

Statistical Analysis of Urban Traffic Flow Using Deep Learning

Quanzhi Liu,^{1,*} Shuang Wu,² Peng Zhang³

¹Department of Mathematics and Statistics, Cangzhou Normal University, Cangzhou, Hebei 061001, China

²Department of Planning and Finance, Cangzhou Normal University, Cangzhou, Hebei 061001, China

³Department of Marxism, Cangzhou Normal University, Cangzhou, Hebei 061001, China

E-mail: xunzhi05qhcjo@yeah.net

*Corresponding author

Keywords: graph convolutional network model, long short-term memory network model, urban road condition, traffic flow

Received: October 31, 2023

In recent years, urbanization has brought about challenges such as population growth, increased demand for traffic, and traffic congestion. To address the need for accurate traffic condition statistics, this paper proposed an improved method that combines graph convolutional network (GCN) and long short-term memory network (LSTM) model for forecasting and statistics of traffic conditions. Through the modeling and analysis of urban road conditions and traffic flow, the combination of GCN model and LSTM model enabled more precise prediction of traffic flow trends. Experiments were carried out on the actual traffic data set of Cangzhou, Hebei. The results demonstrated that the proposed method achieved high accuracy and reliability in predicting traffic flow. By using the LSTM model to improve the GCN model, it effectively adapts to changes in urban traffic conditions while providing dependable predictions.

Povzetek: Članek predlaga izboljšano metodo, ki združuje grafično konvolucijsko mrežo (GCN) in dolgoročno kratkoročni model spominskega omrežja (LSTM) za napovedovanje in statistiko prometnih razmer.

1 Introduction

In the field of transportation, there has been notable advancements in deep learning technology in recent times, among which graph convolutional network (GCN) has attracted much attention as an effective graph data modeling method. However, the traditional GCN model encounters challenges such as low accuracy and substantial prediction errors when dealing with traffic prediction. Consequently, research focused on enhancing the prediction accuracy of the GCN model has emerged as a prominent topic within the current academic community. In order to improve the traditional GCN model and improve the accuracy of urban traffic flow statistics, this paper introduced the long short-term memory (LSTM) model into the GCN model. By integrating urban road conditions and traffic flow data, this approach achieves more precise and reliable traffic flow prediction. Furthermore, empirical evaluation using real-world traffic dataset from Cangzhou, Hebei Province demonstrates that the proposed GCN-LSTM model outperforms other models in terms of prediction accuracy and precision.

2 Related works

Table 1: A summary of related works

Literature	Model type	Recognition method	Finding
Peng et al. [1]	A memory augmented graph	Predict traffic	It solved the problem of

	convolutional network (MA-GCN) model	networks by combining graph convolutional network (GCN) and differential neural computer (DNC)	long-term dependence and spatial dependence in traffic prediction.
Bao et al. [2]	Prior knowledge enhanced time-varying graph convolutional network (PKET-GCN)	Considering dynamic and static features and using prior knowledge to improve the correlation of nodes, combined with convolution and a time-varying feature extraction module	PKET-GCN showed superior results over existing methods on multiple real datasets.
Diao et al. [3]	A novel deep learning	Using a graph with specific local	EC-GCN outperformed the state-of-the-

	framework called EC-GCN	structure to represent each encrypted traffic type, and introducing a lightweight layer and a graph pool learning layer	art methods on three real data sets, improving the classification accuracy by 5%-20%.
Zhang et al. [4]	Combine fuzzy C-means (FCM) and GCN to create a hybrid model	FCM was used to cluster the trajectory data to generate the air traffic flow network graph structure, and GCN was used to detect trajectory deviation and delay.	FCM-GCN was superior to other models in medium and long-term traffic prediction tasks, which is helpful to optimize the highly interconnected and interdependent air traffic flow network.

capture structural information among nodes, which in turn enables representation and inference of node features. Through multi-layer graph convolution operations, the GCN model can learn more advanced features, thus enhancing robustness in tasks such as edge prediction, node identification, and network analysis of graph data. The GCN model can be applied to various graph structures, including social networks, recommendation systems, bioinformatics, and other fields.

