Soft Computing (2022) 26:5363–5375
https://doi.org/10.1007/s00500-022-07096-7

APPLICATION OF SOFT COMPUTING

Deep learning method for traffic accident prediction security

Zhun Tian[1,2,3] · Shengrui Zhang[1,3]

Accepted: 2 September 2021 / Published online: 21 April 2022
© Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Since frequent traffic accidents bring great losses to people's safety and social property, this paper takes the results of traffic accident risk prediction as the basis so that it can help city managers to reasonably deploy police force, relieve traffic pressure, avoid traffic accidents and provide safe guidance to pedestrians. Based on this paper, a deep learning framework including spatiotemporal attention mechanism is proposed to solve the problem of traffic accident prediction in urban areas, and the experimental simulation shows that the accuracy of traffic accident risk prediction proposed in this paper reaches 94%.

Keywords Deep learning · Traffic flow · Risk prediction · Algorithm

1 Introduction

Deep learning has been called "technology changing the world" (Zihua, et al. 2020), and in addition to breaking records in image recognition and speech recognition competitions, deep learning has produced a range of exciting results in a variety of tasks in natural language processing, notably topic classification, sentiment analysis, question answering systems, and machine translation. At the same time, deep learning has beaten other machine learning techniques in many natural science fields and has become a major force in advancing science. Representative tasks include predicting the activity of potential drug molecules, analyzing particle gas pedal data, reconstructing brain circuits, and predicting the effect of mutations in non-coding DNA on gene expression (Hao 2020). Deep learning has proven to be very adept at discovering complex structures in high-dimensional data, as it allows computational models to learn data representations with multiple levels of abstraction through multilevel processing.

The extensive success of deep learning comes from the rapid development of deep neural network. In 1957, Rosenblatt proposed the concept of perceptron to solve the simple linear classification problem, which is the beginning of neural network. In 1986, Hinton (Yongqiang and Xiaofan 2020) proposed a backpropagation algorithm to solve the training problem of multi-layer perceptron. In 2006, Hinton (Xiangyu et al. 2020) proposed the concept of "deep belief network" for the first time, using the pre-training method to let the neural network find an initial better solution, and then run the fine-tuning technology to train the whole network. This training method proposed by Hinton provides an empirical method for the training of deep neural networks and speeds up the research of various deep learning networks. At present, the widely used deep learning models are recurrent neural network and convolutional neural network (Liu, et al. 2020).

Yuan Hongtao (2020) proposed signal timing optimization based on short-term traffic flow prediction. The method is good, but how to optimize is problem. Yuan Er Bao (2020) proposed traffic flow prediction modeling and calculation based on neural network algorithm. For the method, I think the method is based the traffic flow prediction modeling, but the model is not accurate. Zheng

✉ Shengrui Zhang
zhangsrchd@126.com

Zhun Tian
tianzhunxauat@126.com

[1] College of Transportation Engineering, Chang'an University, Xi'an 710064, Shaanxi, China

[2] School of Civil Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, Shaanxi, China

[3] Key Laboratory of Transport Industry of Management, Control and Cycle Repair Technology for Traffic Network Facilities in Ecological Security Barrier Area, Chang'an University, Xi'an 710064, Shaanxi, China

Youkang (2020) proposed short-term road traffic flow prediction based on deep learning, and the fault of the method is the short-term road traffic flow. Huang Zijing (2020) study on multi-period exit flow prediction method of Expressway multi-toll station based on spatiotemporal attention mechanism.

Yan Yang (2020) short-term traffic flow prediction based on deep learning. Yan Yuchan (2019) proposed short-term traffic flow prediction using MPSO optimized SVR. The author used the different method to optimize SVR, and the data are little.

Jiang Fucai, et al. (2018a, b) proposed gray model in ship traffic flow prediction of Dongying Port, the idea is good, but the model is incorrect. Zhang Guo (2019) used the Markov random to realize the short-term traffic flow prediction, as the data are a little small, so the prediction is incorrect.

Wu Feng (2019) researched on intelligent traffic flow prediction technology based on spark.

Chen Shiju (2019) researched on urban vehicle traffic flow analysis algorithm based on deep learning.

Xu Ruiguang, and Liang Shidong (2018) proposed traffic flow prediction method based on bilinear recurrent neural network.

Liu Mingyu, et al. (2018) used the deep learning for traffic flow prediction. The flow is correct, but the prediction is incorrect.

Wen Feng, Zhang Guo (2018) proposed the SVR algorithm for short-term traffic flow prediction. The idea is good.

Shen Guicheng, Guan Shuicheng, and Sun Fangyu (2018) also proposed short-term traffic flow prediction of Expressway considering optimal time delay factor spatiotemporal model. The effect is positive. Timothy Qiu (2020) researched on predictive control strategies for distributed economic models of nonlinear vehicle queueing systems. According to the author thinking, the simulation effect is great. Yu Yang, Liang Jun, Chen Long, Chen Xiaobo, Zhu Ning, and Hua Guodong (2020) proposed an emergency lane change behavior prediction method based on Gaussian hybrid hidden Markov model and artificial neural network. The proof is great, but the simulation may be incorrect. Wang and Pan (2020) proposed on the coupled traffic flow Aw-Rascle model Riemann problem. Aung (2020) researched on traffic optimization and driving safety enhancement based on vehicle networking. Zhang (2020) researched and analysis of highway accident model with high bridge-tunnel ratio based on multiple nonlinear regression method. Yang Pengfei, Sun Xianbo (2020) researched on road icing prediction model based on support vector machine algorithm Tang Zuogam, Sun Yeyao, Guo Ying, Yu Fengquan (2020) proposed modeling the number of fatalities and injuries in street traffic accidents based on multi-source data fusion.

The paper has following arrangement.

Sect. 1 is introduction. Sect. 2 is model design. The calculation and experimental analysis are for Sect. 3. Sect. 4 is example analysis. Sect. 5 is conclusions.

2 Ta-Stan model design

2.1 Design of spatial attention mechanism

The reason why spatial attention mechanism is proposed is that for each specific time slice, the input of the model is all traffic indicators of the whole urban area, and the data volume is large. However, we also hope that the hidden unit in the encoder can learn the different importance of local traffic indicators and external traffic indicators better and faster. Therefore, this paper designs local spatial attention mechanism and global spatial attention mechanism to help the model establish the connection between the historical state and all traffic index inputs at the current time (Wenqin 2020).

