Dean&Francis

ISSN 2959-6157

Research Progress and Future Prospects of Large Language Models in the Field of Traffic Flow Prediction

Yang Zhang

Department of Electronic Engineering Fuzhou Institute of Technology, Fuzhou, China
YangZhang051102@outlook.com

Abstract:

With the enhancement of computer performance, the predictive capabilities of large language models have also improved, and these models have been applied in many fields. Currently, the application of large language models in the field of traffic flow prediction has entered a rapid development stage. Thanks to its powerful feature learning and generalization capabilities, it has provided new ideas and methods for traffic flow prediction, and a series of achievements have been achieved. This article reviews the research progress of mainstream large language models in the field of traffic flow prediction. This introduces the background and significance of applying large language models to traffic flow prediction, as well as the current mainstream large language models, and elaborates on the relevant model architectures and methods in detail. Analyzed its advantages in enhancing predictive accuracy and interpretability, as well as the challenges it faces. At the same time, an outlook on the future development directions of large language models in this field has been provided, aiming to offer a reference for further research and application of traffic flow prediction in intelligent transportation systems.

Keywords: Large language models; Traffic Forecasting; Data Transformation; Model Scale

1. Introduction

With the current surge in the development of new energy technologies, the significant increase in the number of vehicles has led to increasingly severe traffic issues. The traditional traffic planning and regulation model is showing signs of fatigue in response to such situations [1]. With the advancement of artificial intelligence technology. Many fields leverage large language models for fine-tuning to achieve predictions of spatiotemporal data [2], and the transportation sector is particularly representative. Large language models achieve data transformation through their own capabilities in natural language processing. Transform abstract and complex traffic flow data into structured natural language, and capture the spatiotemporal traffic flow data for subsequent traffic flow prediction. The aim is to visualize the processing of spatiotemporal sequence tasks, enabling users to conduct more rational and efficient planning through the interpretability of large language model processing methods. With the advancement of research, the integration of multiple disciplines to achieve autonomous intelligent driving of vehicles will reduce accident rates while enhancing travel efficiency [3]. Certainly, many requirements have been raised for large models, such as key aspects including input features, model types, datasets, and evaluation metrics. It emphasizes the trade-offs between practicality and performance, as well as the challenges of model interpretability and scalability in different applications [4]. The following text will summarize the research advancements and future prospects of large language models in the field of traffic flow prediction. This aims to introduce readers to the main research directions and progress of current large language models in the field of traffic flow prediction, as well as to anticipate future research directions and provide relevant research suggestions.

2. Research and Development Progress and Key Issues of Large Language Models in the Field of Traffic Flow Prediction.

2.1 Research and Development Progress

Currently, significant progress has been made in the research of large language models in the field of traffic flow prediction. Taking TP-LLM as an example. Compared to previous large models, it has certain advantages. As shown in Table 1 [5], In the past two years, many teams have successively published articles introducing new methods or improving existing ones. For example, STGCN is a model specifically designed for spatiotemporal data, which integrates Graph Convolutional Networks (GCN) with Temporal Convolutional Networks (TCN). Its core principle lies in modeling the spatial correlations of traffic road networks using graph structures (if the segments are regarded as nodes and the connectivity relationships as edges), At the same time, capture the temporal dynamic characteristics of traffic flow through temporal convolution [6]. In traffic flow prediction, it can accurately uncover the 'impact of traffic flow in adjacent segments' and 'temporal patterns', However, the adaptability to unstructured data is relatively low, and it is necessary to optimize the graph structure design when generalizing to complex road networks (such as multiple intersections); LSTM, an improved type of recurrent neural network (RNN) that utilizes gating mechanisms at its core (Input gate, forget gate, output gate), addresses the issues of gradient vanishing/explosion in traditional RNNs. Can effectively capture long-distance dependencies in temporal data. In traffic flow prediction, it can handle continuous time series data such as traffic volume and speed, learning traffic patterns at different time steps (As the peak occurs periodically). However, the modeling ability of spatial correlation is weak, and it is difficult to directly capture the conduction effect of traffic flow in adjacent road sections, which is suitable for single-sequence prediction scenarios based on time dimension [7]. The currently more advanced model is ST-LLM, to address the challenges of insufficient spatio-temporal modeling, weak cross-domain generalization, and high computational costs in traffic forecasting, significant upgrades have been achieved through the use of graph-enhanced spatio-temporal LLM and efficient fine-tuning techniques such as LoRA, resulting in explicit spatial modeling, parameter efficiency, and robust performance across domains. Experiments conducted on real traffic datasets indicate that ST-LLM outperforms most current advanced models, demonstrating robust performance in both few-shot and zero-shot prediction scenarios [8].