The GCN model has been widely used in traffic-related tasks, including flow prediction, road condition analysis, and road network modeling. By modeling the topology structure and node characteristics of the traffic network, it offer more accurate traffic prediction and analysis results for road network. Through iterative updates to node representations, the direct neighbors of the node are considered in each iteration, and their feature information is aggregated and updated. The way of neighbor aggregation and multi-layer stacking enables nodes to use their location relationships within the graph to acquire richer information. Consequently, it becomes possible to capture contextual relationships among nodes within the graph structure while effectively addressing data analysis tasks involving graph structures [7]. The update formula for each node and layer of the model is as follows:

$$H^{l+1} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l \right),$$

$$\tilde{A} = A + I_N,$$

$$\tilde{D} = D + I_N,$$

where H^{l+1} is the node representation of the l+1-th layer, $H^{(l)}$ is the node representation after the processing of layer l, and W^l represents the weight matrix of the l-th layer. \tilde{A} represents the sum of the adjacent matrix and unit matrix, i.e., $\tilde{A} = A + I$. \tilde{D} stands for the degree matrix of \tilde{A} .

This way of information transmission employed by the GCN model facilitates the comprehensive analysis of traffic data patterns and characteristics from both a global and local perspective, taking into account the topological structure of the traffic network as well as node relationships [8]. However, it falls short in effectively capturing long-term time dependencies when dealing with time series data. However, the LSTM model is very good at processing time series data and can effectively capture long-term time dependencies, enabling retention of past information for accurate future predictions. Therefore, the LSTM model is incorporated into the GCN model to effectively capture and retain historical traffic flow information, enabling consideration of previous states and trends in prediction. This comprehensive modeling of spatio-temporal relationships allows for better adaptation to dynamic changes and complexity in traffic flow data, resulting in accurate predictions of traffic flow and road conditions. Ultimately, this improves traffic efficiency and reduces congestion.

3 Traffic flow and GCN model improvement

3.1 Traffic flow

The term “traffic flow” refers to the density and flow degree of moving objects within a transportation network during a specific time period. It serves as an important indicator for measuring the traffic congestion levels and road utilization. Traffic flow exhibits spatio-temporal correlation, periodicity, and uncertainty. It is possible to achieve data processing and feature engineering with appropriate model algorithms and evaluate and optimize predictive models by using historical traffic flow data as training samples in conjunction with on-site observation data, environmental factors, and road network structure information [5]. These models can analyze historical data and train models using statistical methods, machine learning, artificial intelligence, and other technologies for predicting future traffic flow indicators. They assist transportation management departments in formulating effective traffic strategies for improving traffic mobility and reduce congestion issues.

3.2 GCN model

The GCN model is a deep learning model designed for the modeling and analysis of graph structure data. By considering the interconnections between nodes and their neighbor nodes, it realizes the feature extraction and information dissemination propagation within the graph data. It applies convolutional operations on the graph to

3.3 LSTM model

The LSTM is a variant of temporal RNN that solves the issues of gradient explosion and vanishing in traditional RNNs by introducing input gates, forget gates, and output

gates [9]. LSTM adjusts the degree of information storage and forgetting in both long-term memory and short-term memory by setting a threshold, while introducing cell state to save long-term memory information [10]. This method effectively overcomes the gradient problem encountered in RNN and improves the stability of the training model.

The forget gate filters the cell information of the previous time step and determines which information needs to be discarded. The formula is:

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

This gate integrates information h_{t-1} transmitted in the previous stage and input data x_t at the current moment, so as to fuse them together in the calculation process, control whether the input information at the current time step enters the memory unit, and thus affect the memory and information flow of the network when processing sequence data. The input gate controls what information from the current input and the hidden state at the previous time step should be added to the cell state, and the corresponding formulas are:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= \text{thah}(W_C \cdot [h_{t-1}, x_t] + b_i), \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t. \end{aligned}$$