2.1.1 Local spatial attention mechanism

The local spatial attention mechanism is a weighting mechanism for the input data of the model in the encoder stage, which pays more attention to the influence of local traffic indicators. For a region, a variety of local traffic flow data and traffic accident risk data have a complex relationship with the future traffic accident risk, and this relationship will change dynamically over time. For example, in the urban area of city a, most of the time, yellow taxi is the main traffic mode, because it can stop at any place (Jie 2020). However, in the morning and evening rush hours, due to the limited number of yellow taxis, their carrying capacity is not enough, so the use of green taxis and online car Hailing will increase greatly, and the importance of these traffic indicators affecting the risk of traffic accidents will change. In order to describe the impact of this local change, given the j-th traffic index feature sequence of the *i*-th region at time *t*, an attention mechanism is designed to dynamically obtain the dynamic relationship between local future traffic accidents and different local traffic indexes

$$a_t^j = w_1 \tanh\left(w_2 h_{t-1} + w_3 c_j^{z_{i,t}} + b_0\right), \quad j \in [1, n_I] \tag{1}$$

$$\alpha_t^j = \frac{\exp\left(\alpha_t^j\right)}{\sum_q^{n_z} \exp(a_t^q)} \tag{2}$$

Formula 1 uses the concat method of formula 2 to calculate the similarity. The h_{t-1} represents the hidden layer state of the historical time in the encoding. a_t^i can depict the importance of the hidden layer unit and the different traffic indexes at the previous time to the hidden layer of t.

2.1.2 Global spatial attention mechanism

For a regional traffic accident, the impact of its external area on the regional traffic accident risk is also important. We observe that the influence of the external region on the current region is dynamic and will change with time. At the same time, the impact of different traffic indicators in the external area on the traffic accident risk of the target area is also dynamic. If we simply take the traffic indicators of all areas that may affect the target value as the input, there will be two problems: (1) the calculation cost is very high; (2) the convergence is very slow and the effect is not good. Therefore, using the idea of multi-perspective for reference, this paper divides the overall spatial impact into two angles: (1) index: all regions of a traffic index have a common index impact; (2) regionality: all traffic indexes of a region have an overall regional impact. For example, in the morning rush hour, the traffic flow in all areas will increase, so the importance of a certain traffic index in all areas will increase, regardless of the region, which is the indicator effect; on the other hand, for the airport, fixed flights will output high flow risk to A office area in the daytime, and high flow risk to B residential area in the evening, so this is the reason regional comprehensive influence is relatively large, which is a regional influence. Inspired by the above facts, this paper designs a new attention mechanism, which can dynamically obtain the impact of other regions on the traffic accident risk of the target region (Hongtao 2020; Bao 2020; Youkang and Honglei 2020; Zijing 2020; Yang 2020; Yuchan, et al. 2019).

$$e_t^k = w_4 \tanh\left(w_5 h_{t-1} + w_6 c_k^t + b_1\right), \quad k \in [1, n_I] \tag{3}$$

$$\beta_t^k = \frac{\exp\left(e_t^k\right)}{\sum_q^{n_I} \exp\left(e_t^q\right)} \tag{4}$$

$$r_t^l = w_7 \tanh\left(w_8 h_{t-1} + w_9 c^{Z_l,t} + b_2\right), \quad l \in [1, n_2] \tag{5}$$

$$\gamma_t^l = \frac{\exp\left(\gamma_t^l\right)}{\sum_q^{n_z} \exp\left(r_t^q\right)} \tag{6}$$

Here, e_t^k and β_t^k represent the impact factors of the global k traffic index and the normalized impact factors, as shown in formulas 3 and 4. r_t^l and γ_t^l are the influence factors and normalized influence factors of the global l-th region, as shown in formulas 5 and 6. Therefore, the global attention mechanism is also divided into global indicator attention and global regional attention.

2.1.3 Spatial attention fusion module

After using the spatial attention mechanism, this paper extracts three kinds of spatial input features, and the next step is to fuse these three features as the new input of the model.

Figure 1 shows the operation process of fusing the three spatial input features. Among them, $c_i^{z_i,t}$, $x_{\text{local}}^t \in R^{n_1}$, $c_k^t \in R^{n_I \times n_z}$, after the operation of convl * 1, $c_k^t \in R^{n_I}$, so $x_{gl-I}^t \in R^{n_I}$; similarly, $c^{z_{l,t}} \in R^{n_z \times n_I}$, after the operation of convl * 1, $c^{\tilde{z}_{l,t}} \in R^{n_z}$, and finally, the input of the encoder at time t is $x_{\text{input}}^t = \left\{x_{\text{local}}^t, x_{l-1}^t, x_{gl-z}^t\right\} \in R^{2 \times n_I + n_z}$.

2.2 Design of time attention mechanism

Similar to neural machine translation, in the framework of traffic accident risk prediction, attention mechanism of time should be added between different time solutions in coder and decoder to fit the dynamic importance of historical time slice to future time slice. This is because when the encoder length is short, the output of the last encoder can be used as the input of the first decoder hidden layer, and the model effect is very good. However, when the length of time slice becomes longer, the backpropagation of time dimension becomes more difficult. Therefore, we need to associate each time slice t' in the decoder with each time slice t in the encoder to add a connection between them to help the model better learn the influence of historical moments on future moments. Through formulas 7, 8 and 9, we can calculate the attention value of each time t' in the decoder and each hidden layer in the encoder, so as to obtain the context information $c_{t'}$ of each time t' in the decoder.

$$e_{t'}^p = h'_{t'-1} W_a h_p + b_d \tag{7}$$

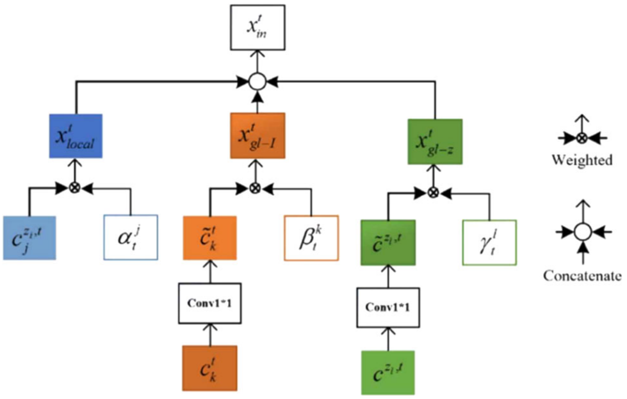

Fig. 1 Space Insurance module

$$\delta_{t'}^{p} = \frac{\exp\left(e_{t'}^{p}\right)}{\sum_{j=1}^{T} \exp\left(e_{t'}^{j}\right)} \tag{8}$$

$$c_{t'} = \sum_{p=1}^{T} \delta_{t'}^{p} h_p \tag{9}$$

2.3 External environment feature fusion module

Traffic accidents are not only closely related to traffic accidents and traffic volume, but also related to the inherent attributes of this area, such as the average speed limit and the number of traffic signs in this area. On the one hand, the point of interest is also very important for traffic accident risk prediction, because the point of interest shows the functionality of this area. After all, the accident risk degree of schools, parks and commercial streets is different. On the other hand, weather and time factors also affect the traffic pattern in this area. For example, in rainy days, sunny days and even storms, the travel demand and the driving mode of drivers will change. Therefore, the previous spatiotemporal traffic indicators should be combined with the weather, time, street design and other external environmental characteristics to better predict the future traffic accident risk model. Inspired by the influence of adding external factors to the traffic accident prediction work (Fucai et al. 2018a, b; Guo 2019), this paper designs a simple and effective component to join the decoder stage to assist the traffic accident risk prediction. Firstly, this paper extracts a series of features from the data of street speed limit, point of interest, weather, area number and time. The corresponding features are described in detail in Sect. 2.2. For all features, this module uses a layer of embedding layer to embed these features and then adds multi-layer full connection to realize the interaction of low-level features and generate higher-level features. In order to prevent overfitting, dropout technology is used between the full connection layers of each layer. Finally, the output of the full connection layer is the external environment feature input that we need. The external environment feature input is a part of the decoder at time t' to help predict the traffic accident risk at time t'.

