Deep Learning with Bidirectional Long Short-Term Memory for traffic flow Prediction

Song Xue1,a, Chunfu Shao2,b*, Shengyou Wang3,c, Yan Zhuang4,d

1College of Transportation, Beijing Jiaotong University, Beijing, China
2College of Transportation, Beijing Jiaotong University, Beijing, China
3College of Transportation, Beijing Jiaotong University, Beijing, China
4College of Transportation, Beijing Jiaotong University, Beijing, China

1090969126@qq.com, bcfshao@bjtu.edu.cn, c18114043@bjtu.edu.cn, d19114069@bjtu.edu.cn

Abstract—With the development of cities, the total number of trucks has increased year by year. Traffic flow forecasting has become an indispensable part of the cargo transportation industry and directly affects the development of the transportation industry. In the field of traffic flow prediction, Long Short-Term Memory (LSTM) model has advantages in processing time series, but it cannot extract the periodicity in time series. Therefore, in this experiment, a Bidirectional Long Short-Term Memory (BLSTM) model was constructed to predict traffic flow in the road network. It is worth mentioning that this article considers the non-parametric model autoregressive integrated moving average model (ARIMA) and the parametric model recurrent neural network (RNN) to compare and analyze with LSTM. Data from Guangwu Toll Station, Zhengzhou city, China were used to calibrate and evaluate the models. The experimental results show that the performance of RNN based on deep learning such as BLSTM and LSTM model is better than that of ARIMA. In order to better illustrate the advantages of BLSTM model, we comprehensively considered the performance effects of four models under morning peak, evening peak and flat peak. Experiments have proved that BLSTM has good nonlinear fitting ability and anti-noise ability, and the average prediction accuracy reaches 92.873%.

1. INTRODUCTION

In the past decade, with the acceleration of urbanization and the rapid development of car ownership, traffic flow prediction plays a key role in intelligent transportation system. In particular, traffic flow information has a strong demand for travelers and commercial companies. At present, many cities have established big data platforms to analyze data, thereby improving road traffic conditions, reasonably assigning driving paths to vehicles, and effectively alleviating traffic congestion.

However, real-time and accurate traffic flow prediction is not an easy task. Previous studies have used linear models to solve traffic flow prediction problems, including ARIMA [1], Kalman filtering [2] and exponential smoothing [3]. The linear model assumes that the future predicted data has the same characteristics as the past data. The model structure is simple and easy to calculate. When processing data with more feature points, the prediction effect will be close to perfect. However, traffic flow will be affected by factors such as the environment and emergencies. Moreover, the flow rate changes with...
time will also show obvious periodicity, so nonlinear models are introduced to solve this kind of problems.

RNN is a widely used nonlinear model, which has the ability to approximate nonlinear continuous functions with arbitrary precision. Many scholars use this method to apply to image recognition, trajectory prediction, feature selection, and have achieved good prediction results [4]. It is worth noting that as the number of network layers increases, the network weights of RNN will be gradually adjusted along the direction of local improvement, which will easily cause the algorithm to fall into local extremes, and the weights will converge to the local minimum.

As one of the most emerging and promising methods, LSTM can solve the above-mentioned problems. Compared with those traditional shallow learning structures, the LSTM framework can be represented by the nonlinear phenomenon of distributed hierarchical features [5].

In order to obtain higher prediction accuracy and deal with the periodicity of time series, this paper establishes BLSTM model. The prediction time interval is 1 hour, and MAE, MSE, MAPE are used for model performance evaluation indicators.

The remainder of this article is as follows: Section 2 illustrates the structure of RNN, LSTM, and BLSTM models, Section 3 explains the proposed model by studying Guangwu toll station dataset. Section 4 illustrates the conclusion

2. METHODOLOGY

2.1 Recurrent Neural Network Architecture

RNN overcomes the situation that general neural networks cannot handle time series changes. In most cases, the order of sample data has a great impact on natural language processing and traffic prediction. In this article, we apply RNN to the time information of traffic data and expand it according to the time node. The unfolded structure is shown in Figure 1.

![Figure 1 Unfolded RNN](image)

Therefore, the traffic flow prediction problem can be expressed by the following formula:

\[
\begin{align*}
    h^{(t)} &= \sigma(V \times q^{(t)} + U \times h^{(t-1)} + h) \\
    o^{(t)} &= f(W \times h^{(t)} + c)
\end{align*}
\]

Where \( q^{(t-1)} \) is used as the input of the sample data, which represents the flow at time \( t-1 \). \( h^{(t-1)} \) is the hidden state at time \( t-1 \), which is determined by \( h^{(t-2)} \) and \( q^{(t-1)} \), \( o^{(t)} \) is the output of neural network at time \( t-1 \), which is determined by the hidden state at the same time. \( U, V, W \) are the linear relationship parameters of input state, hidden state and output state respectively, which are shared in the whole neural network. \( b, c \) are the bias terms. \( \sigma, f \) are activation functions, representing \( \tanh \) and \( softmax \) respectively.

