The Applications and Prospects of Large Language Models in Traffic Flow Prediction

Yuxuan Liu

School of Business, Hong Kong Baptist University, Hong Kong, 999077, China

Abstract. Predicting traffic flow is crucial for the functionality of intelligent transportation systems. It is of critical importance to relieve traffic pressure, reduce accident rates, and alleviate environmental pollution. It is an important part of the construction of modern intelligent road networks. With advancements in deep learning (DL), DL models have made notable strides in prediction. However, due to the complexity and non-transparency of DL models themselves, there are still problems of low accuracy and interpretability in traffic flow prediction (TFP). Leveraging large language models (LLM) helps to improve the negative conditions caused by other DL models in prediction. This paper first briefly summarizes the basic characteristics of LLM and their advantages in TFP; then conducts relevant research and analysis in the order of experimental design steps comparison and results and conclusions comparison; then analyzes and discusses the current problems and challenges faced by LLM; finally, it looks forward to future research directions and development trends, and summarizes this paper.

1 Introduction

Traffic congestion is one of the main causes of "urban disease" faced by many fast-growing cities. Therefore, traffic flow prediction (TFP) has a vital role in alleviating urban traffic congestion and providing a basic basis for the planning, design, construction, and operation of intelligent transportation systems in smart cities. The aspects it can affect include but are not limited to resource allocation and manpower planning, emergency incident handling, public travel choices, etc. There are two methods for TFP. The first is traditional regression analysis prediction. Generally, in this case, limited data will be used, and the model will be fitted after data processing and parameter tuning. For example, Kumar and Vanajakshi used the seasonal Autoregressive Integrated Moving Average (ARIMA) Model for short-term TFP and used limited input data to improve the applicability of the ARIMA model [1]. However, traditional regression analysis has considerable limitations for extremely complex traffic flow problems. Kumar and Vanajakshi mentioned in the article that the ARIMA model uses a large historical database for model development. For example, Stathopoulos and Karlaftis used approximately 60,000 traffic flow observations at 3-minute intervals, covering 106 days [2]. Williams and Hoel used more than 2 months of traffic volume observations [3].

Corresponding author: 22439021@life.hkbu.edu.hk

Model building based on a database may limit its application in places where data availability may be an issue. Storage and maintenance of historical databases can be a daunting task. The second type is neural network model prediction based on DL. Convolutional neural networks (CNN)are generally used for traffic prediction. For example, Zhao used a method combining CNN and LSTM to build a TFP model [4]. Researchers use CNN to extract spatial features from trajectory traffic flow and use the memory properties of LSTM to extract temporal features of trajectory traffic flow. Finally, the extracted information is fused through the fully connected layer, and the prediction result is obtained. Experimental results show that on the JT-T809-2011 data set used, the proposed CNN-LSTM model has better performance than other methods in root mean square error (RMSE), mean absolute percentage error (MAPE), and Spearman correlation coefficient. There is a significant improvement of 1%-2% in all aspects. However, DL neural networks still have some shortcomings. The most important thing is that they cannot achieve a better balance between prediction accuracy, data usage, and economic costs. CNN relies on a large amount of historical data to train the model. The quantity and quality of data directly affect the accuracy of the model. High-precision TFP usually requires at least several years of historical data as a training set. Missing data or outliers can also affect model performance. For example, in traffic monitoring on a certain highway section, if 1% of the daily data records are missing, the model may not be able to learn the complete periodic pattern. The acquisition and storage of data will undoubtedly increase the cost of training and prediction. Limited by the economic development level of each region, its applicability in different fields will be reduced. Focusing training on specific tasks can result in overfitting, which diminishes the model's generalization capability [5]. In recent years, LLM has thrived and shown significant advancement across various domains. For example, in software engineering development, LLM can help researchers generate more complete specifications faster and conduct testing [6]. Although LLM was originally designed for processing natural language, its characteristics give it the advantages shown in Table 1 in TFP.

This paper explores the application of LLM in TFP and its potential advantages. This study compares several existing models, reveals the advantages and challenges of LLM in practical applications, and looks forward to future research directions. This paper aims to provide a new methodological reference for TFP and promote the widespread application of LLM in the transportation field.