2.4 Traffic accident risk prediction model based on deep spatiotemporal attention mechanism

Ta-stan model includes three modules, namely model input, encoder stage and decoder stage, and the decoder stage also has external environment feature input. The specific steps of ta-stan model are as follows:

Step 1 Model input: according to the time dimension and space dimension, the multi-source heterogeneous data related to traffic accidents are calculated.

(1) The target city is divided into sub-regions according to the traffic administrative region, and the traffic flow and traffic accident volume of different models in the sub-region at historical time are counted; the traffic flow and traffic accident volume data of different models are taken as the multiple traffic index $C = \left(c^{z_i,1}, c^{z_i,2}, \ldots, c^{z_i,n_I}\right)$, n_I Cof the i region z_i, which is called the number of traffic indexes.

(2) To set the time window size T, you can choose 1 h, 2 h or 1 day

(3) In a fixed time window T, the local historical traffic Index Series $\left\{C_{t_0-(p-1)'}^{z_i}, C_{t_0-(p-2)'}^{z_i}, \ldots, C_{t_0}^{z_i}\right\}$ of region z_i will be formed by superposing the data of historical p timestamps adjacent to region a according to time sequence, where $C_{t_0-t}^{z_i} = \left(C_{t_0-t}^{z_i,1}, C_{t_0-t}^{z_i}, \ldots, C_{t_0-t}^{z_i,n_I}\right)$ represents the distance between region z_i and current time t_0, the historical traffic index vector of t timestamps, $t \in [0, p-1]$;

(4) In the time window r of (3), the data p of historical a timestamps adjacent to all regions set Z are superimposed together according to the time sequence to form a global historical traffic index sequence $\left\{C_{t_0-(p-1)}^{z}, C_{t_0-(p-2)}^{z}, \ldots, C_{t_0}^{z}\right\}$, where $z = \left\{z_1, z_2, \ldots, z_{m'}\right\}$, m is the number of regions, which $C_{t_0-t}^{z} = \left(C_{t_0-t}^{z_1}, C_{t_0-t}^{z_2}, \ldots, C_{t_0-t}^{z_m}\right)$ represents the historical traffic index vector of the distance t_0 timestamps between all m regions and the current time t, $t \in [0, p-1]$.

Step 2 Encoder: learn the local temporal changes and the temporal and spatial effects of surrounding areas.

According to the first mock exam time index of traffic data in the first step of the model, the traffic risk of B C in the future A is predicted. The basic deep learning framework is encoder–decoder. The encoder is used to encode the traffic index of D historical moments, and the decoder is used to decode the traffic accident risk of e future moments. Among them, the basic unit of the encoder uses several gating loop units (Grus) which can model long-term time dependence.

Step 3 decoder: timing prediction using temporal attention mechanism and external features.

In the decoder stage, the predicted value of traffic accident risk in the next seven moments is generated, and the external environment features are integrated to improve the prediction accuracy.

3 Calculation and experimental results

3.1 Experimental setup

3.1.1 Sample generation and data set division

In recent years, due to the rise of online car hailing, the number of online car hailing accounts for an increasing proportion in some regions (such as B city), Kennedy Airport and LaGuardia airport. Therefore, the composition structure of traffic flow in G city is in a pattern of long-term change. The research object of this paper is that there is only dynamic adjustment of traffic flow proportion in G city from 2017 to 2018, but there is no big change, so it can ensure the timeliness of the article model (Fucai et al. 2018a, b; Guo 2019; Feng 2019). Next, we discuss the process of sample generation and data set partition.

(1) Data sample generation

The problem to be studied in this paper is to predict the future 12 o'clock traffic accident risk based on historical 12 o'clock traffic accident data, traffic flow data and other external data. First of all, we use the sliding window method, which is commonly used in time series problems, to process all the data into a series of 24 slices, with the first 12 hours as the input of training, and the last 12 hours as the label. Therefore, the data of 2017-201 are transformed into (731 * 24−12) * 236 = 4610196 sequence inputs and outputs.

(2) Data set partition

Data set is usually divided into three parts: training set, verification set and test set. Generally speaking, training set takes up the largest proportion. It is used to train the model in machine learning to determine the structure and parameters of the model. The validation set is used to evaluate whether the super parameters and structure in the model have strong generalization ability. Because the parameters in the model are determined by the training set, it will lead to the overfitting of the parameters for the training set data. The test set does not participate in the training of the model, nor in the determination of super parameters, but a separate data set to evaluate the performance of the model. Generally, the test set is divided separately, and then, the remaining data set is divided into training set and verification set by "leave one" method. Finally, the model is trained several times by combining with cross-validation. This can make every piece of data be trained, and the method of averaging is also an idea of model fusion. The idea of "stay method" is shown in Figure 3, which means that the data set is divided into parts averagely, one of which is randomly selected as the verification set and the other data as the training set during each training, so that the model will train times to get the sub-model, and the mean value of the prediction result of the sub-model is taken as the final prediction result.

The traffic accident risk problem in this paper is a kind of time series problem. Because the data are generated in sequence, the statistical characteristics of the data will change with the time. Because of this, we cannot train with the future data and evaluate with the historical data, which violates the law of time. At the same time, considering the large scale and long running time of the data set, this paper adopts the improved time series data set division method, as shown in Fig. 4–4, renumbers the months, and the month represents the third month from December 2016, for example, January 2017 is the first month, January 2018 is the 13th month, and so on. Therefore, the data set has a total of 18 months from January 2017 to June 2018. In order to increase the generalization ability of the model and speed up the training of the model, another sixfold cross-validation is carried out in 18 months. According to the time, the first 3 months are used as the training set, and the last 3 months are used as the validation set. Six models with the same structure are obtained. Then, the six models are used to predict the traffic accident risk from July to December in 2018, and the final prediction result is the average of the predicted values of the six models (Shiju 2019; Ruiguang and Shidong 2018; Mingyu, et al. 2018).

