Prediction of short-term passenger flow of subway based on LSTM model

Jiahan Hu

College of Transportation Engineering, Tongji University, Shanghai, China

2152981@tongji.edu.cn

Abstract. With the continuous development of urban rail transit, the subway is becoming the travel choice for more and more people. And the increase in travel demand also brings about the problem of subway congestion. This paper will use the subway card swipe data of Hangzhou City from January 1st to 20th, 2019, which were counted after data cleaning. With root mean square error (RMSE), mean absolute percentage error (MAPE), coefficient of determination (R^2), and loss values as standards, the long short-term memory's (LSTM) parameters were adjusted. The LSTM model is trained with time and passenger flow data, and the model is used to predict the short-term passenger flow in and out of subway stations. From the prediction results, the LSTM model has a good adaptability to the subway station short-term passenger flow. The trend of the predicted line is the same as the real value, and the difference is also controllable. The short-term passenger flow prediction of subway stations is beneficial to reduce the loss caused by congestion and improve the travel efficiency of residents.

Keywords: Metro, LSTM, Data Cleaning.

1. Introduction

As a kind of rail transportation, the subway has been paid attention to by many countries for its convenient, fast and large volume of traffic. The development of the subway system began in the 19th century. As the subway has grown in sophistication, it has increasingly replaced other modes of urban transportation as people's first option. As of February 2023, Hangzhou Metro has 12 operating lines and 260 stations. On April 28, 2023, Hangzhou Metro ushered in the highest daily passenger flow of the line network, 4,616,400 people.

While the subway brings convenience to urban residents, it also faces many difficulties and challenges. One of the most prominent problems is crowding. Crowding will not only reduce the travel efficiency and travel experience of travelers but also bring economic losses. Although the limited resources of the subway are one of the reasons for the congestion, it is important to note that the utilization rate of some sections is low. If only by increasing the number of subway lines to alleviate the congestion problem will bring many problems [1]. For example, higher infrastructure spending and waste of resources during off-peak periods. Therefore, it is necessary to seek improvements in management. Among them, predicting the short-term subway passenger flow is crucial. This can help metro authorities understand travel demand and optimize schedules [2]. At the same time, this can also help travelers optimize travel decisions and improve travel efficiency [2].

In the past few decades, the prediction of short-time traffic has attracted the attention of researchers around the world. As the research progresses, different types of methods and models are developed. These techniques can be categorized into three groups: naive methods, parametric methods and non-parametric methods [3]. Deep learning has been a research hotspot in recent years and is widely used in various scientific research fields. The application of deep learning in short-term passenger flow prediction of subways has been relatively mature, and many targeted models and methods have been formed [4]. Different research methods and models are combined and improved during the research process. Their prediction accuracy has been greatly improved compared to the original model. GM-RBF neural network prediction model combines the characteristics of gray model (GM) with less sample cloth demand and radial basis function neural network (RBF) with good adaptive ability [5]. The CNN-GRU combination model combines the convolutional neural networks model (CNN) and the gate recurrent unit model (GRU) and synthesizes the influence of time factors and space factors on passenger flow [6]. Another study used long short-term memory neural networks (LSTMNN) with stacked autoencoders (SAEs) to create the best prediction model with compressed data of various dimensions. [7]. It can be seen from these research examples that it is a trend of traffic forecasting to build a mixed model using different advantages of different models.

It is possible to think of the prediction of subway passenger flow as a time series problem that evolves over time [2]. Long short-term memory (LSTM) can adapt to the random and nonlinear characteristics of subway passenger flow [8]. The gradient disappearance issue in recurrent neural network (RNN) can also be solved by LSTM [9].

This paper will refer to the existing research, based on LSTM to forecast the subway passenger flow. The data used for training the model are the swipe card data of Hangzhou City from January 1 to 20, 2019. After the models have been constructed, their predictive power will be assessed using charts and evaluation metrics. Through the study of this paper, it can provide a reference for passenger travel and the management of subway stations.

2. Method

2.1. Data source

The data used in this paper are the subway card swipe data records of Hangzhou City for a total of 20 days from January 1 to 20, 2019. The data include 81 stations on three subway lines. The source of the data is the Tianchi Big Data Competition.