2 Basic features of the LLM and advantages in TFP

Different from other neural networks, the LLM is a natural language processing technology based on DL, which has excellent performance in language tasks, including text generation, answering questions, and language translation. This section will introduce in detail the basic features of the LLM and its advantages when applied to TFP.

2.1 Basic features of LLM

The main features of LLM include a self-attention mechanism, Transformer-based architecture, pre-training and fine-tuning, and large parameter scale. A key characteristic of LLM is their self-attention mechanism, which allows the model to selectively attend to various segments of the input sequence during processing. The self-attention mechanism allows the model to take into account the information of the entire input sequence, that is, the global picture, rather than just the most recent words when generating content. For the Transformer architecture, it is the basis of modern LLM. It abandons traditional CNN and LSTM and relies on the above-mentioned attention mechanism to process the input data sequence. This architecture can process all elements in the sequence in parallel, so the

training speed becomes faster. LLM is usually pre-trained on large-scale unlabeled text to learn the general laws of language. After pre-training, the model can be fine-tuned on specific tasks to adapt to the knowledge of specific fields. In terms of parameter scale, LLM are usually very large, with hundreds of billions of parameters, which enables them to capture extremely complex language structures and patterns in pre-training [7].

As shown in Fig. 1, it covers the basic characteristics of LLM.

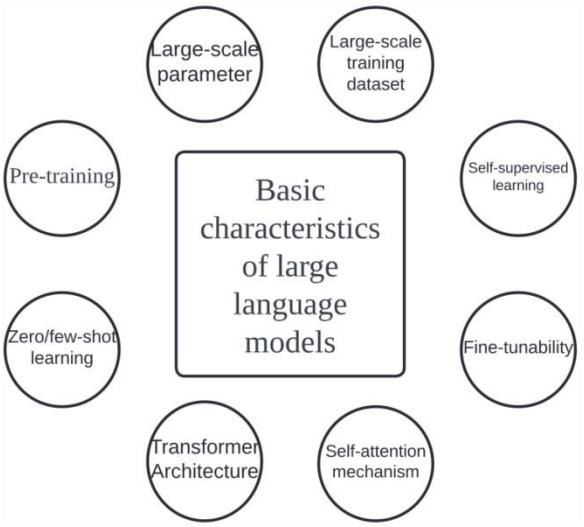

Fig. 1. Characteristics of LLM. The structure of the autoencoder (Photo/Picture credit: Original).

2.2 Advantages of LLM in TFP

The model can capture the trend and periodic characteristics of traffic flow over time. Fusion of multiple information TFP often needs to consider the influence of multiple factors, such as weather, holidays, special events, etc., and usually this information appears in the form of natural language. LLM can integrate this information to improve the accuracy of predictions. Contextual understanding. LLM is good at processing context-related information. In TFP, this ability can help the model better understand the changes in traffic patterns, such as traffic congestion in a specific period. Long-distance dependency in TFP, the traffic conditions between different locations may have long-distance mutual influence. LLM can capture this long-distance dependency, which is crucial for predicting changes in traffic flow.

Scalability and generalization ability. LLM have strong generalization ability after large-scale data training, which means that even in the face of new scenarios or unseen data, the model can make relatively accurate predictions. Practical application cases in practical applications, LLM can be combined with other types of machine learning models to form hybrid models to further improve prediction performance. For example, a CNN can be combined to process image data (such as satellite images, surveillance camera videos, etc.), or a Recurrent Neural Network (RNN) can be combined to enhance the processing capabilities of time series data. Table 1 covers the basic advantages of LLM.

Table 1. The basic advantages of LLM.

Explainability	Transforming multimodal traffic data into natural language descriptions is more intuitive and explainable
Multimodal Input	Comprehensively capture multimodal factors: time, space, accidents, etc.

