Prediction of Traffic Flow at Urban Intersections using LSTM Model

Haiyi Hu *

School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, 510000, China

* Corresponding Author Email: 202130200200@scut.edu.cn

Abstract. The growth of population and economy is resulting in an increase in the usage of cars in cities, which is causing issues. Traffic congestion has become a normal phenomenon. To better deal with the problem, it is of great significance to forecast the traffic flow. This problem is addressed by proposing the use of the LSTM model in this paper. Using the data from the Shanghai Public Data Open Platform, two sets of data are chosen to do the experiment. The data was collected by loop and accurate to second, which makes it pretty unstable. In this situation, the LSTM model still has a good result. The mean absolute error of Loop 1, Loop 4, and Loop 9 prediction is 3.579, 6.515, and 4.253. R^2 all reaches over 0.8. Furthermore, the LSTM model can add more features to improve its accuracy to fit future traffic conditions. Through this research, it is concluded that the LSTM model is suitable for predicting time-series statistics and performs a high accuracy.

Keywords: Traffic flow prediction; LSTM; intersections.

1. Introduction

With the development of cities and population growth, the problem of urban traffic congestion is constantly exacerbating. Traffic congestion not only increases travel time but also energy consumption and carbon emissions. To study more efficient traffic management strategies, more reasonable allocation of traffic resources, and provide better traffic services, especially, traffic flow prediction is of great significance for Intelligent Transportation System (ITS).

In early researches, many statistical methods were used in solving this problem, for example, Historical Average (HA) [1], Autoregressive Integrated Moving Average Model (ARIMA) [2], Vector Autoregressive Model (VAR) [3], etc. However, these methods normally should follow some hypothesis to have a better result. Nowadays, the traffic condition is more complex. These methods are no longer suitable for solving real traffic problems.

In the past ten years, benefiting from the rapid development of artificial intelligence, machine learning has become a popular method in this issue. Hong et al proposed a three-stage framework to predict short-term traffic flow, their K-Nearest Neighbor (KNN) model with the Least absolute shrinkage and selection operator (Lasso) outperforms the traditional one [4], Duan used the particle swarm optimization (PSO) algorithm to select parameters to achieve optimal Support Vector Machine Model (SVM), finding it superior than neural network training and SVM model [5]. Machine learning methods can construct models to address more complicated data, but they may not provide accurate predictions when addressing a large amount of data.

As deep learning is being utilized in a variety of fields, researchers are applying this method in traffic prediction, like Convolutional Neural Networks (CNN), Ma et al. used CNN to predict transportation network speed [6], Azad and Islam used Long Short-Term Memory (LSTM), they found the average relative error is between 8.25% -14.09% and the accuracy remains stable when it is freeway and the traffic flow are identical [7]. Cheng et al. found a Gate Recurrent unit (GRU) network based on the Improved Particle Swarm Optimization (IPSO) algorithm outperforms the other two manually adjusted parameter models [8]. GRU and LSTM are basically used in time series prediction, and CNN considers the connection of spatial parts. Zhang et al combine both methods to get a more accurate prediction (CNN-LSTM) [9, 10]. Also, the methods combined with some algorithms can consider more factors and get a more accurate result that can best fit the changeable

and complex road conditions nowadays. Lan et al built an improved model Cuckcoo Search (CS) - SVM used in short-term prediction got a smaller error from the actual value. The CS-SVM model is found to have a lower prediction error and better stability than other traditional models, as evidenced by the results [11].

The LSTM model is used in this paper to analyze traffic flow data in Shanghai. The objective is to demonstrate the accuracy of this method and compare different types of data sets to find its advantages.

2. Methods

2.1. Data Source

This study utilizes the objective and accurate data source of an intersection's 5-minute traffic flow data in Shanghai from Shanghai Public Data Open Platform from November 1, 2017, to November 15, 2017. The data was collected by loops, containing the vehicle number every second. Some data is missing on some days and every 5 minutes. As a result, there are only 288 statistics.

2.2. Indicator Selection and Description

The original data contains different types of cars, before the experiment the author conducted preliminary processing on it. The author added all the types of vehicles together and named it Total. Also, all the data was collected in Node 1. Node ID is not needed in this study. The variables used in this study are shown in Table 1 with their types and explanations. Loop ID provides different sources of data. Date ID and Time ID offer time series data. The data was accurate to seconds, making the line in the graph fluctuate greatly. Such unstable statistics will affect the accuracy of prediction. Thus, the author added every two seconds vehicle numbers together, to simplify the data.