By filtering the forgotten information, the input gate combines the input data of the current time step with the hidden state of the previous time step at time step t to obtain candidate cell state C_t . The output of the input gate is then multiplied by the candidate cell state, and the obtained result is added to the previous memory state to update the memory state at the current time step. The output gate controls the flow of information between the input at the current time step and the memory at the previous time step, and between the input at the current time step and the output at the current time step. The corresponding formulas are:

$$\begin{aligned} O_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\ h_t &= O_t * \tanh(C_t), \end{aligned}$$

where σ and \tanh are both activation functions, x_t represents the input data at time t , h_{t-1} is the hidden state at time step $t-1$, C_t is the cell state at time t , and \tilde{C}_t is the cell state value at time step $t-1$. The forget gate, input gate, output gate, and neuron state all have parameter matrices with values between 0 and 1, which are denoted by W_f , W_i , W_C , and W_o , respectively. b_f , b_i , b_o , and b_e are bias vectors corresponding to matrices. f_t , i_t , and O_t represent threshold values corresponding to the three gates [11].

3.4 Improved GCN model

The GCN model performs well in handling nonlinear data and capturing complex dependencies between nodes, while the LSTM model is good at capturing long-term dependencies in time series data. These two models are integrated to form a new approach known as the GCN-LSTM model. This model operates by iteratively alternating between the GCN and LSTM models, continuously updating node representations and time-dependent information. By calculating the gradient of the parameters to the loss function and updating the model parameters according to the gradient, the model

progressively adapts to training data and improves performance [12]. The steps are as follows.

(1) The GCN model is used for conducting representation learning on graph data, and the representation of each node is updated by the adjacency matrix and node features.

(2) The spatial feature representation extracted by the GCN model is fed into the LSTM model, which uses its memory unit and gating mechanism to capture the time dependence in sequence data, thereby further extracting crucial time-related information.

(3) The GCN-LSTM model undergoes training with temporal data across multiple time steps. At each step, the hidden state is updated and adjusted to make predictions regarding future traffic flow or other relevant information.

4 Experimental analysis

4.1 Data acquisition

In order to evaluate the GCN-LSTM traffic flow prediction model constructed in this paper, experimental verification was conducted. Data on vehicle flow from Qinxue Road, Wenlan Street, and Fengfan Road in the central urban district of Cangzhou City were collected from the website of the Ministry of Communications of the People's Republic of China during October 1 to October 31, 2021. The statistical time period ranged from 00:00 to 24:00, with a time interval of every 10 min. Finally, 1,500 records were obtained as the data set. The actual road images were collected from several angles and time periods, and the resulting images were subjected to frame-splitting and data cleaning, and manually labeled with automobiles, pedestrians, and non-motorized vehicles. The GCU and GCN models were selected as control objects, and the mean absolute percentage error (MAPE) was used to evaluate the prediction performance of the test models.

4.2 Evaluation indicators

Due to the complexity of the traffic system, the MAPE can be used as the evaluation indicator for assessing the performance of the GCN-LSTM prediction model. MAPE represents the average degree of deviation between predicted and actual vehicle flow, with a smaller value indicating a stronger correlation [13]. The formula is:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100\%,$$

where n is the number of observations, \hat{y}_i is the predicted value of the model, and y_i is the true data value.

4.3 Experimental results

By comparing the MAPE of the GRU, GCN and GCN-LSTM models as illustrated in Figure 1, it is evident that the improved GCN-LSTM model exhibited exceptional prediction performance for Cangzhou's traffic flow dataset with a significantly low MAPE value.

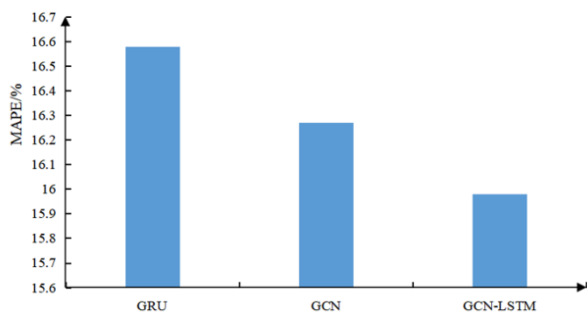


Figure 1: Performance comparison results of various models.