Table 1. Comparison of the performance of three relevant models under different time lengths and datasets.

Dataset	Model	Metrics	30 min (T'=6)	60 min (T'=12)	Average
PeMS04	LSTM	MAE	27.25	32.28	29.77
		RMSE	43.66	49.57	46.62
		MAPE(%)	19.26	22.45	20.86
	STGCN	MAE	25.31	30.29	27.8
		RMSE	40.39	46.7	43.55
		MAPE(%)	17.14	20.74	18.94
	TP-LLM	MAE	19.43	21.49	20.46
		RMSE	31.76	34.81	33.29
		MAPE(%)	12.67	14.2	13.44

ISSN 2959-6157

PeMS08	LSTM	MAE	22.6	26.94	24.77
		RMSE	36.16	41.05	38.61
		MAPE(%)	17.18	20.07	18.63
	STGCN	MAE	21.24	24.78	23.01
		RMSE	32.28	37.1	34.69
		MAPE(%)	13.72	15.86	14.79
	TPLLM	MAE	15.43	17.22	16.33
		RMSE	25.34	28.43	26.89
		MAPE(%)	9.83	11.05	10.44

2.2 Main Issues

Currently, large language models exhibit certain potential in the field of traffic flow prediction; however, they still possess certain limitations. There are still numerous technical challenges and scene adaptation issues to be addressed before the actual implementation can be realized [9].

2.2.1 Noise Interference

The issue of noise is currently one of the core problems affecting the performance of large language models. The presence of noise can easily lead to a decrease in the predictive accuracy of large models, causing them to learn false associations and affecting the reliability of actual traffic decision-making. This mainly manifests in three aspects: data input, model training, and prediction output. Some discrepancies arise from the information collected by sensors, some are due to overfitting of the attention mechanism, and there may also be conflicts with pre-trained knowledge.

2.2.2 The Contradiction Between Predictive Accuracy and Interpretability.

The prediction of traffic flow by large models requires the reading and processing of massive amounts of traffic data to produce results. More accurate predictions necessitate the output of embedding layers and the capability of processing multidimensional data in pre-trained large language models. The increase in dimensions leads to a more complex model, and thus, the interpretability of the model inevitably decreases as the complexity of the model increases due to improved prediction accuracy.

2.2.3 Barriers to Practical Implementation.

In order to involve large language models in traffic prediction, it is necessary to convert the traffic data transmitted through the various traditional model interfaces of the traffic system into a structured language recognized by large language models, and to process the output into results that users can directly understand. During the process of converting traffic data into structured language, there is a risk of data loss, which may even include the loss of crucial information, leading to prediction bias and consequently resulting in traffic decision-making issues.