3.1.2 Cost function and evaluation function

The loss function is used to measure the difference between the single predicted value and the real value, or the error degree of the model. It is a non-negative value. If the errors of all the predicted samples are summed up, the cost function of the whole model can be obtained. The purpose of iterative training of deep learning model is to find an appropriate optimization method, such as gradient descent method, by calculating the overall cost function to update the parameters of the model and minimize the cost function. When the current valence function is the smallest, the parameters of the model can be considered to be optimal. For regression problems, the commonly used loss functions are mean absolute loss function, log likelihood loss function and mean square loss function. In this paper, MSE is used as the cost function of the model:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_L \overset{\wedge}{-} y_i\right)^2 \tag{10}$$

At the same time, different evaluation functions are used to evaluate the final prediction results. In this paper, rootmean-square error and mean absolute error are used to evaluate the prediction results of the model. RMSE and Mae have their own advantages in evaluating a model. RMSE is sensitive to large errors, while Mae can directly observe the real average error. Therefore, this paper uses these two evaluation indicators to evaluate, in order to get the most comprehensive and intuitive evaluation results. The formula of the two evaluation indexes is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i \overset{\wedge}{-} y_i\right)^2} \tag{11}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left| y_l \overset{\wedge}{-} y_i \right|^2 \tag{12}$$

3.2 Super parameter optimization experiment

3.2.1 Optimizer and learning rate

Optimizer refers to the strategy to minimize the cost function. Generally speaking, the optimizer in deep learning is based on the idea of gradient descent, that is to say, the change direction of all model parameters is along the direction of the gradient of the parameters relative to the cost function, where the gradient direction is the value of the first derivative, and the value of the parameter change is the step set in advance in the experiment, which is also called the learning rate.

The commonly used optimizers are random gradient descent, batch gradient descent and small batch gradient descent. The optimizers with adaptive learning rate are adagrad, rmsprop, adadelta and Adam. Among them, SGD can often get the best optimization effect, but it will update the parameters of each sample once, so the running time of the algorithm is too long. On the contrary, bgd algorithm converges fast, but the optimization result is not very good. Therefore, MBGD is a trade-off between SGD and bgd, which adopts fixed batch optimization and takes into account both time efficiency and algorithm effectiveness. Then, adagrad uses the optimization method of adaptive learning rate; the learning rate will decay with the number of parameter iterations; it can automatically adjust the learning rate, but there will be the problem that the learning rate is too small to train in the later stage of training. Rmsprop, adadelta and Adam use different methods of historical gradient attenuation factor to optimize adagrad's disadvantage of low learning rate. In most cases, Adam optimizer can achieve better optimization effect. In this experiment, Adam optimizer is also used for training.

For model training, we not only need to select the appropriate optimizer, but also need to give the optimizer a suitable learning rate. If the learning rate is too large, the optimization will vibrate near the extreme value, and the model will not converge; if the learning rate is too small, the convergence speed will be too slow, and it is easy to fall into the local optimal point, and the model effect is not good. Therefore, four learning rate values of 0.01, 0.001, 0.0001 and 0.00001 were set up in the experiment, and the optimal learning rate value of the model was determined through the experiment. The experimental results are shown in Fig. 2.

According to the experimental results, when the learning rate is too large, such as 0.01, it will cause the loss value of the model to vibrate and cannot converge. If the learning rate is too small, take 0.00001, the model can converge, but the effect is very poor, obviously in the local extreme point. In conclusion, 0.001 is the most suitable learning rate.

3.2.2 Batch size

In the previous section, we described the disadvantages of SGD and bgd. It is not a good strategy to use one sample or all samples when updating parameters. Therefore, a super parameter batch size will be set before training to determine the number of samples used by the optimizer when modifying model parameters. Due to the diversity of samples, the gradient results calculated by different number of samples are different. Therefore, the optimizer takes different optimization routes under different batch size, which leads to that for different data sets and different models, batch size will affect the final optimization results of the model. On the whole, the maximum value of batch size is limited by GPU video memory, and the minimum value of batch size is limited by the longest training time that the experimenters can bear. Therefore, this experiment uses four batch sizes 64, 128, 256 and 512 to determine the optimal batch size of the model. The experimental results are shown in Table 1.

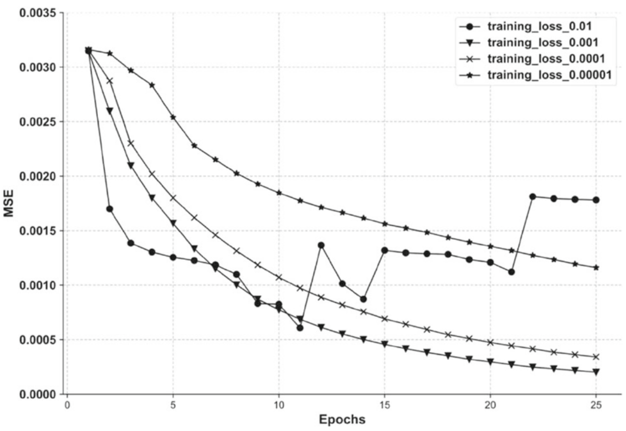

Fig. 2 Loss curve of different learning rates

Table 1 Experimental results of different batch sizes

Batch size	MSE	RMSE	MAE	时间/Epoch (min)
64	1.77E–4	0.01331	0.0084	65
128	1.71E–4	0.01309	0.0080	30
256	1.72E–4	0.01312	0.0082	13
512	1.80E–4	0.01342	0.0085	7

Through the analysis of the experimental results, we can find that the batch size is reduced from 512 to 128, the three evaluation functions are smaller and smaller, and the training results are better and better. However, when the minimum value of batch size is 64, the experimental results are not optimal, indicating that the smaller the batch size is, the better. From the time spent on one epoch, from 512 to 64, the size of batch size is reduced exponentially, and the time spent on training is increased exponentially. Although the training effect of batch size=128 is the best, the effect of batch size=256 is not much different from it, and the time spent is much less. Therefore, the experimental setting is batch size=256.

3.2.3 Circulating nerve unit

Because of gradient disappearance or gradient explosion in RNN, the cells far away from the current position in the sequence cannot get gradient update, and the sequence cannot get long-term dependency. Therefore, long-term and short-term memory network LSTM and gating neural network Gru are proposed as two important variants of RNN. Compared with the classical RNN, the common innovation of them is to set up a memory module, and update the state between neurons at adjacent times. In LSTM and GPU, forgetting gate and update gate control the memory module, respectively. However, LSTM and Gru are also different. The biggest difference is that the memory component of LSTM is partially visible, while the memory component of Gru is completely visible. The degree of visibility is a measure of the amount of this component that can be used by neurons at a later time. The visibility of LSTM memory module is controlled by output gate, while Gru has no output gate. Therefore, there are some differences between them in the expression of the model. In this section, experiments are set to explore the performance of LSTM and Gru in the experiment. The experimental results are shown in Table 2, where epochs represent the number of rounds used in model convergence, and the RNN layers of encoder and decoder are two (Feng and Guo 2018; Guicheng et al. 2018).

It can be found from the experiment that under the same number of neurons, the effect of LSTM model is almost the same as that of Gru model, but Gru converges in less training rounds, and the amount of parameters used is less than 20%, and the training time of the model will be relatively less. Therefore, the Gru model was selected as the basic circulatory nerve unit.