2.2. Index selection and description

The following table shows the 7 variables included in the data (Table 1). This data's information is comparatively extensive and can give a reliable representation of the inbound and outbound traffic of passengers at each station. During the research process, transaction type, card number and transaction date were used for data cleaning. After type filtering, sequential logic filtering and missing filtering, the clean data can be used for machine learning. The primary inputs utilized in training the model are the inbound and outgoing passenger flow statistics for each station.

Table 1. Name and definition of the variable.

Variable Name	Variable Definition	Data Type
time	Transaction date time	Time character
lineID	Line number	Character string
stationID	Station number	Character string
deviceID	Device coding	Character string
status	Type of card used	Character string
userID	Card number	Character string

Table 1. (continued).

payType	Payment type	Character string

2.3. Model Introduction

LSTM is a variant of RNN for processing sequence data, which is specifically designed to solve the problems of disappearing gradients and exploding gradients present in traditional RNN. RNN has a gradient problem when processing sequential data, which results in performance degradation on long sequences. LSTM solves this problem by introducing a special internal department control mechanism.

The core of LSTM is the memory unit, also known as the LSTM unit. It has a cell state and a hidden state. This cell state transmits information between different time steps, allowing the LSTM to capture long-term dependencies.

The LSTM consists of three gated units that determine how the flow of information flows between the cell state and the hidden state. The Forget Gate determines what information is retained in the cell state and what information is discarded. It produces an output between 0 and 1 via a Sigmoid function, which is used to control retention or forgetting of information. The Input Gate controls how fresh information is added to the cell state. It uses a Sigmoid function to produce an output between 0 and 1, and a tanh function to produce a new candidate. The Output Gate chooses how much data should be derived from the cell state. It produces an output between 0 and 1 via a Sigmoid function, passes the cell state to the tanh function for compression, and then multiplies it with the output of Sigmoid to produce the final hidden state. The following diagram shows the internal structure of LSTM (Figure.1) [8].

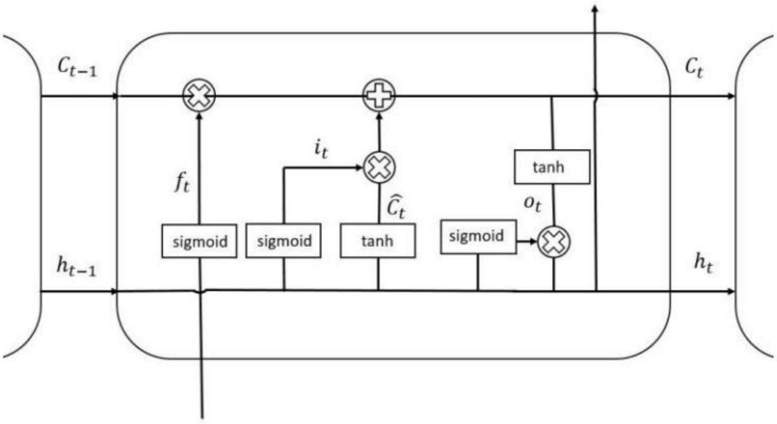

Figure 1. The internal structure of LSTM [8].

At each time step, the LSTM receives hidden states and cell states from the previous time step and updates these states based on the output of the gated unit. These states are then passed on to the next time step. LSTM is designed to capture long-term dependencies effectively, making it useful when dealing with tasks that require understanding context and time dependencies.

2.4. Indicators introduction

This paper used mean square error (MSE) as loss during model training. The MSE between the predicted value and the actual value directly represents the model's predictive power. At the same time, the MSE loss function is usually convex and there is a unique global minimum. Its graph is convenient for observing the change of prediction error during training.

Root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2) were used as evaluation indicators for the model. The model's prediction error is measured intuitively by the RMSE statistic, which averages the difference between the model's predicted value and the actual value. MAPE measures the percentage prediction error of the model, making it easier to explain the overall effect of the model's prediction. R^2 reflects how well the model can explain changes in the observed data.