As shown in Figure 2, the vehicle flow in the central urban area showed obvious periodic changes. With the exception of holidays, this cycle occurred consistently throughout all seven days of the week, and each cycle displayed similar characteristics regarding vehicle flow. Notably, weekdays witnessed a substantial volume of vehicles, while non-weekdays experienced relatively lower traffic levels.

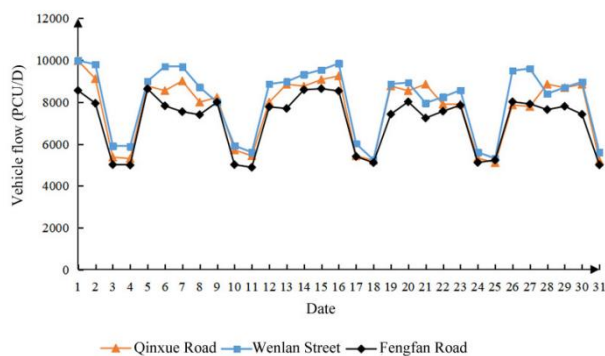


Figure 2: The change trend of traffic flow in Cangzhou City in October.

After analyzing the monthly change characteristics of vehicle flow in the central urban area, it can be found that the change of the vehicle flow in the central urban area was periodic from Monday to Sunday. The vehicle flow data of the central urban area of Cangzhou City from October 18 to 24, 2021, which represents a complete week excluding holidays, were selected to analyze the variation distribution characteristics of vehicle flow in the central urban area within a week (from 00:00 to 24:00) (Table 2).

The data in Table 2 reveals that during the period from October 18 to 22, there was a low traffic flow in the early morning hours, which remained relatively stable at around 2,000 PUC/D. However, it experienced a significant surge after 06:00 and reached its peak between 06:00 and 08:00. Subsequently, there was a decrease in traffic flow from 10:00 to 16:00 followed by an increase reaching its maximum at 18:00. On October 23 and 24, compared to regular working days, the peak traffic volume was delayed by two hours. Consequently, it can be concluded that there

were two distinct peaks of traffic flow in the central area. The periods of 06:00-08:00 and 16:00-18:00 exhibited the highest traffic flow on weekdays, while the peak traffic flow on non-weekdays occurred between 08:00-10:00 and 18:00-20:00. Generally speaking, weekdays witnessed a relatively substantial volume of vehicle flow.

Table 2: The seven-day change distribution of the vehicle flow in the central urban area of Cangzhou City

Date Time slot	10/ 18	10/ 19	10/ 20	10/ 21	10/ 22	10/ 23	10/ 24
00:00-02:00	11 08	10 91	11 95	10 67	11 52	98 5	99 6
04:00-06:00	23 59	24 60	14 27	23 96	24 05	24 56	23 90
06:00-08:00	94 59	96 30	96 48	96 33	97 16	75 11	76 08
08:00-10:00	86 48	87 51	87 06	88 00	86 04	87 15	86 12
10:00-12:00	75 49	78 36	80 61	83 02	82 87	89 60	88 95
12:00-14:00	72 23	73 65	71 62	75 71	76 00	72 93	72 47
14:00-16:00	78 75	77 58	76 89	82 01	84 30	72 23	73 16
16:00-18:00	97 42	98 97	98 00	99 15	96 75	79 87	80 02
18:00-20:00	81 71	81 00	83 65	80 09	84 12	86 50	88 32
20:00-22:00	77 12	80 91	83 21	82 09	84 12	61 53	62 21
22:00-24:00	31 12	32 08	33 69	36 1	32 19	22 01	21 17

Because there is a time-dependent relationship between vehicle flow, vehicle flow on weekdays and non-weekdays by extracting time features [14]. Therefore, when the GCN-LSTM model makes a prediction, it is necessary to determine the input and output of the model beforehand [15]. The input data is mainly composed of the following features: historical traffic flow, month, specific date, travel time period, working day indicator, and legal holidays (0 means yes, 1 means no). The known traffic data for historical dates was selected as the input, and the date to be predicted was input. For example, the traffic flow related characteristics of Qinxue Road from January 1, 2023 to July 31, 2023 was selected as input to predict the vehicle flow between 6:00 and 20:00 on this road section on August 1, 2023. Subsequently, the predicted vehicle flow value was compared with the actual value (Figure 3).