3.2.4 Dropout

Because the network layer of deep learning model is deep and the parameters are large, it is easy to produce overfitting phenomenon. There are three common methods to solve the overfitting phenomenon: (1) increase the data set to make the data distribution more consistent with the real distribution. The method of increasing data set is commonly used in the field of image recognition, and a series of sample enhancement methods appear, which are not used in the experiment; (2) regularization is added to the loss function, and the size of parameters is explicitly added to the optimization objective. In the experiment, 12 regularization is used to prevent the parameter value from being too large, setting=0.001; (3) dropout [80]. Dropout is the most obvious way to prevent overfitting effect from improving. This paper also uses dropout and adjusts its parameters.

Dropout refers to setting a probability in a certain layer of the network, then the probability of the neurons in this layer of the network will be screened out, and the screened neurons will not be associated with the neurons in the next layer, which is equivalent to making some neurons work and the other neurons inactivate. The addition of dropout makes the connection between layers of the network independent of fixed relationship neurons and makes the weight update independent of some special features. To some extent, it increases the generalization ability of each layer structure. From the perspective of model fusion, we can think that dropout is similar to the bagging process of model fusion. Because the experiment in this paper is a mini batch training method, and each update parameter selects part of the samples to calculate the gradient, dropout makes the neurons updated each time not fixed, which is equivalent to a number of small networks with different structures, and the whole dropout process is equivalent to a number of different models average by network, so there is the idea of model fusion in it. At the same time, dropout can reduce the training parameters of the model, improve the generalization ability, prevent overfitting, and shorten the training time of the model. Therefore, the experiment of dropout is designed in this section. The main purpose is to add dropout to the RNN of encoder and decoder. Here, the number of neurons is set to 256, and the circulating nerve unit uses Gru. The experimental results are shown in Table 3.

Through the analysis of the experimental results, it can be seen that if the dropout value is too large, the effective

Table 2 Experimental results of different circulating neurons

Circulating neuron	Number of neurons	RMSE	Epochs	Parameter quantity
LSTM	128	0.01527	16	221w
GRU	128	0.01533	14	177w
LSTM	256	0.01348	23	886w
GRU	256	0.01312	19	708w

Table 3 Results of different dropout experiments

Dropout	MSE	RMSE	MAE
0	1.76E–4	0.01446	0.0087
0.1	1.7E–4	0.01524	0.0094
0.2	1.72E–4	0.01312	0.0082
0.3	1.79E–4	0.01516	0.0087
0.4	1.83E–4	0.01704	0.0098
0.5	1.82E–4	0.01611	0.0089
0.6	1.86E–4	0.01663	0.0101

connection between neurons cannot be formed, and the experimental effect is not good, but if the dropout value is too small, it cannot prevent overfitting. To sum up, when the dropout value is 0.2, the model effect is the best.

3.3 Comparison experiment with traditional machine learning model

In order to compare the effect of the model proposed in this paper, the following models are used as the basic models:

(1) Ha historical average.
(2) Gru, classic gating neural network.
(3) Seq2seq. A Gru layer is used to encode input to generate context information, and then the context information and another Gru layer are used to decode the target sequence.

3.3.1 Historical average of HA

The overall trend or centralization trend of traffic accident risk reflected by historical average is a basic statistical characteristic quantity. Because the risk of traffic accident has an obvious time cycle, which is 24 h, this paper tests a total of three historical average values based on the hours, which are given as follows:

(1) HA-1.The average traffic accident risk of the same hour in the same training area is taken as the prediction value.
(2) HA-2.The average traffic accident risk of the same area and hour in the previous month is taken as the prediction value.
(3) HA-3.The average traffic accident risk of the same area and hour in the previous week is taken as the prediction value.
(4) HA-4. The average traffic accident risk of the same area and hour in the first three days is taken as the prediction value.

The test results are shown in Table 4. According to the experimental results, HA-4 is the best among the three indicators, and the experimental performance is improved from HA-1 to HA-4, which indicates that the traffic accident risk changes with time, but the level of traffic accident risk is basically stable in the near future.

3.3.2 GRU gated neural network

GRU can predict time series directly, but the length of input series and output series should be equal. In the experiment, the length of time series is 12, and the schematic diagram of GRU network is shown in Fig. 3. The bottom layer of GRU network is the input layer. The input layer dimension of each time slice is 5, which, respectively, represents the five traffic index data of a region. What RU network learns is the impact of local historical traffic index on future traffic risk. Similarly, the network structure and super parameters are adjusted to optimize the GRU model.

The experimental results are shown in Table 5. The experiment mainly adjusts the number of network layers and neurons, and the dropout value is set to 0.3. The experimental results show that the model with 2-layer Gru and 512 channel elements has the best effect. The increase

Table 4 Experimental results of historical average value of HA

Model	MSE	RMSE	MAE
HA-1	5.071E–3	0.07121	0.0483
HA-2	4.515E–3	0.06794	0.0455
HA-3	4.08E–3	0.06388	0.0431
HA-4	3.481E–3	0.05902	0.0409

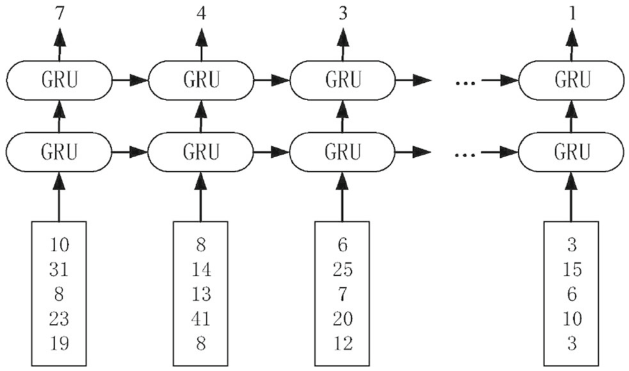

Fig. 3 Gru network diagram

Table 5 Gru network experiment results

Number of network layers	Number of neurons	Dropout	RMSE
1	256	–	0.0525
1	512	–	0.0502
2	256	0.3	0.0471
2	512	0.3	0.04423
3	256	0.3	0.0463
3	512	0.3	0.0454

in the number of network layers and neurons brings nonlinear fitting ability, but it is also easy to overfit.

3.3.3 Comparison of experimental results

This section compares the proposed model with the basic model. In order to be fair, the best results of each model are shown. The experimental results are shown in Table 6. According to the experimental results, LSTM is better than ha, which indicates that RNN structure with memory unit can capture the connection of different time slices before and after processing time series problems, showing strong advantages. And seq2seq is better than LSTM, which indicates that the decoder structure plays a role, and seq2seq has better performance in long-term prediction. The ta-stan model proposed in this paper has achieved the best results in MSE, RMSE and Mae, and has been greatly improved. This is because the ta-stan model can well capture the dynamic impact of local traffic indicators, other regional traffic indicators and external factors on the future traffic accident risk.