Avoid complex data programming	Fine-tune the basic model to simplify the prediction process
Input Dependency Explanation	Generate input dependency explanations to help decision makers better understand and make decisions
Zero-shot generalization capability	Low sample or even zero sample prediction, strong generalization ability

3 Analysis and comparison of research status

3.1 Application and performance evaluation of the R2T-LLM model

Guo and other researchers built the R2T-LLM model based on the open source LLMLama2-7B-chat [8]. In this process, the researchers used LoRA technology to efficiently fine-tune the parameters. The fine-tuning parameters include learning rate, warm-up step size, gradient accumulation step size, etc. Dataset: The researchers used a subset of the LargeST dataset to construct the multimodal traffic prediction dataset CATraffic. The CATraffic dataset contains traffic volume data, weather information, point of interest (PoI) data, and holiday information for various regions in California, with a focus on Greater Los Angeles (GLA) and the Greater Bay Area (GBA). The data from 1/1/2018 to 12/30/2019 are sampled every hour. Sensor selection is clustered into 1,000 categories based on nearby PoI features, with only one sensor retained in each category. PoI data is acquired using OpenStreetMap and the Overpass Turbo API. "Meteorological data is gathered from the National Oceanic and Atmospheric Administration (NOAA). Experimental setup: The data was divided into a training set (2018 data) and a validation set (2019 data). The experiment was set up to forecast the next 12 hours of traffic flow based on the previous 12 hours of historical data. Samples that were zero for 24 consecutive hours were filtered out, which may be due to sensor failure. A zero-shot dataset comprising data from 100 sensors in San Diego (SD) was assembled to assess the model's generalization capability. The researchers used the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to evaluate the accuracy of the prediction results. Among them, RMSE emphasizes larger errors, MAE treats all errors equally, and MAPE provides interpretability in terms of relative accuracy. Baseline model: Extensive comparisons were made with 9 advanced baseline models, including LSTM, DCRNN, STGCN, ASTGCN, etc. The baseline models cover the popular research directions of time series prediction between 2018 and 2020. Responsible output generation: Initial attempts to have the model directly output text explaining the prediction were not successful. The solution is to adopt few-shot learning techniques to provide several explanation examples in a conversational setting. Few-shot explanatory texts are generated through ChatGPT by providing ground truth traffic flow sequences. Conclusion: The R2T-LLM model consistently outperforms the baseline models at multiple prediction steps, demonstrating substantial benefits in both short and long-term TFP. The R2T-LLM model performs particularly well in MAE and MAPE, which are 18.37% and 34.00% higher than the best baseline model, respectively. The R2T-LLM model is effective in capturing complex temporal patterns and can make more accurate and reliable predictions. Although it faces challenges in capturing subtle traffic fluctuations, R2T-LLM provides textual evidence for its predictions, indicating its high sense of responsibility and reliability. Future work will include in-depth research to enable LLM to use spatial information more effectively and consider more external factors to achieve more accurate predictions. The R2T-LLM model helps advance effective and responsible traffic prediction methods, which are crucial for making well-informed decisions in urban transportation planning and management.