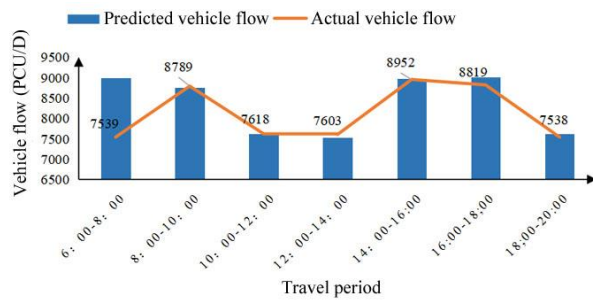


Figure 3: Predicted and actual traffic flow values of Qinxue Road in Cangzhou on August 1.

The predicted vehicle flow of Qinxue Road in Cangzhou City during the morning peak hours of 06:00-08:00 on August 1, as depicted in Figure 3, exhibited a significant deviation from the actual value. It is worth noting that this time period typically experiences heavy traffic volume with an average of approximately 9,600 PCU/D based on historical data. However, it can be seen from Figure 3 that the actual traffic flow on August 1 was 7,539 PCU/D, which exhibited a decrease compared to the previous flow, and the morning peak time was also postponed. Consequently, it can be inferred that congestion occurred on Qinxue Road between 06:00 and 08:00 on August 1st.

5 Discussion

Urban transportation is easily affected by various factors such as human intervention and traffic control. To cope with congestion caused by peak travel and traffic restrictions, it is essential to have real-time monitoring of traffic flow and predictions for future time periods. Traditional GCN models have several limitations in the field of transportation. For instance, these models can only extract spatial features in a transportation network and may not effectively consider the temporal changes in feature processing for spatiotemporal data. Traditional GCN models tend to provide deterministic predictions, which limits their robustness in practical applications due to the existence of uncertainties in real traffic networks. However, the experimental results showed that by leveraging the temporal characteristics of LSTM, it effectively modeled the time dimension data and reduced the average absolute percentage error between predicted and actual traffic flow. This approach better captures the temporal variations in traffic volume, comprehensively considers both spatial and temporal features, enhances understanding of dynamic patterns in traffic data, and ultimately improves the accuracy of traffic flow prediction. In actual urban traffic flow prediction, by comparing with historical data, we can more accurately assess the future changes in city road conditions. By comparing historical data, future changes in city road conditions can be more accurately assessed in actual urban traffic flow prediction. Therefore, the GCN-LSTM model was used to predict the traffic flow in real time, and the result was compared with the historical data. The change

of road conditions was judged by analyzing the difference between the predicted results and the actual observation values. If the difference between the predicted value and the actual value is large, it may mean that there is congestion or other abnormal conditions. Based on these prediction results and difference analysis, traffic management departments can take timely measures to optimize road traffic safety and efficiency and maximize the utilization of road resources. According to the predicted traffic conditions, people can adjust the appropriate means of transportation and formulate more reasonable travel routes at any time, so as to avoid congestion and improve travel efficiency.

6 Conclusion

In this paper, the LSTM model was introduced to improve the GCN model. An experimental application was carried out using real traffic data of Cangzhou, Hebei Province. By training and verifying the GCN-LSTM model, the traffic flow in different time periods was successfully predicted, and the statistical analysis was carried out. The results showed that the improved GCN-LSTM model significantly improved the prediction accuracy and precision, enabling more accurate forecasting of traffic flow fluctuations and providing valuable support for travel route and mode planning. Furthermore, it can also assist traffic management departments in conducting traffic dispatch and monitoring more effectively, thereby reducing congestion and improving road efficiency. Additionally, it provides valuable reference data for urban transportation planning and management. In the future, we will further optimize and improve the GCN-LSTM model to verify its applicability in a wider range of scenarios, aiming to continuously enhance the efficiency and safety of urban traffic.