3. Solutions to Major Issues

3.1 Solutions to Noise Problems

From the perspective of data input: The collection of traffic data by sensors may be affected by sensor device aging, extreme weather conditions (such as heavy fog and rain), resulting in data missing or anomalous values. Multi-layered sensors can be utilized for monitoring, employing various types of monitoring equipment such as exhaust concentration sensors and cameras. This will be combined with weather records from meteorological departments and traffic accident records from traffic management authorities to perform multi-layered data comparison. Ultimately, this will facilitate the integration of data for input;From the perspective of model training: Utilize a small amount of annotated clean traffic flow data alongside a large volume of unannotated noisy data for semi-supervised learning. For example, by utilizing the pseudo-labeling method, we first train the model on a small amount of labeled data, and then use the trained model to generate pseudo-labels for a large volume of unlabeled data. Afterwards, these pseudo-labeled data are added to the training set for joint training, allowing the model to learn more generalized features and reduce noise interference, thereby emphasizing the need to customize specific noise tailored to the transportation domain [10]; From the perspective of predictive output: Build multiple prediction sub-models based on large language models, train multiple large language models in parallel, with each model trained on a different subset of training data, and finally average or vote on the prediction results of these models to reduce prediction bias caused by noise in a single model [11, 12].

3.2 Methods for Resolving the Contradiction between Prediction Accuracy and Interpretability

Improving the attention mechanism, we design an attention weight allocation rule based on knowledge in the transportation sector. In calculating attention weights, more reasonable weights are artificially assigned based on factors such as the topological relationships and distances between road segments. This allows the distribution of attention weights to more intuitively reflect the propagation paths and influence ranges of traffic flow, facilitating the understanding of the model‘s decision-making basis and enhancing prediction accuracy. At the same time, during the model training, some practical traffic scenarios that may occur, such as reduced speed in rainy or snowy weather and surge in traffic during early morning and evening rush hours, are artificially set. This, combined with real-time data, validates the triggering conditions of these rules, making the explanation more aligned with actual logic and enhancing interpretability.

3.3 Practical Application and Implementation Solutions.

By defining clear semantic rules and mapping relationships, data features are matched one-to-one with natural language descriptions, avoiding ambiguity. For quantitative data such as flow, speed, and congestion levels, establish hierarchical semantic labels and clarify the correspondence between numerical ranges and language descriptions. For example, stipulate that different speeds correspond to different congestion situations. Set up a structured semantic template to clarify the correspondence between event attributes, scope of influence, and degree, for example: „At xx location, on xx date, xx event occurred, resulting in xx degree of congestion in the xx area.“ By constraining with the template, unstructured event data can be transformed into logically clear language descriptions, avoiding information loss or misinterpretation. In practical applications, incorporate a human feedback learning mechanism to assist in optimizing the model.

4. Future Perspectives of Large Language Models in the Field of Traffic Flow Prediction

4.1 Model Structure Optimization

Continuously promote innovations in model architecture, deepen the integration mechanism between graph neural networks and large language models, optimize the allocation of attention weights for spatial and temporal features, and enhance the model‘s ability to capture the topological relationships of road networks and the laws of spatiotemporal interactions. Further explore the design of lightweight models to reduce computational costs while ensuring prediction accuracy, thereby meeting the demands of real-time traffic flow prediction.

4.2 Upgrade of Noise Reduction Technology

Develop more intelligent noise detection and processing methods by integrating knowledge from the transportation field and multimodal data fusion technology, in order to enhance the ability to differentiate between noise and genuine abnormal events. Strengthen the research on the cumulative effects of noise under long-term sequence input, design targeted denoising algorithms, and enhance the robustness of models in complex noise environments.

4.3 Enhanced Interpretability

Research the balancing mechanism between interpretability and predictive accuracy, and develop more effective attention mechanism visualization methods that correspond to the physical logic of traffic scenarios. By integrating knowledge graphs and domain rules, a more comprehensive model explanation framework can be constructed, making the reasoning process of prediction results more transparent and enhancing user trust in the model.

4.4 Data Conversion and Integration Optimization

Improve the semantic mapping rules between transportation data and structured language, establish a more comprehensive semantic template library for the transportation domain, and reduce information loss during the data conversion process. Strengthen the research on the interface adaptation with existing transportation systems, promote the seamless integration of model output formats with the requirements of traffic management systems, and reduce the integration difficulties of practical applications.