3. Results and discussion

3.1. Data description and filtering

Prior to training the model, the model must be cleaned to confirm the validity of the data. First of all, the data used is checked for gaps or missing values. No vacancy value was found in the data, which indicates that the data is relatively clean. Secondly, dealing with duplicate values in the data. Duplicate values were handled by preserving unique values. Thirdly, checking the timing logic in the data to make sure the data is logical. In order to improve the processing efficiency and ensure that there is no wrong data, the data that violates the timing logic was deleted. Last of all, the scatter plots were used to check whether there were outliers in the data. For example, Figure.2 shows the scatter plot of credit card type at site 9 on January 1st. As can be seen from the figure, there are no discrete points, and the data are standardized, so the next statistical operation can be carried out.

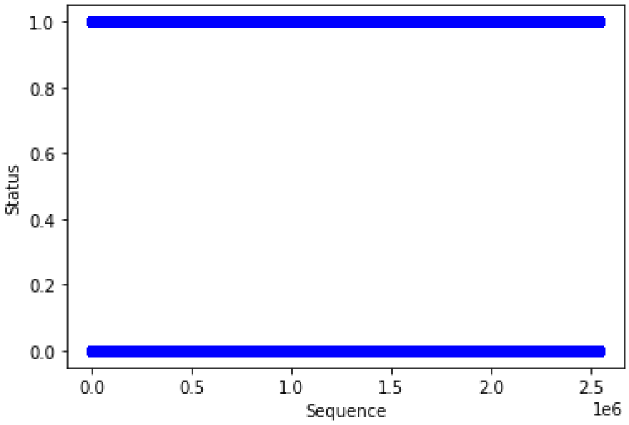

Figure 2. Scatter plot of credit card type at site 9 on January 1st.

After data cleaning was complete, the data was collected by the site. Figure.3 shows the passenger flow of each station on January 1st. It can be seen that the passenger flow of station 9 and station 15 are more prominent than other stations'. In order to ensure there were sufficient data for training, data from these two sites were selected for model training.

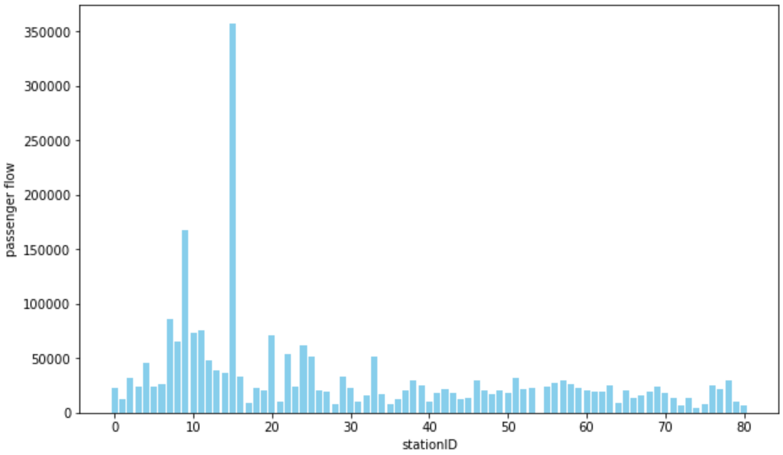

Figure 3. The passenger flow of each station on January 1st.

The data were collected at intervals of ten minutes to obtain the passenger flow data of the two stations from January 1st to 20th. Figure.4 shows the change in passenger flow at Station 9. It can be seen that the passenger flows of stations have regularity and change. The passenger flow on January 1st was significantly higher than on other days, which was consistent with the fact that that day was a holiday. Special values should be left out to verify the model's ability to fit data. The data from the 1st day was thus eliminated from the subsequent training procedure, which used the data from the subsequent 19 days instead.