Therefore, the output of the neuron can be ‘saved’ and used as an input layer for other time periods. In other words, the output in the nth layer at time \( t \) is not only related to the input at time \( t \) of this layer, but also related to the output of other N-1 layers, which is the result of the joint action of historical time
relationship. It can be seen that the RNN used to predict the traffic flow also meets the temporal correlation. However, with the number of layers and the prediction interval increase, RNN model can easily make the calculated partial derivative close to 0, which will lead to the disappearance of gradient.

2.2 Long Short-Term Memory Architecture

Ordinary recurrent neural network is difficult to train, which makes it difficult to deal with long distance dependencies in practical applications. Therefore, the LSTM model was developed to solve various deficiencies in the RNN model, mainly by adding three kinds of 'gates' to control information, namely input gates, forget gates and output gates. The input gate calculates the probability of adding some new information through the sigmoid function, and the forget gate can calculate the probability of discarding some historical information through the sigmoid function.

\[
\begin{align*}
    r_t &= \sigma(h_{t-1} \times W_{hr} + X_t \times W_{xc} + b_r) \\
    i_t &= \sigma(h_{t-1} \times W_{hi} + X_t \times W_{xc} + C_{t-1} \times W_{ci} + b_i) \\
    f_t &= \sigma(h_{t-1} \times W_{hf} + X_t \times W_{xf} + C_{t-1} \times W_{cf} + b_f) \\
    o_t &= \sigma(h_{t-1} \times W_{ho} + X_t \times W_{xo} + C_{t-1} \times W_{co} + b_o) \\
    C_t &= f_t \times C_{t-1} + r_t \times i_t \\
    h_t &= o_t \times \sigma(C_t)
\end{align*}
\]

In the prediction process, MSE is used as a loss function to judge the error value in each cycle. The corresponding formula is as follows:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2
\]

Where \(p_i\) represents the predicted value, and \(x_i\) represents the true value.

Historically, the common choice for model optimization algorithms is the gradient descent method. In this paper, we use the Adam optimization algorithm. The Adam optimization algorithm is an extension of the stochastic gradient descent algorithm. Compared with other optimization algorithms, it is faster and requires less memory, and Adam can also calculate different adaptive learning rate for different parameters.

LSTM model has a strong ‘memory’ ability, but it will inevitably produce over-fitting. In order to reduce the probability of over-fitting, we use the dropout method to modify the number of neurons in
the hidden layer. In each cycle, dropout will randomly close some hidden layer neurons, and use the remaining neurons to update the parameters. This can reduce the relationship between features, thereby increasing the robustness and generalization ability of the model.

2.3 Bidirectional Long Short-Term Memory Architecture
At present, all neural network models make predictions based on historical data, that is, use the data at time t-2 and t-1 to predict the data at time t. But in reality, there are many random factors in the data. Therefore, the prediction accuracy of the one-way LSTM model will be affected due to the inability to extract random factors and periodic changes in the data. BLSTM is a kind of double-layer LSTM, but the direction of the two layers is completely opposite. It can pass through the input sequence from two directions at the same time. The upper layer calculates the forward hidden layer output from time 1 to time t, obtains and saves the output of the forward hidden layer at each time. The lower layer is calculated in reverse from time t to time 1, and the output of the hidden layer behind each time is obtained and saved. Finally, the average of the two results at the corresponding time is the final output.

3. EMPIRICAL RESEARCH
The data set used in this article comes from the Guangwu toll station in Zhengzhou, China. We analyzed the traffic data for one month from June 1, 2019 to July 1, 2019, and sampled the data every hour. In order to eliminate the adverse effects caused by singular sample data, it is necessary to standardize the data so that it is within the range of [0, 1]. The normalization formula is shown in (10). In addition, we divide the data set into two parts, where 75% is used as training data and 25% is used as validation data. In order to better explain the performance of the BLSTM model, this paper uses BLSTM, LSTM, RNN and ARIMA four models to predict the morning peak, flat peak and evening peak. The morning peak is from 7 am to 9 am, the evening peak is from 5 pm to 7 pm, and the rest of the time is flat peak. At the same time, MAE, RMSE, MAPE and R Squared ($R^2$) are used as evaluation indicators to judge the predictive performance of the model. The corresponding formula is shown in (11)-(14).

\[
X^c = \frac{X - \min(X)}{\max(X) - \min(X)}
\]

(10)

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |p_i - x_i|
\]

(11)

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - x_i)^2}
\]

(12)

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{|p_i - x_i|}{x_i} \right) \times 100
\]

(13)

\[
RSquared = 1 - \frac{\sum_{i=1}^{n} (p_i - x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}
\]

(14)

Where $P_i$ is the predicted value, $x_i$ represents the actual value, $\bar{x}_i$ denotes the average of the actual values.