References

- [1] Peng D, Zhang Y (2023). MA-GCN: A Memory Augmented Graph Convolutional Network for traffic prediction. *Engineering Applications of Artificial Intelligence*, 121. <https://doi.org/10.1109/TITS.2019.2935152>
- [2] Bao Y, Liu J, Shen Q, Cao Y, Ding W, Shi Q (2023). PKET-GCN: Prior knowledge enhanced time-varying graph convolution network for traffic flow prediction. *Information Sciences*, 634, pp. 359-381. <https://doi.org/10.1016/j.ins.2023.03.093>
- [3] Diao ZL, Xie GG, Wang GX, Ren R, Meng XY, Zhang GX, Xie K, Qiao M (2023). EC-GCN: A encrypted traffic classification framework based on multi-scale graph convolution networks. *Computer Networks*, 224. <https://doi.org/10.1016/j.comnet.2023.109614>
- [4] Zhang YN, Lu Z, Wang J, Chen L (2023). FCM-GCN-based upstream and downstream dependence model for air traffic flow networks. *Knowledge-Based Systems*, 260. <https://doi.org/10.1016/j.knsys.2022.110135>

- [5] Wei G, He S, Ma J (2012). Review on Traffic Flow Phenomena and Theory. *Journal of Transportation Systems Engineering and Information Technology*, 12, pp. 90-97. [https://doi.org/10.1016/S1570-6672\(11\)60205-5](https://doi.org/10.1016/S1570-6672(11)60205-5)
- [6] Lee K, Rhee W (2022). DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting. *Transportation Research Part C: Emerging Technologies*, 134. <https://doi.org/10.48550/arXiv.1905.12256>
- [7] Yu B, Lee Y, Sohn K (2020). Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN). *Transportation Research Part C: Emerging Technologies*, 114, pp. 189-204. <https://doi.org/10.1016/j.trc.2020.02.013>
- [8] Zheng Y, Wang S, Dong C, Li W, Zheng W, Yu J (2022). Urban road traffic flow prediction: A graph convolutional network embedded with wavelet decomposition and attention mechanism. *Physica A: Statistical Mechanics and its Applications*, 608. <https://doi.org/10.1016/j.physa.2022.128274>
- [9] Li Z, Xu H, Gao X, Wang Z, Xu W (2023). Fusion attention mechanism bidirectional LSTM for short-term traffic flow prediction. *Journal of Intelligent Transportation Systems Technology Planning and Operations*. <https://doi.org/10.1080/15472450.2022.2142049>
- [10] Narmadha S, Vijayakumar V (2023). Spatio-Temporal vehicle traffic flow prediction using multivariate CNN and LSTM model. *Materials Today: Proceedings*, 81, pp. 826-833. <https://doi.org/10.1016/j.matpr.2021.04.249>
- [11] Yang BL, Sun SL, Zhang KL, Li JY, Tian Y (2019). Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing*, 332, pp. 320-327. https://doi.org/10.1007/978-981-10-0451-3_1
- [12] Wang S, Zhang Y, Hu Y, Yin B (2023). Knowledge fusion enhanced graph neural network for traffic flow prediction. *Physica A: Statistical Mechanics and its Applications*, 623. <https://doi.org/10.1016/j.physa.2023.128842>
- [13] Song J, Kang H, Hyun S, Jee E, Bae DH (2022). Continuous verification of system of systems with collaborative MAPE-K pattern and probability model slicing. *Information and Software Technology*, 147. <https://doi.org/10.1016/j.infsof.2022.106904>
- [14] Hou R, Wang Z, Ren R, Cao Y, Wang Z (2023). Multi-channel network: Constructing efficient GCN baselines for skeleton-based action recognition. *Computers & Graphics*, 110, pp. 111-117. <https://doi.org/10.1016/j.cag.2022.12.008>
- [15] Wang G, Zhang Z, Bian Z, Xu Z (2021). A short-term voltage stability online prediction method based on graph convolutional networks and long short-term memory networks. *International Journal of Electrical Power & Energy Systems*, 127, pp. 1-9. <https://doi.org/10.1016/j.ijepes.2020.106647>