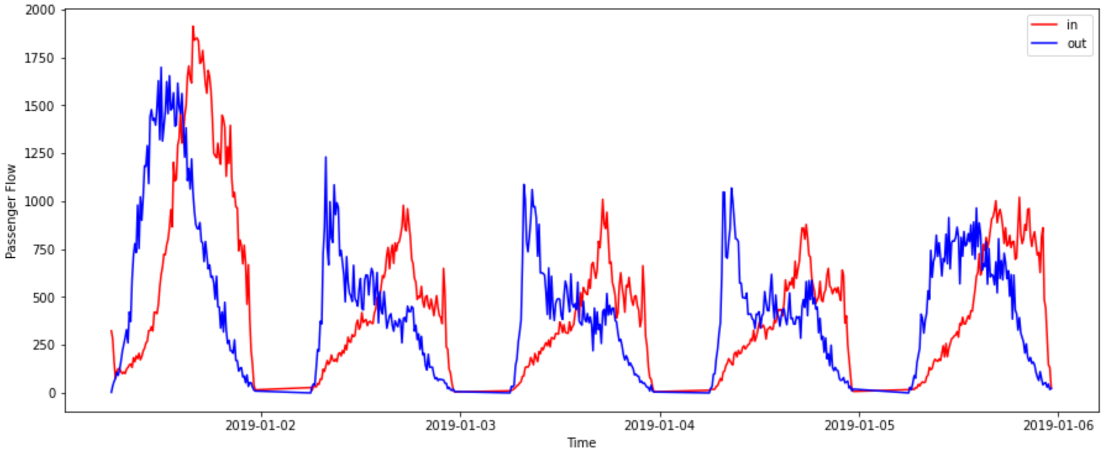

Figure 4. The passenger flow of station 9 in 5 days.

After data cleaning was completed and the range of data was selected, 19-day passenger flow data of Station 9 and Station 15 were obtained. Figure.5 shows the comparison of passenger flow in and out of station 9 on January 2nd. The model was trained with the data from January 2nd-19th and tested with the data from January 20th as the validation set.

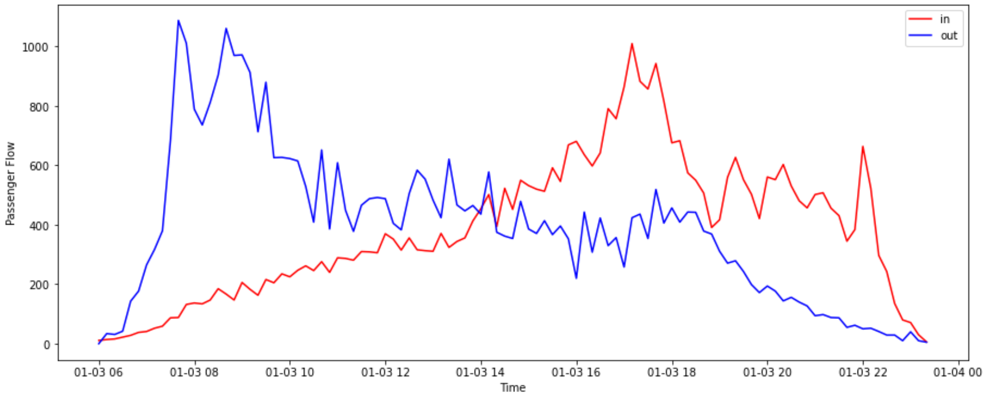

Figure 5. Comparison of passenger flow in and out of station 9 on January 2nd.

3.2. *Selection of parameters*

The LSTM model used in this paper set the batch number to 1 and also needed to determine the quantity of training times and the quantity of neurons. The quantity of training times is not the more the better, too many training times may lead to over-fitting and reduce training efficiency. Using MSE as loss, Figure.6 shows the change of loss during 300 training sessions of 32 layers of neurons. The image illustrates how, at the start of training, the loss value decreased rapidly with the increase in training times, and then the decline rate slowed down, and finally, the loss value increased with the increase in training times.

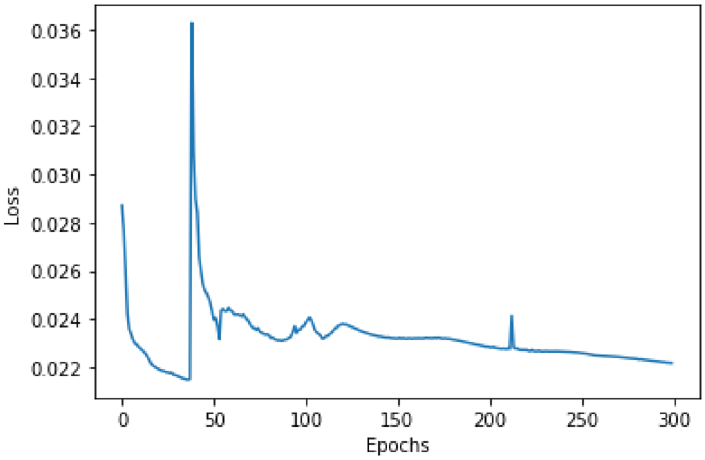

Figure 6. Change of loss during 300 training sessions of 32 layers of neurons.

