Short-term Passenger Flow Prediction of Subway based on Neural Network LSTM Algorithm

Minyou Wei [1], Tianle Yang [2,*] and Yu Zheng [3]

[1] Aberdeen Institute specializing in the fields of Data Science and Artificial Intelligence, South China Normal University, Guangzhou, 528225, China

[2] Chang'an Dublin International College of Transportation, Chang'an University Xi'an, 710064, China

[3] Artificial Intelligence, Shanghai Normal University TIANHUA College, Shanghai, 201815, China

* Corresponding Author Email: tianle.yang1@ucdconnect.ie

Abstract. The present study investigates the application of the Long Short-Term Memory (LSTM) neural network algorithm for short-term passenger flow forecasting in subway systems. The dataset, sourced from Hangzhou subway stations, records passenger flow at 10-minute intervals over 30 days, distinguishing between weekdays and weekends. The LSTM model was trained using 80% of the available data, while the remaining 20% was reserved for testing purposes. Results showed the model achieved a mean squared error (MSE) of 19.33 on training data, indicating a good fit. However, the test data MSE was significantly higher at 460.78, suggesting overfitting. This implies the model captures training data patterns well but struggles with generalizing to new data. The findings underline the potential of LSTM networks in handling long-term dependencies and gradient vanishing issues inherent in traditional Recurrent Neural Networks (RNNs). Despite these advancements, the study highlights the need for model refinement to address overfitting and improve generalizability. Future research should focus on data augmentation, model enhancement, and regularization techniques to develop more robust prediction models. Accurate short-term predictions of passenger flow play a crucial role in optimizing train scheduling, resource allocation, and overall service quality within urban rail transit systems.

Keywords: Revising the prediction of passenger mobility; LSTM; Artificial neural network.

1. Introduction

Given the rapid pace of urbanization, the implementation of urban rail transit has emerged as an indispensable measure for alleviating traffic congestion in major metropolitan areas. Forecasting short-term traffic flow—quantitatively anticipating the traffic situation in the near future—has become a crucial area of study in transportation research with the advent of intelligent transportation systems in recent decades [1]. Predicting the flow of passengers in subway transportation using big data is of immense importance for devising effective strategies and enhancing urban services. By leveraging vast datasets and advanced analytical techniques, cities can better understand patterns and trends in subway usage. This understanding enables the development of targeted interventions to optimize operations, reduce congestion, and improve the overall efficiency and quality of urban transit systems. Additionally, accurate passenger flow predictions contribute to better resource allocation, emergency response planning, and long-term infrastructure development, ultimately leading to a more resilient and responsive urban environment [2]. Passenger flow forecast is the essential part of the transportation organization, only according to accurate passenger flow predictions can help operators develop scientific train scheduling plans, optimize resource allocation, enhance service quality, and reduce operational costs [3]. Additionally, short-term passenger flow prediction can provide passengers with more precise travel advice, improving their travel experience.

Early research evaluated the historical average, time series, and neural network and found that the nonparametric regression model outperformed the other models for this topic [4]. Zhang and Yang suggested a spline function for flow prediction. These prediction models consistently produce predictions that are not satisfactory due to their irregular and stochastic features [4]. The increasing

popularity of deep learning has led to numerous studies attempting to apply state-of-the-art models such as long short-term memory networks (LSTM), encoder-decoder LSTM, convolutional LSTM (ConvLSTM), and deep residual networks (ResNet) in the field of STPFF [5]. Most of these researches are based on LSTM. The LSTM has been used in feedback neural networks for 21 years, which is a type of recurrent neural network (RNN) that has been developed and optimized based on RNN architecture [5, 6]. The success of a typical RNN is attributed to its ability to establish connections between earlier information and the current task. However, if there is a significant gap between the relevant information and the point where it is needed, the RNN struggles to learn how to make these connections due to a critical issue known as long-term reliance. This issue manifests itself through the vanishing and inflating of gradients [5].