The number of hidden layers and time lag are important factors in neural network, which directly affect the generalization ability, running speed and performance index of the model. Too few neurons in the hidden layer lead to under fitting. Conversely, using too many neurons can also cause problems. First of all, excessive number of neurons in the hidden layer may cause overfitting. When the neural network has too many nodes, the limited information contained in the training set is not enough to train all the neurons in the hidden layer, thus resulting in overfitting. In this paper, the optimal number of hidden layers and time lag are determined by experiments. The corresponding simulation results are listed in Table 1, 2 and 3 respectively.
Table1 Parameter selection of BLSTM model

| UNITS | lags | MAE  | RMSE | MAPE | R²  |
|-------|------|------|------|------|-----|
| 5     | 3    | 7.785| 8.941| 6.705| 0.956|
| 5     | 4    | 11.521| 14.073| 10.416| 0.891|
| 5     | 5    | 11.962| 14.548| 9.958| 0.883|
| 5     | 6    | 14.826| 17.098| 12.730| 0.839|
| 10    | 3    | 7.345| 8.255| 5.978| 0.962|
| 10    | 4    | 12.129| 14.242| 10.365| 0.888|
| 10    | 5    | 14.346| 16.870| 12.355| 0.843|
| 10    | 6    | 16.274| 19.815| 13.717| 0.783|
| 15    | 3    | 9.969| 11.729| 8.498| 0.924|
| 15    | 4    | 12.813| 15.220| 10.772| 0.872|
| 15    | 5    | 12.924| 15.947| 10.710| 0.860|

Table2 Parameter selection of LSTM model

| UNITS | lags | MAE  | RMSE | MAPE | R²  |
|-------|------|------|------|------|-----|
| 5     | 3    | 8.715| 11.544| 7.344| 0.927|
| 5     | 4    | 11.684| 14.324| 9.964| 0.887|
| 5     | 5    | 12.227| 15.494| 11.094| 0.868|
| 5     | 6    | 17.507| 20.169| 14.456| 0.776|
| 10    | 3    | 7.986| 9.509| 6.828| 0.950|
| 10    | 4    | 12.520| 15.266| 10.830| 0.872|
| 10    | 5    | 14.414| 17.360| 12.412| 0.834|
| 10    | 6    | 16.369| 19.263| 13.579| 0.796|
| 15    | 3    | 9.165| 11.040| 7.848| 0.933|
| 15    | 4    | 13.159| 16.142| 11.204| 0.857|
| 15    | 5    | 14.103| 17.790| 12.295| 0.826|

Table3 Parameter selection of RNN model

| UNITS | lags | MAE  | RMSE | MAPE | R²  |
|-------|------|------|------|------|-----|
| 5     | 3    | 12.237| 14.766| 9.227| 0.880|
| 5     | 4    | 9.531| 11.409| 7.560| 0.928|
| 5     | 5    | 10.535| 12.649| 8.132| 0.912|
| 5     | 6    | 11.311| 14.485| 9.450| 0.884|
| 10    | 3    | 15.478| 17.412| 12.513| 0.834|
| 10    | 4    | 13.396| 16.332| 10.638| 0.853|
| 10    | 5    | 12.221| 14.839| 10.169| 0.879|
| 10    | 6    | 10.288| 12.349| 8.330| 0.916|
| 15    | 3    | 9.276| 11.494| 7.566| 0.927|
| 15    | 4    | 13.879| 16.768| 10.956| 0.845|
| 15    | 5    | 14.861| 17.668| 11.533| 0.828|

It can be concluded from the above table that BLSTM and LSTM are similar in basic structure. Therefore, the variation trend of the model is basically similar when the parameters are chosen. If the number of hidden layers is less, the model complexity is low, and it is difficult to remember the historical data, unable to summarize the laws in the historical data, and the prediction effect is poor. If the number of hidden layers is too large, the model is too complex, resulting in good prediction effect on the training set, but poor prediction effect on the verification set. Therefore, the number of hidden layers in both BSLTM and LSTM models is 10. For time lag, it can be clearly seen from Table 1 and 2 that the prediction effect is not necessarily the best with more historical data of prediction, so the
The optimal time lag is 3. The parameter selection results of the RNN model are shown in Table 3 which proved that when the number of hidden layers and the time lag are 5 and 4 respectively, the model has the highest accuracy.