Besides the number of training times, the setting of the number of neurons will also have an impact on the training results. According to Table 2, the training results of different neuron numbers at 100 training times. The criteria used are RMSE, MAPE and R^2.

Table 2. The training results of different neurons numbers at 100 training times.

number of neurons	RMSE	MAPE(%)	R^2
2	89.1305	35.198	0.9658
4	83.7301	32.942	0.9697
8	82.6444	38.890	0.9706
16	82.4281	20.361	0.9707
32	81.7522	16.981	0.9712
64	84.0113	16.664	0.9696

As can be seen from Table 2, all three parameters perform better when the neurons are 32. Although the MAPE was smaller when the neurons were 64, its RMSE and R^2 were both worse than when the neurons were 32. Considering the three criteria comprehensively, the number of neurons in the model was determined to be 32. At the same time, by observing Figure 6, it can be found that the minimum loss occurs when the training times are 33. Therefore, 33 is selected as the suitable training number for the inbound model at station 9. Through the same method, the suitable training times of the four models can be obtained, as shown in Table 3.

Table 3. The training times of the four models.

Station ID	Inbound or outbound	Suitable training times
9	Inbound	33
9	Outbound	9
15	Inbound	49
15	Outbound	32

3.3. *Result analysis*

After determining the parameters of the LSTM models, the models were trained using the data from January 2nd to 18th and verified with data from January 20th. The forecast results are shown in Figure.7 and Figure.8.

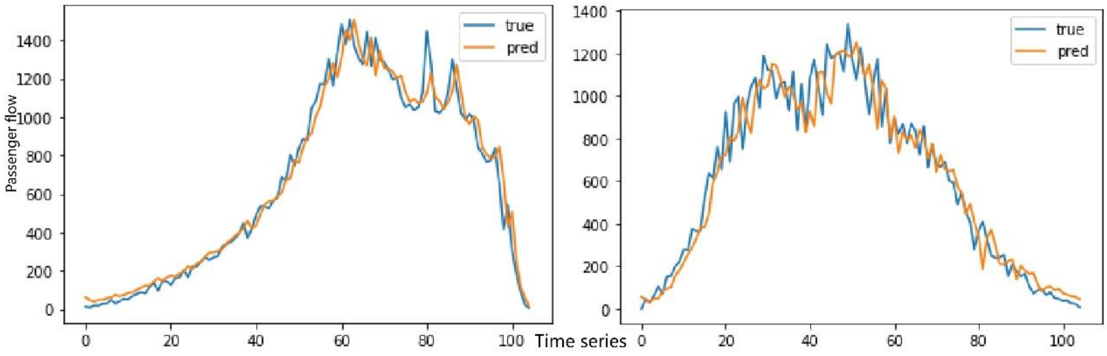

Figure 7. Predicted results of station 9 (left for inbound and right for outbound).