Unlike traditional RNNs, the LSTM can fortunately avoid the problem of long-term reliance and gradient vanishing by incorporating memory cells and forgets gates [6, 7]. The activation of the cell will not be overwritten by new inputs arriving in the network as long as the input gate remains closed (i.e. has an activation near 0). Therefore, it can be preserved and made available to the network at a later stage in the sequence through the operation of opening the output gate [8]. In another research, the result of study also confirmed that all convolution-based deep learning models outperform recurrence-based models in the condition of utilizing a single model for network predictions [9]. However, LSTM performed better than other basic machine learning methods when the data dimension was small. It’s hard to consider the prediction accuracy of LSTM is the best when importing huge amounts of big data [10].

LSTM neural networks are highly advantageous for experiential learning and similar tasks due to their effective regulation of information flow within the cell state. The objective of this article is to develop an appropriate model that can enhance the precision of forecasting short-term fluctuations in passenger traffic within subway stations. The primary approach involves utilizing the LSTM algorithm to construct a model capable of forecasting short-term variations in subway passenger flow using relevant station data. Following data cleansing and preprocessing, LSTM will be employed to establish the model, which will then undergo evaluation to ensure its precision.

2. Methods

2.1. Data Source and Description

The subway ridership data comes from publicly available data from the Hangzhou subway station (download from AI studio website), which records the number of people entering and exiting a station of one line every 10 minutes for 30 days in a month, and distinguishes between weekdays and weekends.

Figure 1 shows the actual figure in the station, the abscissa is the time (the nth ten minutes), and the ordinate is the passenger flow as time going by. The table demonstrates the passenger flow in a week. After initial processing of the data, this paper keeps a seven-day record of footfall data. Some of the useless metrics in the data are culled, and only the time (every ten minutes) and inbound traffic are retained, we use the tool in excel to create the line chart (Fig. 1) the figure shows that footfall shows a double-peak trend on weekdays and a single-peak trend on weekends, and in the midnight, the number of footfalls is 0. This data is a good example of the trend of classic short-term passenger flow data, and There won't be much to change, and the characteristics of the data are very obvious, so it can be used for the data in this study.

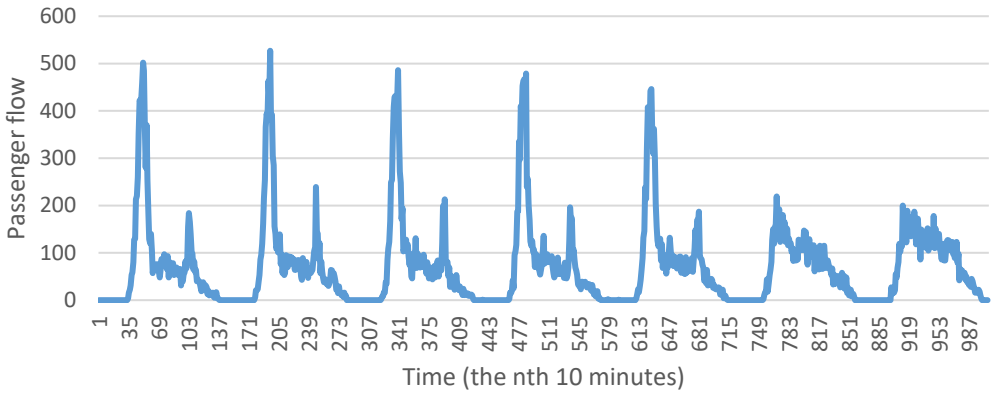


Fig 1. subway ridership passenger flow in a week

2.2. Indicator Selection and Description

The selection of time granularity presents a significant challenge in predicting short-term passenger flow, as it directly impacts the accuracy of predictions. Opting for a 10-minute time period enhances the representation of short-term passenger flow, aligning with the raw dataset's intervals and reducing workload. Moreover, employing appropriate time separation and leveraging sufficient data volume effectively improves algorithm efficiency while minimizing training time.

2.3. Method Introduction

For this research, this paper will utilize the memory with long and short-term capabilities recurrent neural networks (LSTM) algorithm to analyze the data, considering its effectiveness in addressing the long-term dependence issue compared to RNN models that face challenges of gradient vanishing or explosion during training. The dataset will be divided into 80% for model training and 20% for prediction purposes. By employing the LSTM algorithm, this paper can obtain predicted values which will be compared with the original data. To assess accuracy and mitigate underfitting or overfitting risks, this paper performs analysis of variance, mean square error (MSE), and Loss Value calculations. This evaluation process enables us to determine if the LSTM model is suitable for achieving passenger flow in the near term forecasting goals. The Mean Squared Error (MSE) is employed as the loss function in this study and can be calculated using the following formula. The formula for the MSE loss function is:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

The training and testing mean squared error (MSE) values are reported, indicating the utilization of this metric for evaluating the performance of the LSTM model on both training and testing datasets. In summary, the loss function (MSE) quantifies the average squared deviation between predicted and actual values, facilitating an assessment of model accuracy and its vulnerability to overfitting.