3.2 Application and performance evaluation of TPLLM model

Ren et al. developed a sequence embedding layer using a CNN and a graph embedding layer utilizing a graph convolutional network (GCN) [5]. The sequence embedding layer extracts time series features, and the graph embedding layer extracts spatial features, and the LoRA method is used for fine-tuning to promote efficient learning and reduce computational requirements, thereby constructing a TPLLM model. GPT-2 served as the base LLM with an embedding dimension of D=768. Through the application of LoRA, the number of trainable parameters was reduced to approximately 0.95% of the total, significantly lowering the computational requirements. The optimizer of TPLLM is set to Adam, with an initial learning rate of 0.001 and a learning rate decay rate of 0.5/100 epochs. The batch size is set to 16 or 32, and the LoRA level is set to 4/8/16/32/48/64. The loss function selects the MAE to deal with the outlier problem in the dataset. Dataset: The researchers used two real-world datasets, PeMS04 and PeMS08, for experiments. The PeMS dataset contains traffic data collected by multiple sensors on major roads in California at a frequency of 5 minutes. All data are normalized to improve the stability of the training process. Experimental setting: The input sequence length T is 12, and the traffic flow in the next 15 minutes, 30 minutes, and 1 hour is predicted. The dataset is divided into training sets, validation sets, and test sets in chronological order. The training set accounts for 60%, and the validation set and test set account for 20% each. To evaluate the small sample learning ability of TPLLM, experiments with only 10% training data are also conducted. The researchers used MAE, RMSE, and MAPE to evaluate the accuracy of the prediction results. Baseline model: Extensive comparisons are made with multiple advanced baseline models such as LSTM, STGCN, and ASTGCN. Conclusion: The TPLLM model outperforms other baseline models in both full-sample and small-sample predictions, demonstrating the effectiveness of its embedding layers and pre-trained LLM in capturing spatiotemporal dependencies in traffic data. Notably, in small-sample predictions (using only 10% of the training set), the TPLLM model maintains high accuracy with minimal performance degradation, indicating its strong generalization capabilities suitable for real-world scenarios. The LoRA adaptation technique has a negligible impact on prediction accuracy, allowing for the selection of a smaller LoRA level to balance accuracy with computational efficiency. The ablation study confirms the positive contributions of the sequence embedding layer, graph embedding layer, and LoRA to the overall predictive accuracy. TPLLM represents a novel approach to traffic prediction, effectively leveraging pre-trained LLM to provide accurate forecasts of graph-structured traffic data.

Future research plans include considering more factors affecting traffic flow to design embeddings to improve prediction accuracy, further exploring PEFT techniques that are more suitable for spatiotemporal prediction tasks, and trying to find an explainable LLM knowledge learning model. The experimental results also show that TPLLM can provide accurate prediction results even in the case of limited data, which is particularly useful for areas where historical traffic data is scarce.

3.3 Application and performance evaluation of LLM-embed TFP model

Huang's research proposes a new method that does not use LLM directly for prediction but uses LLM to handle textual information and derive embeddings [9]. These embeddings are then merged with historical traffic data and fed into established spatiotemporal forecasting models. Application scenarios include regional-level scenarios and node-level scenarios. In regional-level scenarios, text information is represented as nodes connected to the entire network; while in node-level scenarios, embeddings from LLM represent additional nodes that are only connected to the corresponding nodes. Dataset: The New York City bicycle-

sharing dataset is used, which records the bicycle pick-up and drop-off volume of each grid every hour. This dataset is widely used in traffic forecasting research. Each record in the dataset includes the time, start point, and end point of a trip. After division, the dataset becomes 169 grids, each with a span of 1 km × 1 km. City-level factors such as air quality, rainfall, temperature, forecast time, day of the week, and holiday information are included. The dataset is divided into a training set (6/1/2023 to 8/7/2023), a validation set (8/8/2023 to 8/22/2023), and a test set (8/23/2023 to 9/20/2023). Experimental setup: The researchers used the traditional version and the enhanced version (i.e., LLM embedding) for prediction. Two tasks are focused on: one is assessing whether the prediction improves accuracy in the whole grids; the other is assessing whether the TFP accuracy of a specified grid will improve after providing information related to a specific grid. Baseline model: Eight basic spatiotemporal prediction models are selected for experiments: AGCRN, DCRNN, STGCN, MTGNN, GWNET, TGCN, STTN, and STSGCN. Conclusion: It is found that after adding LLM embedding, the prediction accuracy for all grids is improved. Special attention is paid to Grid 84, which is close to the Barclays Center in New York City and often hosts events such as concerts and sports games. Errors in all areas are reported, especially those in area 84. After the embedding of LLM, the performance of the model is improved, especially in the prediction of all grids and specific grids 84. LLM performs well in handling special situations (such as holidays and special events), which can effectively improve prediction accuracy. LLM can effectively integrate non-numerical background information (such as weather forecasts) to improve the accuracy of predictions. Future research can further explore how to better utilize the knowledge of LLM to improve the accuracy and robustness of traffic predictions.