Table 6 Comparison between ta-stan and basic model

Model	MSE	RMSE	MAE
HA	3.481E–3	0.05902	0.0409
GRU	1.95E–3	0.04423	0.0259
Seq2seq	1.376E–3	0.03712	0.0220
TS-STAN	1.72E–4	0.01312	0.0082

3.4 TA-stan component effect evaluation experiment

This section explores the role of each component of ta-stan through experiments, including five components: local spatial attention, global indicator attention, global regional attention, temporal attention and external environment feature fusion module. Add one component at a time and then observe the different impacts of the model in the five regions of New York. There are five variant models.

(1) TA-1: L
(2) TA-2: L+GI
(3) TA-3: L+GI+GZ
(4) TA-4: L+GI+GZ+T
(5) TA- STAN: L+GI+GZ+T+E

The experimental results of the variant model are shown in Table 7. Next, the article evaluates the effect of each component by analyzing the experimental results of the variant model.

3.4.1 Evaluation of local spatial attention mechanism

As shown in Fig. 3, in the five regions of G City, the performance of Ta-1 model is better than that of seq2seq, especially in B, the loss value decreases most. This is because there is less interaction between B and other large M, and the regional traffic condition is more affected by local traffic flow and traffic accidents. However, the decrease in D area is relatively small, which may be due to the fact that D is a tourist destination with less local residents and less affected by local traffic flow. Through the above analysis, we can see that the error of Ta-1 model is much lower than that of seq2seq. On the one hand, it shows

Table 7 RMSE values of variant model in five districts of G City

Model	A	E	C	B	D
Seq2seq	0.0371	0.0398	0.0422	0.0388	0.0417
TA-1	0.0302	0.0341	0.0361	0.0283	0.0393
TA-2	0.0259	0.0281	0.0312	0.0265	0.0351
TA-3	0.0203	0.0237	0.0256	0.0246	0.0287
TA-4	0.0125	0.0143	0.0151	0.0154	0.0152
TA-STAN	0.0103	0.0132	0.0136	0.0141	0.0138

that the impact of various traffic flows on the prediction of future traffic accident risk is obviously different. On the other hand, it shows that the local traffic attention mechanism does capture the dynamic change of the impact of different traffic flows on future traffic accident risk.

3.4.2 Evaluation of global attention mechanism

As shown in Fig. 4, compared with Ta-1, the error of model Ta-2 after adding component GI decreases to a certain extent; TA-3 after adding group Gz on the basis of Ta-2 also has a significant loss reduction in all areas. The experimental results agree with the fact that the traffic conditions of the five regions affect each other. Especially in area a, with the addition of GI and GZ, the improvement of the model is particularly obvious. This may be due to the frequent commuting between area a and surrounding areas every day. Therefore, the dynamic impact of other areas on the traffic accident risk in area a can be well captured through the two spatial attention mechanisms.

3.4.3 Measurement of time attention mechanism

Continue to analyze Fig. 4. Compared with TA-3, the temporal attention mechanism in TA-4 greatly improves all regions, indicating that the temporal attention mechanism enables each moment in the future to obtain the correlation with historical moments. Since the temporal attention mechanism is to align the hidden layer units of the decoder and the encoder, the length of the encoder's time slice will affect the alignment effect. Therefore, this paper also carries out the experiment of recording the effect of temporal attention mechanism under different encoder lengths. The experimental results are shown in Fig. 4, where the decoder time length is 12. It can be found from Fig. 4 that basically the six models reach the lowest RMSE value when the encoder length is 24, indicating that there is obvious daily periodicity in traffic accidents. When the length of the encoder is greater than 30, the RMSE error increases with the increase in the length. However, the error values of TA-4 and ta-stan models are relatively stable and will not increase continuously, because both models contain temporal attention mechanism. It also shows that the temporal attention mechanism designed in this paper is very effective in finding the temporal correlation between encoder and decoder hidden layer.

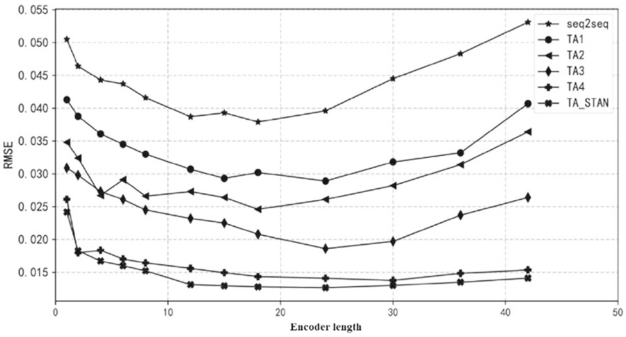

Fig. 4 Comparison of experimental results of variant models with different encoder lengths

3.4.4 Evaluation of external environment feature fusion module

As an empirical feature component, this part provides many additional features to improve the accuracy of experimental results. According to Fig. 4, the paper adds external features on the basis of TA-4. Under the condition of very low loss and low loss, the model decreases a part, and the decline range of each region is relatively close, which indicates that external factors are secondary factors relative to traffic indicators, but the effect is also obvious in improving the accuracy of the model.

3.5 Attention function experiment

This section discusses the effectiveness of different attention mechanisms (L, GI, GZ and T) using two different attention functions general and concat, hereinafter referred to as "general approach" and "concat approach". Since l, GI and GZ are the characteristics of different perspectives that affect the risk of traffic accidents, their mutual influence is not considered. In this experiment, when testing each attention function, other attention mechanisms use the optimal parameters and structure. Specifically, the four attention structures L, GI, GZ and T were evaluated under the four models of Ta-1, Ta-2, TA-3 and TA-4, corresponding to four sub-experiments. The experimental results are shown in Table 8.

By analyzing sub-experiment 1, sub-experiment 2 and sub-experiment3 with L, GI and GZ attention mechanisms, concat is better than general because general attention function itself is a single-layer neural network, which can better establish the nonlinear importance expression between encoder hidden layer and input data. Sub-Experiment 4 shows that for T attention mechanism, the loss of RMSE using General method is smaller, and General method is better than Concat method. T is considered to be the temporal attention mechanism, which is closely related to the length of the series prediction. Generally speaking, the encoder length is fixed, so the experiment further explores the relationship between T attention mechanism and decoder length. As shown in Fig. 5, there are two curves t-general and t-Concat in the figure, which, respectively, show the experimental results of T attention mechanism in the way of general and Concat. Through the

Table 8 The effect of attention function on attention mechanism

Experiment number	Model	Attention mechanism	Attention function	RMSE	MAE
Sub-experiment 1	TA-1	L	General	0.0314	0.0223
			Concat	0.0307	0.0217
Sub-experiment 2	TA-2	GI	General	0.281	0.0192
			Concat	0.0273	0.0183
Sub-experiment 3	TA-3	GZ	General	0.0239	0.0158
			Concat	0.0232	0.0149
Sub-experiment 4	TA-4	T	General	0.0156	0.0117
			Concat	0.0170	0.0125