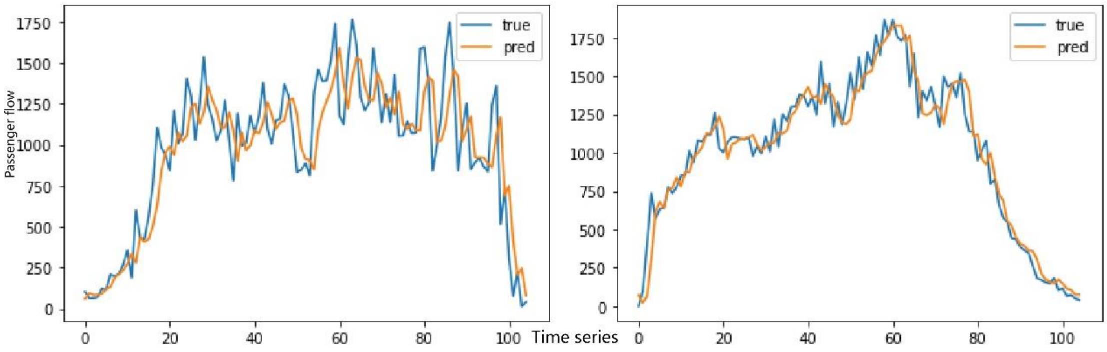

Figure 8. Predicted results of station 15 (left for inbound and right for outbound). Predicted results of station 15 (left for inbound and right for outbound).

As shown in the figures, the forecast results of the four models differ. Their criteria are shown in Table 4.

Table 4. The four models' criteria.

Station ID	Inbound or outbound	RMSE	MAPE (%)	R^2
9	Inbound	82.6097	28.535	0.9706
9	Outbound	106.0249	inf	0.9326
15	Inbound	228.5114	38.655	0.7407
15	Outbound	123.8333	inf	0.9363

It can be seen from the results that the predicted results of the four LSTM models are different under the condition of setting the same number of neurons. Intuitively, this model predicted the inbound passenger flow of station 9 best. From the three criteria, the effects of the four models are also different. Among them, the reason why the MAPE of outbound passenger flow predictions is infinite (inf) is that there is a value of 0 in the real value, which leads to infinite MAPE calculation. In addition, it is apparent from the prediction outcomes that the LSTM forecast line will move to the right. This is consistent with the lag of LSTM's prediction [10].

All in all, the four anticipated passenger flow trends are compatible with the trend of the real value, which can reflect the change in passenger flow with time correctly. However, some models can not accurately predict the specific value of passenger flow, such as the peak value of passenger flow. There are several reasons for the differences between the model's predictions. First of all, the LSTM model has different adaptation effects to different data. Secondly, the selection criteria of parameters are different. In this paper, the three criteria and loss values were used to select parameters. Different selection criteria can lead to different results. Thirdly, although the passenger flow of subway stations is regular, some special circumstances will lead to special passenger flow, which will cause different deviations in the models. Lastly, there is randomness in the machine learning algorithm, even if all the parameters are the same, the resulting prediction results will be different.

4. Conclusion

This paper-trained LSTM and used the model to predict passenger flow. According to the findings, the LSTM model has a good predicting impact on the volume of passengers in subway stations. With RMSE, MAPE, R^2 and loss values, the LSTM model's parameters can be changed to reflect the current scenario. The models can depict the trend of passenger flow accurately. This demonstrates that the LSTM model can accurately forecast data which is strongly tied to time.

At the same time, because the model is single, the LSTM model has certain differences in the predictions of different data. Combining the LSTM model with other machine learning methods, a better passenger flow prediction model can be obtained. All in all, the LSTM model has a significant impact on the field of predicting subway passenger flow, which benefits both residents' travel experiences and the effectiveness of urban rail transit operations. Higher prediction accuracy will be achieved by the upcoming LSTM-based passenger flow prediction model, which has a guiding significance for reducing congestion and improving efficiency.

References

[1] Ding X et al 2018 *J.Eur. Transp. Res. Rev.* 10(2) pp 1-11
[2] Liu Y, Liu Z and Jia R 2019 *J.Transp. Res. Part C: Emerg. Technol.* 101 pp 18-34
[3] Xie C et al 2021 *J. Adv. Transp.* 2021 pp 1-8
[4] Zhang S 2021 *Lanzhou Jiaotong University*
[5] Chen Y et al 2021 *The world transportation engineering technology BBS (WTC2021)* (People's traffic press co., LTD) p 9
[6] Luo J 2023 *Guangdong University of Technology*
[7] Jia F et al 2019 *IET Intelligent Transport Systems* 13(11) pp 1708-16
[8] Chang H 2023 *Xijing University*
[9] Yang L et al 2018 *J. Comput. Appl.* 38(S2) pp 1-6+26
[10] Huang T and Yu L 2019 *J.CEA.* 55(01)142-148