4 Problems and challenges

LLM has many advantages in the field of TFP, but it also has some objective defects. LLM still relies on DL models, which makes it difficult to explicitly understand its internal reasoning chain and process of decision-making. It only can be analyzed through input and output. In addition, LLM is prone to overfitting during training, especially when the label noise is large. These cannot be ignored. As shown in the research and analysis comparison in Section 3 of this article, the use of fine-tuning techniques (such as Lora) to adjust the model on the original basis to adapt to different situations has become the main feature of using LLM for TFP research. Therefore, what technology to use for adjustment, how accurate the adjusted model is, and whether it has better generalization performance have become important challenges in current research.

5 Future development direction

LLM still has a lot of room for improvement in the field of TFP. In terms of multimodal data fusion, LLM can further explore how to better integrate multimodal traffic data, such as traffic sensor data, weather information, various traffic flow data, etc., and improve the accuracy of prediction by integrating different data sources. In terms of spatiotemporal relationship, LLM can consider the spatiotemporal relationship of data in more detail. For example, the interaction between different time scales (hours, days, weeks) and spatial scales (different regions, road sections). In terms of the introduction of external factors, LLM could further explore the influence of external factors in TFP, such as special events, road construction, etc. In this way, traffic flow information can be predicted more accurately. In the field of transportation, it is not limited to ground transportation but plays a greater role in

the prediction of air traffic. For example, Ma et al. have built a TEMPT model based on a Transformer to identify and predict cross-flow in the air [10].

6 Conclusion

Since the LLM represented by ChatGPT was proposed and achieved results in many fields, many studies have brought it into the field of transportation after recognizing its generalization ability. For example, LLM is used to predict taxi usage during special events, and LLM is used to analyze text descriptions of traffic accidents to determine the cause of the accident and the responsible party. In addition, some articles suggest using LLM to analyze users' potential transportation choices based on their attributes and the costs and utilities of various modes of transportation. In summary, the LLM has not only made certain research progress in TFP but has also been greatly applied to other fields of transportation. In the future, in the mutual learning, mutual benefit, and mutual inspiration between different fields, many problems existing in the LLM itself will also be optimized and solved by new methods and new research that are constantly proposed, and achieve a complete transformation.

References

1. S. V. Kumar, L. Vanajakshi, Short-term TFP using seasonal, Eur. Transp. Res. Rev. **7**(3), (2015)
2. A. Stathopoulos, M. G. Karlaftis, A multivariate state space approach for urban traffic flow modeling and prediction, Transp. Res. **11**, 121-135 (2003).
3. B. M. Williams, L. A. Hoel, Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results, J. Transp. Eng. 129(6), 664 – 672 (2003)
4. Z. Zhao, Z. Li, F. Li, Y. Liu, *CNN-LSTM Based Traffic Prediction Using Spatial-temporal Features*, in Proceedings of Journal of Physics: Conference Series, J. Phys.: Conf. Ser. 2037, 012065 (2021)
5. Y. Ren, Y. Chen, S. Liu, B. Wang, H. Yu, Z. Cui, TPLLM: A Traffic Prediction Framework Based on Pretrained Large Language Models, arXiv preprint arXiv:2403.02221 (2024)
6. Alshahwan N., Harman M., Harper I., Marginean A., Sengupta S., & Wang E. Assured LLM-Based Software Engineering. In Proceedings of the InteNSE 24: ACM International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering Lisbon, Portugal. arXiv:2402.04380 (2024)
7. S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-Attention with Linear Complexity, arXiv preprint arXiv:2004.05150 (2020)
8. X. Guo, Q. Zhang, J. Jiang, M. Peng, H. (F.) Yang, M. Zhu, Towards Responsible and Reliable TFP with Large Language Models, arXiv preprint arXiv:2404.02937v4 (2024)
9. X. Huang, Enhancing Traffic Prediction with Textual Data Using Large Language Models, arXiv preprint arXiv:2405.06719v1 (2024)
10. C. Ma, S. Alam, Q. Cai, D. Delahaye, Text-Enriched Air Traffic Flow Modeling and Prediction Using Transformers, IEEE Trans. Intell. Transp. Syst. **25**(7), 7963-7976 (2024)