Table 1. Name and explanation of variables

Name	Data type	Explanations
Loop ID	INT	Collected by which loop
Date ID	Object	The date when data was collected
Time ID	INT	Every second in 5 minutes
Total	INT	The total vehicle number

2.3. Method Introduction

The basic LSTM structure is shown in Figure 1, the introduction of memory cells, input gates, output gates, and forget gates allows LSTM to effectively solve long sequence problems. The initial stage of LSTM is to determine which information should be discarded. This is achieved by what is called "forget gates". It receives h_{t-1} and x_t, inputting them into σ (Sigmoid). Then outputting a vector f_t:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

Its value is between zero and one. Zero means we totally abandon it. One means the opposite (totally accept it). Then comes to the input gate. There are two parts here. In part I, the information goes through σ and outputs i_t to identify which information will be upgraded.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

In the next stage, a new vector $\tilde{C}_t$ is created through $\tan h$. It can be used in the next stage.

$$\tilde{C}_t = \tan h\,(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

Then, the cell C_{t-1} can be upgraded to C_t.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

In the final stage, the output value will be confirmed in the output gate. First, information will also go through the σ and output o_t:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

With that vector, through the under-processing, the output value h_t can be calculated.

$$h_t = o_t * \tan h\,(C_t) \tag{6}$$

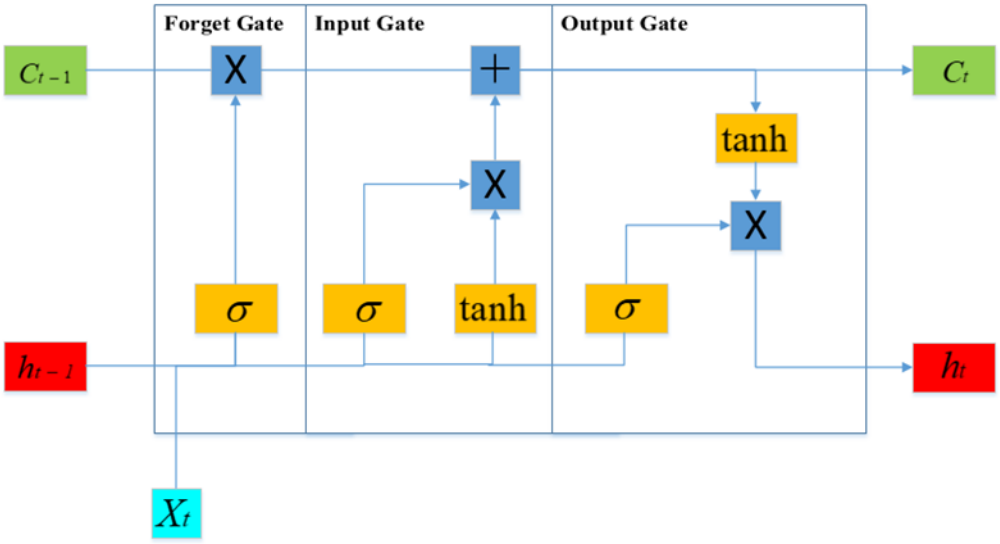

Fig. 1 LSTM structure

2.4. Evaluation Index

In this model, accuracy is evaluated using three criteria, containing Mean absolute error (MAE), Mean-square error (MSE), and R^2. The formula below shows how the parameters are calculated.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y_i| \tag{7}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y_i|^2 \tag{8}$$

$$R^2 = 1 - \frac{\sum_i(\hat{y}_i - y_i)^2}{\sum_i(\bar{y}_i - y_i)^2} \tag{9}$$

3. Results and Discussion

3.1. Data Preprocessing

The data contains 18 loops' statistics, to have a better result, the author chooses the data from Loop 1, Loop 4, and Loop 9. Because the data is more complete. In the data from loop 1, there are eleven sets of 144 total vehicle numbers every two seconds. It is divided into 10 sets for training and one set for testing. When the training epoch is around 500 during training, the results indicate that there are fewer losses. The author decided to choose 500 as this experiment's training epoch. Also, only two features, time and total vehicle numbers are considered.

3.2. Model Fitting

In the graph, the x-label is second, showing the change in traffic volume in 5 minutes. The actual data is well-matched by the LSTM forecasting result, as demonstrated in Figure 2, 3 and 4. Even if the data is pretty unstable, the LSTM can predict the trend of traffic volume. Similarly, in the prediction result of Loop 4 and Loop 9, the LSTM can follow the trend of changing, except for some specific statistics like the data around the second 100, the data fluctuates greatly, and the errors are bigger than the others.

