access*, vol. 7, pp. 42946–42955, 2019.

JUN YANG received the bachelor's and master's degrees from Harbin Engineering University, Harbin, China, in 2000 and 2003, and the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2014. He is currently a Professor Level Senior Engineer with the China University of Mining and Technology, Beijing, and the Head of the Big Data and Internet of Things Research Center. His current research interests include big data, the Internet of Things, and intelligent transportation systems.

XUCHEN DONG received the bachelor's degree from the Kunming University of Science and Technology, Kunming, China, in 2014. He is currently pursuing the master's degree with Beijing Jiaotong University. His research interests include short-term traffic flow prediction and recommendation systems.

SHANGTAI JIN received the bachelor's, master's, and Ph.D. degrees from Beijing Jiaotong University, Beijing, China, in 1999, 2004, and 2009, respectively. He is currently an Associate Professor with Beijing Jiaotong University. His current research interests include model-free adaptive control, data driven control, learning control, and intelligent transportation systems.