3. Results and Discussion

3.1. Model Results

After conducting 300 training sessions, this paper obtained the results shown in Figure2. The mean square error (MSE) for training was found to be 19.329, while the MSE for testing was 460.779. Additionally, the Loss Value is approximately 0.002. It is worth noting that the training MSE falls below 100, indicating a strong fit between the training data and actual data. However, it should be acknowledged that the test MSE exceeds 400, suggesting a suboptimal fit between predicted and actual data and some degree of overfitting.

The findings suggest that although the LSTM model demonstrates some capability in forecasting short-term passenger flow, its elevated test MSE raises doubts regarding its applicability. The overfitting problem implies that the model may have learned patterns specific to the training data, which do not generalize well to new, unseen data.

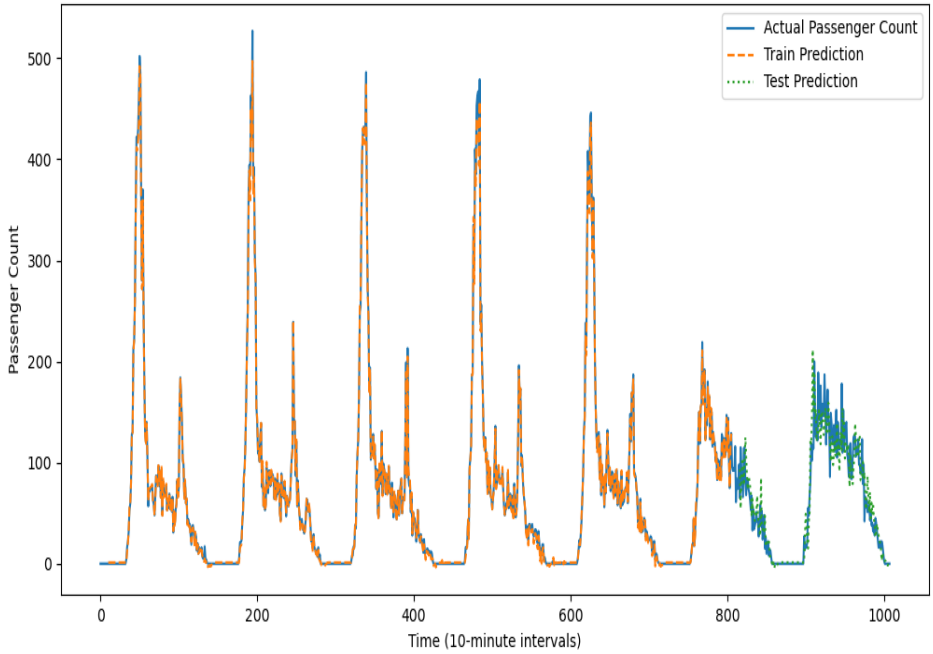


Fig 2. LSTM training and prediction graph

In Figure 2, the horizontal axis denotes the time (every 10 minutes), while the vertical axis represents the passenger count. The continuous line corresponds to the real-time data of passenger flow, whereas the long-dashed line indicates 80% of the training dataset and the short-dashed line signifies projected data. The total training time of 30 minutes for 300 epochs is acceptable. In fact, for algorithms with fewer layers and larger time spans, although they have an advantage in training speed, their two MSE values are above 300 on average, and the loss degree is above 0.06, which is a bad result. The three levels taken in this experiment are the best results obtained by combining training time, fit and loss value.