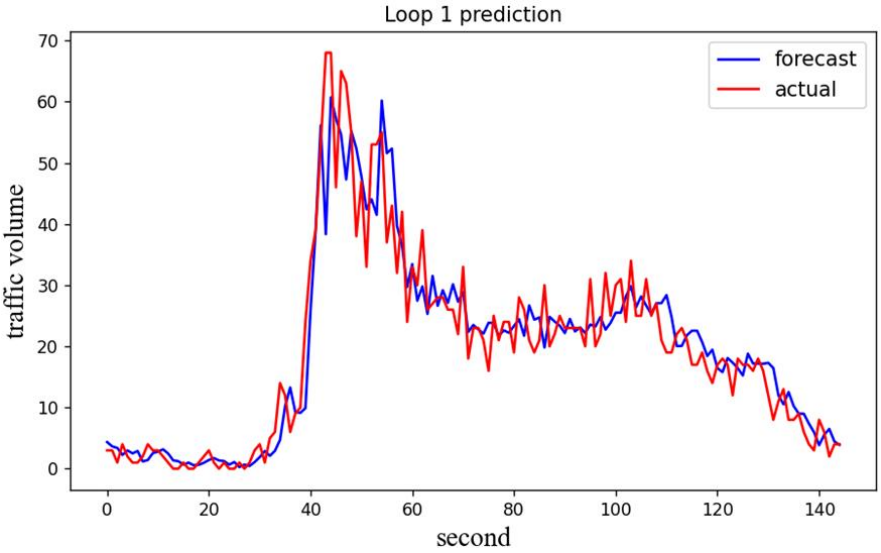

Fig. 2 Loop 1 traffic flow prediction result

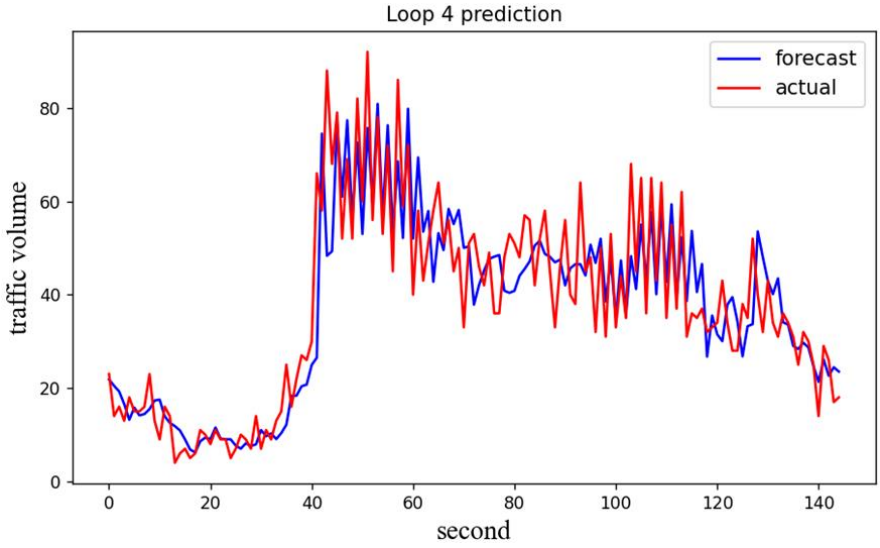

Fig. 3 Loop 4 traffic flow prediction result

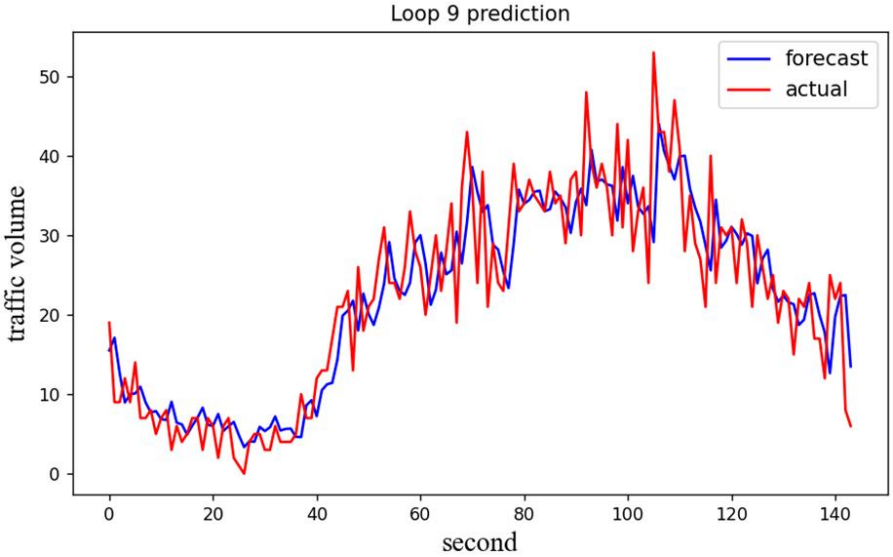

Fig. 4 Loop 9 traffic flow prediction result

To better find the accuracy of the LSTM model, the evaluation index should be analyzed. In this experiment, three loops' data are chosen. Table 2 shows the evaluation index of these three loops.