In the experiment, three neural networks and the ARIMA model were trained for performance evaluation under different peak conditions. Unlike previous work that only calculated losses, we also compared the accuracy of the models. The corresponding index comparison results are shown in Table 4, 5, 6, and 7.

Table 4: The prediction performance of BLSTM model

| Time       | MAE   | RMSE  | MAPE   | R²    |
|------------|-------|-------|--------|-------|
| 24-h       | 7.345 | 8.255 | 5.978  | 0.962 |
| Morning peak | 10.962 | 10.805 | 10.254 | 0.889 |
| Evening peak | 7.382  | 8.044 | 5.981  | 0.921 |
| Flat peak  | 7.420  | 8.822 | 6.296  | 0.941 |
| Mean       | 8.277  | 8.982 | 7.127  | 0.928 |

Table 5: The prediction performance of LSTM model

| Time       | MAE   | RMSE  | MAPE   | R²    |
|------------|-------|-------|--------|-------|
| 24-h       | 7.986 | 9.509 | 6.828  | 0.950 |
| Morning peak | 10.329 | 10.417 | 10.071 | 0.886 |
| Evening peak | 7.404  | 8.115 | 6.178  | 0.918 |
| Flat peak  | 8.801  | 10.830 | 7.867  | 0.924 |
| Mean       | 8.63  | 9.718 | 7.736  | 0.920 |

Table 6: The prediction performance of RNN model

| Time       | MAE   | RMSE  | MAPE   | R²    |
|------------|-------|-------|--------|-------|
| 24-h       | 9.531 | 11.409 | 7.560  | 0.928 |
| Morning peak | 14.274 | 15.787 | 13.938 | 0.857 |
| Evening peak | 7.745  | 8.672 | 6.601  | 0.911 |
| Flat peak  | 9.794  | 9.533 | 7.971  | 0.920 |
| Mean       | 10.336 | 11.350 | 9.018  | 0.904 |

Table 7: The prediction performance of ARIMA model

| Time       | MAE   | RMSE  | MAPE   | R²    |
|------------|-------|-------|--------|-------|
| 24-h       | 18.349 | 22.994 | 13.890 | 0.709 |
| Morning peak | 25.872 | 27.618 | 24.138 | 0.468 |
| Evening peak | 29.522 | 35.484 | 18.386 | 0.587 |
| Flat peak  | 18.349 | 22.994 | 13.890 | 0.614 |
| Mean       | 23.023 | 27.273 | 17.576 | 0.595 |

Tables 4, 5, 6, 7 further examine the predicted performance in a more intuitive way. We can see the detailed prediction results: in the whole time period, BLSTM is superior to other methods both in terms of loss value and accuracy, indicating that BLSTM model can significantly improve the prediction effect. It is worth noting that there is no significant difference between the BLSTM model and the
LSTM model during peak hours. Through research, it is found that road traffic will be close to saturation during peak hours, the periodicity is not obvious, and traffic flow tends to be stable. Conversely, when the peak period gradually dissipates and the traffic flow begins to fluctuate, BLSTM will be better than LSTM in prediction effect. In general, BLSTM can significantly improve the accuracy of traffic prediction. The prediction performance is also better than the general neural network model and ARIMA model, the error index MAPE is reduced by at least 0.6%.

4. CONCLUSIONS
In this contribution, we developed the BLSTM model, which mainly relies on historical and future data to predict traffic flow. Compared with other mainstream models, BLSTM can well extract the periodicity in the traffic flow, and predict the medium and long-term traffic flow with obvious periodicity. In order to better illustrate the advantages of BLSTM, we use the real data of Zhengzhou Guangwu toll station for testing and analysis, and build 4 models to predict the data in different time periods. The comparison proves that the average accuracy of BLSTM is 0.6% higher than that of LSTM, and 1.891% higher than RNN. In the future work, the influence of characteristics such as the flow of upstream and downstream sections and the average speed of the sections on the prediction effect will be mainly considered.

REFERENCES
[1] Guoxian, T. . (2005). Prediction model of short-term traffic flow at crossroads in cities. Computer & Communications.
[2] Hai-Feng, G. , Liang-Jun, F. , & Li, Y. U. . (2013). A short-term traffic flow prediction model based on fuzzy kalman filtering. Journal of Zhejiang University of Technology.
[3] Qi, C. , & Hou, Z. . (2012). Application of adaptive single-exponent smoothing for short-term traffic flow prediction. Kongzhi Lilun Yu Yinyong/Control Theory and Applications, 29(4).
[4] Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. "Generating text with recurrent neural networks." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.
[5] Shao, H. , & Soong, B. H. . (2016). Traffic flow prediction with Long Short-Term Memory Networks (LSTMs). TENCON 2016 - 2016 IEEE Region 10 Conference. IEEE.