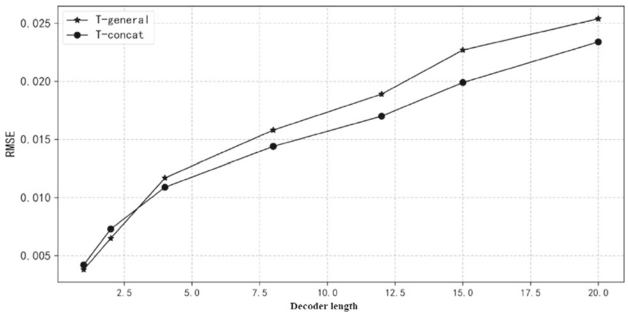

Fig. 5 Effect of attention function under temporal attention mechanism

analysis of Fig. 5, we can get two conclusions: (1) the two curves have the same trend, that is, with the increase in decoder length, the RMSE error values of the two curves are doubled, and the effect is the best when the encoder length is 1. This shows that the model is the most accurate in predicting the risk value of traffic accidents at the next moment in the future. With the increase in the prediction sequence length, the prediction effect is getting worse and worse. (2) When the length of decoder is short, Concat is better than General, but with the increase in decoder, General is better than Concat. This shows that General's method is more accurate in capturing the importance between history and future time slices. Considering the General method, a weight matrix is added between the two time hiding units, which is similar to the weight matrix structure of similarity calculation to improve the effect of temporal attention mechanism more obviously.

4 Example analysis

This section focuses on area numbers 25 and 144, showing the case of 6:00–18:00 on May 17, 2018. By putting the data of 6:00–18:00 into the encoder, the traffic accident risk in the next 12 h can be predicted. District 25 is the core financial district of Xiacheng A, and it is also the area where people gather during the day. District 144 is the residential district of B, so a large number of people commute between district 25 and 144 every day. Next, through the analysis of local spatial attention mechanism and temporal attention mechanism cases, to explain the practical significance of attention weight. Case 1: the visualization effect of the weight of local spatial attention mechanism is shown in Fig. 4–10. In the figure, the vertical axis represents the time stamp of the encoder, including 12 time stamps from the 0 th moment to the 11 th moment. The horizontal axis is the historical five traffic indicators, and from left to right are traffic accident risk, yellow taxi flow, green taxi flow, online car flow and bicycle flow. The depth of the color in the graph reflects the size of the attention coefficient. The darker the color, the greater the coefficient and the stronger the correlation. The local spatial attention mechanism studies the correlation between the time hidden unit in the encoder and the local traffic index input. The first is the common trend of the two regions: (1) the importance of traffic accident risk itself in the two time periods of 0–1 and 6–10 is similar to that of traffic flow, indicating that they are equally important. These two time periods correspond to the early morning of 6:00–7:00 and noon of 12:00–16:00 on the same day, respectively. There is no obvious directional flow trend of people. The current hidden layer state is determined by the local risk value and traffic flow. It is decided by value. (2) Time periods 2–6 and 10–12, corresponding to 8:00–12:00 (Fig. 6).

The importance of the four modes of transportation has increased in the morning peak and the evening peak from 16:00 to 18:00, while the importance of traffic accident risk has decreased, indicating that the current hidden layer state is mainly determined by the traffic flow. The different trends of the two regions are as follows: in area 25, the importance of green taxis is very low as a whole, while the importance of yellow taxis and online booking rate is higher; in area 144, on the contrary, the importance of yellow taxis is lower, and the importance of green taxis is higher. This is because the service volume of green taxis in a lower city is relatively small, and that of yellow taxis in Brooklyn is also relatively small. Through the above analysis, we can see that the local spatial attention

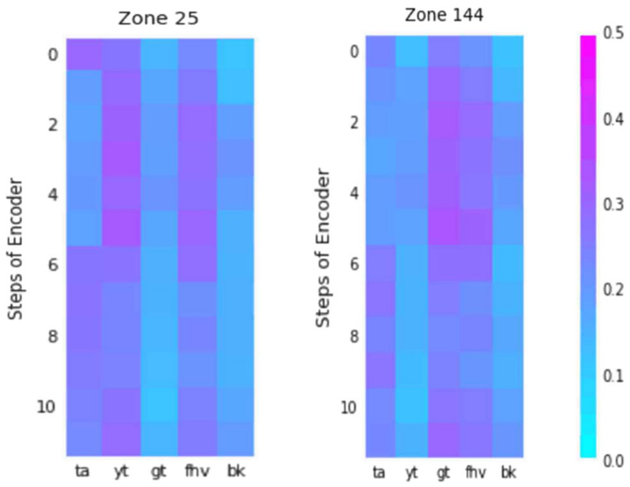

Fig. 6 Weight visualization of local spatial attention mechanism

mechanism can dynamically obtain the importance of local traffic indicators.

Case 2 the visualization effect of temporal attention mechanism weight is shown in Fig. 7, where the horizontal axis represents the decoder's time slice and the vertical axis represents the encoder's time slice, and the range of time slice is 0–11.

The closer the prediction time is, the more important the historical moment is. Of course, there are some special points. When the encoder time is 1–4, corresponding to 7–10 o'clock rush hour, and the decoder time is 0–2, corresponding to 18–20 o'clock rush hour, the time between encoder and decoder is also very important. This is because the traffic volume in rush hours can also reflect the traffic volume and traffic accident risk in rush hours. The decoder 6–11 is the midnight period from 0:00 to 5:00 the next day, which focuses on two parts: (1) encoder 10–11, corresponding to 17–18:00, focuses on the evening peak travel volume of the day. (2) Encoder 0–1 time corresponds to the previous morning. This may be the reason why the risk of traffic accidents in the early morning is similar to that in the middle of the night. Through the above analysis, we can see that the temporal attention accurately captures the dynamic temporal correlation of different time slices of decoder and encoder. In conclusion, by visualizing the attention weight value, people can make a reasonable explanation, which improves the interpretability of the model in this paper. From the results, we know our prediction is accurate, and the method is feasible and reliable.