5. Conclusion

The aforementioned work focuses on the application of large language models in the field of traffic flow prediction. Firstly, this study clarifies the research objective of outlining its value and development in order to serve intelligent transportation research. Subsequently, it combines the emerging trend of the surge in new energy vehicles, Point out the limitations of traditional transportation planning and elucidate the advantages of data transformation in large language models and their ability to capture spatial and temporal features. Subsequently, we will sum-

ISSN 2959-6157

marize the research progress and introduce the principles and characteristics of mainstream models such as LSTM, STGCN, TP-LLM, and ST-LLM. We will quantify the performance differences of these models using metrics such as MAE, relying on the PeMS04 and PeMS08 datasets. Analyze the three key issues and their causes: the impact of noise, the contradiction between accuracy and interpretability, and the obstacles to implementation. Propose targeted solutions at the data layer, model layer, and application layer, and finally outline future research directions from four perspectives. With the development of new energy technologies, the rapid increase in the number of vehicles has exacerbated traffic issues, leading to the gradual ineffectiveness of traditional traffic planning models. Leveraging natural language processing capabilities, large language models transform complex traffic flow data into structured language, capturing spatial and temporal features to enable predictions, thus providing new insights for intelligent traffic planning.

At present, research in this field has made certain progress. The TP-LLM model demonstrates advantages compared to traditional models, while the advanced ST-LLM achieves explicit spatial modeling, parameter efficiency, and robust cross-domain upgrades through graph-enhanced spatiotemporal LLM and LoRA fine-tuning techniques. It outperforms most existing models in real datasets as well as in few-shot and zero-shot scenarios.

However, research still faces key issues: noise affects the accuracy of predictions and the reliability of decisions at the levels of data input, training, and output; improvements in prediction accuracy come with increased model complexity, leading to a decrease in interpretability; information loss can easily occur during the data transformation process, hindering practical application implementation.

In response to these issues, the study proposes a multi-level data integration approach, semi-supervised learning, and ensemble prediction to address noise problems; an improved attention mechanism is employed to balance accuracy and interpretability; and data transformation is optimized through semantic rule mapping and human feedback. In the future, efforts will focus on optimizing model structures, upgrading noise handling, enhancing interpretability, and optimizing data integration, to promote the practical development of large language models in the field of traffic flow prediction.

References

[1] Shanmugam A et al. Integrated Traffic Management System: A review. *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing* 2024 Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology. 2011.

[2] Chang Y P et al. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology 15.3 2024: 1-45.

[3] Miglani A and Neeraj K. Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. Vehicular Communications 20 2019: 100184.

[4] Wu Y Y et al. A survey on origin-destination flow prediction. *2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2024.

[5] Ren Y L et al. TP-LLM: A traffic prediction framework based on pretrained large language models. arXiv preprint arXiv:2403.02221. 2024 .

[6] Gao N et al. Generative adversarial networks for spatio-temporal data: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13.2 2022: 1-25.

[7] Yu Y et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation 31.7 2019: 1235-1270.

[8] Liu C X et al. ST-LLM+: Graph Enhanced Spatio-Temporal Large Language Models for Traffic Prediction. *IEEE Transactions on Knowledge and Data Engineering.* 2025 .

[9] Zhao Y & Zhou Y. B. Research on the application of reinforcement learning in traffic flow prediction. *Information and Computers (Theoretical Edition)*, 36(03) 136-138. 2024.

[10] Feriance M et al. Navigating noise: A study of how noise influences generalisation and calibration of neural networks. *arXiv preprint arXiv:2306.17630*. 2023.

[11] Gloeckle F et al. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737.* 2024.

[12] Yuan X L, Xin L and Jian Z. The development of new energy vehicles for a sustainable future: A review. Renewable and Sustainable Energy Reviews 42 2015: 298-305.