3.2. Discussion

From the results obtained from various datasets and algorithms, there are several limitations to consider, including overfitting, data granularity, limited data scope, computational efficiency, and evaluation metrics. Here are some of the more common limitations observed in this study. The presence of overfitting is indicated by a significant disparity between the mean squared error (MSE) calculated on the training set and that on the test set. This phenomenon can occur when the model becomes excessively complex relative to the available training data or when regularization techniques are not adequately utilized and noise points are frequently picked up during training. While it is considered standard practice to select a 10-minute interval for prediction purposes, it should be noted that the model's performance may vary depending on the chosen granularity. Opting for shorter intervals allows for capturing more intricate patterns, while longer intervals tend to smooth out noise but might overlook finer details. It is important to mention that this study solely relies on data from a single metro line and covers only a period of 7 days. The limited scope of this study may not fully capture the variability in passenger flow patterns across different metro lines, seasons, or special events. The analysis is based solely on data from a single metro line and for a restricted duration of only 7 days. This narrow focus might fail to encompass the complete range of observed passenger

flow patterns across various lines, seasons, or special events. While the primary evaluation metric is Mean Squared Error (MSE) which provides valuable insights, it would be beneficial to incorporate additional metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) for a more comprehensive understanding of the model's performance. The study recommends LSTM over RNN due to concerns regarding gradient vanishing in RNNs; however, it does not compare LSTM with other potential models such as ARIMA, Prophet, or alternative machine learning techniques suitable for time series prediction.

4. Conclusion

The aim of the study was to develop a prognostic framework for forecasting passenger movement in subway networks by utilizing the Long Short-Term Memory (LSTM) neural network algorithm. The model underwent thorough data cleaning and preprocessing before being trained and assessed on Hangzhou subway station's ridership data. Results showed that while the LSTM model performed well during training with a mean square error (MSE) of 19.33, it exhibited overfitting when tested, resulting in a significantly higher MSE of 460.78. This suggests that although the LSTM model can identify patterns within training data, it struggles to apply these findings to new or unseen information.

This study contributes to the growing body of literature on urban rail transit and intelligent transportation systems. The ability to predict short-term passenger flow accurately can help transit authorities optimize train scheduling, enhance resource allocation, and improve overall service quality. The use of LSTM networks, which address the issues of long-term dependencies and gradient vanishing inherent in traditional RNNs, represents a significant advancement in the field of transportation forecasting. However, the observed overfitting highlights the need for further refinement and validation of the model. This finding underscores the importance of balancing model complexity with generalizability, particularly when dealing with real-world data that may contain noise and other irregularities.

Future investigations should prioritize various key aspects such as the augmentation and diversification of data, along with enhancing models and implementing regularization techniques. These measures aim to enhance both the accuracy and robustness of short-term passenger flow prediction models. By addressing these areas, future studies can build upon the discoveries made in this research to develop more reliable and precise models for predicting short-term passenger flow. Ultimately, this will lead to improved efficiency and service quality within urban rail transit systems.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Tang L, et al. Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro, IEEE Transactions on Intelligent Transportation Systems, 2018.

[2] Zhang Kaiwen, He Yong. Passenger volume prediction of urban orbital transportation stations based on EMD-SSA-LSTM model. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 2023: 1-9.

[3] Zhou G, Tang J. Forecast of urban rail transit passenger flow in holidays based on support vector machine model 2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT). IEEE, 2020: 585-589.

[4] Zhang B, Li S, Huang L, Yang Y. An Improved Feedback Wavelet Neural Network for Short-Term Passenger Entrance Flow Prediction in Shanghai Subway System, Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, 2017.

[5] Zhang J, Chen F, Shen Q. Cluster-Based LSTM Network for Short-Term Passenger Flow Forecasting in Urban Rail Transit, IEEE Access 7, 2019.

[6] Wang Q. Research on Smart Tourism Based on Big Data Analysis. (Doctoral dissertation). Changchun University of Technology, 2021.

[7] Feng B. Research on Short-Term Passenger Flow Prediction for Urban Rail Transit Based on CNN-LSTM Combination Model. East China Jiaotong University, 2020.

[8] Graves A, Graves A. Long short-term memory. Supervised sequence labelling with recurrent neural networks, Supervised sequence labelling with recurrent neural networks, 2012.

[9] Zhao Q, Feng X, Zhang L, Wang Y. Research on Short-Term Passenger Flow Prediction of LSTM Rail Transit Based on Wavelet Denoising. Mathematics, 2023, 11(19).

[10] Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward, PloS one, 2018.