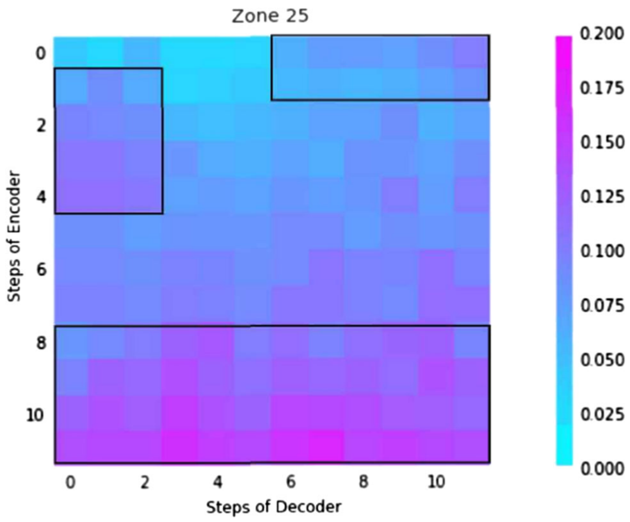

Fig. 7 Weight visualization of temporal attention mechanism

The part code is as follows (Table 9):

Table 9 The part code

```
from random import choice
from numpy import array, dot, random
1_or_0 = lambda x: 0 if x < 0 else 1
training_data = [ (array([0,0,1]), 0),
  (array([0,1,1]), 1),
  (array([1,0,1]), 1),
  (array([1,1,1]), 1), ]
weights = random.rand(3)
errors = []
learning_rate = 0.2
num_iterations = 100

for i in range(num_iterations):
  input, truth = choice(training_data)
  result = dot(weights, input)
  error = truth - 1_or_0(result)
  errors.append(error)
  weights += learning_rate * error * input

for x, _ in training_data:
  result = dot(x, w)
  print("{}: {} -> {}".format(input[:2], result, 1_or_0(result)))
```

5 Conclusion

This paper discusses a city-level traffic accident risk prediction model based on multiple data sources using deep learning modeling. Different from the previous study of traffic data using grid division method, this experiment directly uses the traffic administrative region as the statistical unit of the data. Meanwhile, in order to model the spatiotemporal characteristics of traffic, an encoder–decoder framework including spatiotemporal attention mechanism is used. In the multi-source heterogeneous data of this experiment, more attention is paid to the traffic volume data with higher correlation of traffic accidents, and the traffic volume is subdivided into multiple traffic volumes based on different vehicles. In order to better capture the dynamic impact of different traffic indicators on future traffic risk, this paper designs three attention mechanisms, namely local spatial attention mechanism, global spatial attention mechanism and temporal attention mechanism. Finally, in the prediction stage of decoder, the model integrates the influence of external environmental factors on traffic risk, which makes the prediction more accurate. Then, a series of experiments are carried out to verify the effectiveness and practicability of the model.

Funding Natural Science Basic Research Plan of Shaanxi Province (Project No. 2020JQ-682). Scientific Research Project of Education Department of Shaanxi Provincial Government (Project No. 18JK0450). Scientific Research Project of Education Department of Shaanxi Provincial Government (Project No. 21JK0718).

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work.

Ethical approval My paper does not deal with any ethical problems.

Informed consent We declare that all authors have informed consent.

References

Aung NT (2020) Research on traffic optimization and driving safety enhancement based on vehicle networking. University of Science and Technology Beijing

Bao YE (2020) Traffic flow prediction modeling and calculation based on neural network algorithm. Mod Electron Technol 43 (10):66-68+75

Feng W (2019) Research on intelligent traffic flow prediction technology based on spark. Shenyang University of Technology, MA thesis

Feng W, Guo Z (2018) Short term traffic flow prediction based on SVR. Sci Technol Innov Guide 15(25):189–190

Fucai J et al (2018) Application of unbiased grey model in ship traffic flow prediction of Dongying Port. J Guangzhou Inst Navigation 26(04):45–49

Fucai et al. (2018) J Meas Sci Instrum 9(04): 326–334

Guicheng S, Shuicheng G, Fangyu S (2018) Short term traffic flow prediction of Expressway considering optimal time delay factor spatiotemporal model. Sci Technol Eng 18(24):149–156

Guo Z (2019) Short term traffic flow prediction based on Markov random field. Shenyang University of Technology, MA thesis

Hao L (2020) Research on efficient prediction method for urban traffic flow at bayonet. 2020. Zhejiang University of Technology, MA thesis

Hongtao Y (2020) Signal timing optimization based on short term traffic flow prediction. Huaqiao University, MA thesis

Jie J (2020) Research and application of short term traffic flow prediction based on Improved BP neural network. Zhejiang University of Technology, MA thesis

Liu Y et al (2020) An improved trajectory planning algorithm based on deep learning. Softw Guide 19(06):15–18

Mingyu L et al (2018) Traffic flow prediction based on deep learning. J Syst Simul 30(11):4100-4105 + 4114

Pengfei Y, Xianbo S. Research on road icing prediction model based on support vector machine algorithm. J Hubei Univ Nationalities (Natural Science Edition), 2020, v.38; No. 107(03):118–123.

Qiu T (2020) Research on predictive control strategies for distributed economic models of nonlinear vehicle queueing systems. Zhejiang University of Technology

Ruiguang X, Shidong L (2018) Traffic flow prediction method based on bilinear recurrent neural network. Compr Transp 40(11):70–75

Shiju C (2019) Research on urban vehicle traffic flow analysis algorithm based on deep learning. Hebei University of Science and Technology, MA thesis

Wang WL, Pan LJ (2020) On the coupled traffic flow Aw-Rascle model Riemann problem. J Yantai Univ (Natural Science and Engineering Edition) v.33; No.123(04):4–12.

Wenqin P (2020) Research on short term traffic flow prediction technology based on deep learning. Chongqing University of Posts and Telecommunications, MA thesis

Xiangyu M et al. (2020) Research on short term traffic flow prediction of intelligent transportation. Proceedings of 2020 Wanzhi Scientific Development Forum (smart engineering II). Ed. proceedings of 2020 Wanzhi Scientific Development Forum (smart engineering II), 2020, pp 1012–1021

Yang Y (2020) Short term traffic flow prediction based on deep learning. Qingdao University of Science and Technology, MA thesis

Yang Y, Jun L, Long C, Xiaobo C, Ning Z, and Guodong H (2020) An emergency lane change behavior prediction method based on Gaussian hybrid hidden Markov model and artificial neural network. China Mech Eng v.31; No. 551(23):106–114+122.

Yongqiang Z, Xiaofan W (2020) Traffic flow prediction model based on complementary integrated empirical mode decomposition and genetic least squares support vector machine. Sci Technol Eng 20(17):7088–7092

Youkang Z, Honglei W (2020) Short term road traffic flow prediction based on deep learning. Software 41(05):72–74

Yuchan Y et al (2019) Short term traffic flow prediction using MPSO optimized SVR. Comput Technol Dev 29(04):133–138

Zhang BX (2020) Research and analysis of highway accident model with high bridge-tunnel ratio based on multiple nonlinear regression method. Highw Eng 2020, v.45; No.200(01):48–53.

Zihua S et al (2020) Short term traffic flow prediction based on improved WNN. Comput Digit Eng 48(07):1617–1622

Zijing H (2020) Study on multi period exit flow prediction method of Expressway multi toll station based on spatiotemporal attention mechanism. South China University of Technology, MA thesis

Zuogam T, Yeyao S, Ying G, Fengquan Y (2020) Modeling the number of fatalities and injuries in street traffic accidents based on multi-source data fusion. The 15th annual china intelligent transportation conference

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.