Table 2. Loop ID and Evaluation index

Result	MAE	MSE	R^2
Loop 1	3.579	28.884	0.881
Loop 4	6.515	78.977	0.810
Loop 9	4.253	33.495	0.832

The mean absolute error of Loop 1 is 3.579 as can be observed, and that of Loop 4 is about 3 higher than Loop 1. That is because the data of Loop 4 fluctuates more greatly than the data of Loop 1. Especially in the middle of the data set, the difference in traffic volume between the previous second and the following second is much greater. The mean absolute error is the highest among the three. Loop 4 is a different data set. In the data of Loop 1, there is a slow downward trend at the end. In the data of Loop 9, there is a comparatively slower upward trend at the beginning.

The Mean Squared Error can also confirm that because MSE is more sensitive to the errors. In Loop 4, the difference is greater, so the MSE is significantly higher than Loop 1 and Loop 9's MSE.

Although, the traffic volume fluctuates greatly, Loop 1 and Loop 9 R^2 reach 0.881 and 0.832. The Loop 4 also reaches 0.81. The traffic condition is more and more complex. These sets of data captured the traffic volume of every second and also collected by loops, making the data complex. That means that the LSTM performs well in predicting traffic flow.

Another finding is that when utilizing the LSTM model to estimate the traffic flow in this experiment, it can accurately find the upward trend. That means the LSTM model can predict the peak time of a traffic flow or a sudden increase in the traffic flow, which is very important in dealing with traffic congestion.

4. Conclusion

The LSTM model was used in this research to predict the traffic flow of urban intersections. By utilizing the data from an intersection of Shanghai and preliminary processing, the author chose the data from Loop 1 (1584 samples), Loop 9 (1296 samples), and Loop 4 (1872 samples). The mean absolute error of Loop 1 is 3.579, 4.253 for Loop 9, and 6.515 for Loop 4. The R^2 reaches over 0.8 for all the data sets. These results are good enough when dealing with such unstable statistics. Because the traffic flow is accurate to the second, the data fluctuates greatly. Comparing the three different types of data sets, the study finds that the LSTM model performs well when there is a relatively slow trend of change. The study also finds that the LSTM model can find the upward trend of traffic flow. It gives a good sample of solving a complex traffic condition. In the future, more and more cars and new types of vehicles will make the condition more serious. To improve the model, more features should be added, for example, the weather, the speed, and the vehicle type. To help the model make a more accurate prediction. In all, the LSTM model is suitable for processing time-series data with high accuracy and can be improved to fit more and more complex traffic conditions in the future.

References

[1] Pan B, Demiryurek U, Shahabi C. Utilizing Real-World Transportation Data for Accurate Traffic Prediction, 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012, 595-604.

[2] Williams B M. Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling. Journal of the Transportation Research Board, 2001, 194-200.

[3] Zivot E, Wang J. Vector autoregressive models for multivariate time series. Modeling Financial Time Series with S-PLUS, 2007, 385-429.

[4] Hong H, et al. Short-term traffic flow forecasting: Multi-metric KNN with related station discovery. 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China, 2015, 1670-1675.

[5] Duan M. Short-Time Prediction of Traffic Flow Based on PSO Optimized SVM. 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, China, 2018, 41-45.

[6] MA X L, et al. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. Sensors, 2017, 17(4): 818.

[7] Azad A K, Islam M S. Traffic Flow Prediction Model Using Google Map and LSTM Deep Learning. 2021 IEEE International Conference on Telecommunications and Photonics (ICTP), Dhaka, Bangladesh, 2021, 1-5.

[8] Cheng Z, Li Y, Zhu H. Improved Particle Swarm Optimization-based GRU Networks for Short-time Traffic Flow Prediction. 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, 2863-2868.

[9] Zhang X, Huang K, Liu C, Xu X. Urban Short-Term Traffic Flow Prediction Algorithm Based on CNN-LSTM Model. 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2023, 214-217.

[10] Feng H, Zhang X, Xu Y. Multi-step Ahead Prediction of Traffic Speed Based on Attention-based CNN-LSTM-BiLSTM. 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2022, 1-6.

[11] Lan T H, et al. Short-term traffic flow prediction model based on the optimization of SVM by CS algorithm. Journal of Qingdao Technological University, 2024, 134-140.