Transactions on Computational and Scientific Methods | Vo. 5, No. 8, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Attention-Enhanced Network Traffic Prediction for Intelligent Load Balancing Optimization

Ziyu Qing[1], Fei Huang[2]
[1]University of Southern California, Los Angeles, USA
[2]University at Buffalo, Buffalo, USA
*Corresponding Author:Ziyu Qing; qingziyu1024@gmail.com

Abstract: This paper focuses on the problem of load balancing optimization in complex network environments and proposes a time series-based network traffic prediction method to achieve more efficient resource scheduling and traffic management. The proposed model integrates gated recurrent units (GRU) with a multi-head self-attention mechanism to capture both short-term local fluctuations and long-term dependencies. In data preprocessing, the model applies a sliding window and normalization strategy to model multi-node traffic sequences, enhancing adaptability in dynamic scenarios. During training, a weighted mean squared error is used as the loss function to guide the model in accurately identifying high-variance regions, thereby improving overall prediction quality. To validate the effectiveness of the method, experiments are conducted on the MAWI network traffic dataset, including a sensitivity analysis on time window length and sampling frequency. Experimental results show that the proposed method outperforms mainstream models across multiple error metrics and demonstrates strong robustness under different parameter settings and data conditions. This work provides accurate and stable predictive support for intelligent load balancing and shows strong potential for deployment in real-world network scenarios.

Keywords: Traffic modeling; sequence prediction; attention mechanism; load scheduling

1. Introduction

In today's highly interconnected digital society, network infrastructure has become a key resource supporting various information services and online applications. With the rapid development of large-scale cloud computing, the Internet of Things, 5G communication, and edge computing, network environments are becoming increasingly complex and dynamic. Network traffic now exhibits features such as high concurrency, strong fluctuation, and time variance[1]. At the same time, user expectations for service quality continue to rise. Performance indicators such as low latency, high bandwidth, and service continuity have become critical metrics for evaluating network systems. Against this background, efficiently scheduling and intelligently allocating network resources has become a core challenge for ensuring the stability of upper-layer applications. As a key mechanism to ensure network performance and improve system throughput and resource utilization, load balancing increasingly relies on accurate perception and forecasting of dynamic network traffic[2].

Traditional load balancing methods often rely on static rules or heuristic strategies based on historical averages. These methods tend to exhibit delayed responses, inaccurate predictions, and uneven resource allocation in highly dynamic and complex network topologies[3]. These limitations are especially evident in large-scale distributed systems, where some nodes become performance bottlenecks due to excessive load

while others remain underutilized. This leads to reduced service quality and overall system efficiency. Therefore, there is an urgent need for a more intelligent and proactive approach that can forecast traffic trends in advance and provide real-time guidance for load balancing strategies. This enables a more adaptive and elastic scheduling mechanism. Under this context, network traffic prediction has gained increasing attention as a critical precursor to intelligent load balancing, with significant research value and application potential[4].

Network traffic prediction is essentially a complex time series modeling problem influenced by multiple factors, including user behavior patterns, service request types, temporal periodicity, and geographic distribution. An effective prediction model must capture short-term fluctuations and model long-term trends, sudden events, and cross-node correlations. Based on this, recent research has explored methods that integrate multi-source heterogeneous data with deep neural network architectures[5]. These approaches aim to enhance the model's capacity for nonlinear modeling, dynamic feature extraction, and multi-scale information fusion. In data-driven intelligent network management frameworks, traffic prediction has become a core module for system optimization. It plays a forward-looking role in load awareness, traffic scheduling, and congestion control.

In load balancing scenarios, traffic prediction not only provides information about future traffic patterns but also directly affects the efficiency of task distribution and resource allocation across nodes. Accurate prediction results allow the system to schedule proactively before peak traffic periods. This avoids congestion and packet loss caused by traffic surges and improves resource utilization during off-peak periods. Furthermore, for load balancing systems based on centralized or distributed control architectures, the responsiveness and deployment flexibility of prediction models are equally important. This requires prediction models to be accurate, lightweight, and transferable[6]. Building an efficient, generalizable, and adaptable traffic prediction model is of great theoretical and practical value for enhancing the intelligence and robustness of load balancing systems.

In summary, as network systems become more complex and the demand for service quality increases, traditional load balancing mechanisms face significant challenges. It is essential to introduce more intelligent sensing and decision-making mechanisms to meet future development needs. As a key enabling technology for optimizing load balancing, traffic prediction plays a fundamental and proactive role in network resource management[7]. By incorporating advanced time series modeling and deep learning architectures, it is possible to accurately capture and respond to traffic patterns in dynamic environments. This transition promotes the shift of load balancing from passive adjustment to proactive optimization, providing solid support for building efficient, intelligent, and adaptive next-generation network infrastructure.

Recent studies further show that predictive load balancing should be connected with adaptive backend scheduling, anomaly-aware operation, and multi-scale temporal modeling. Reinforcement learning-based elastic scaling for microservices [8], structure-guided attention and multi-scale behavior modeling for anomaly detection [9], and uncertainty-aware backend failure detection [10] indicate that load-aware network management can benefit from adaptive control and reliability-sensitive modeling. Learning-based intelligent agents for backend resource scheduling [11] further connect prediction outputs with operational decision making. Cross-domain temporal modeling studies, including enterprise audit risk identification with temporal context and entity relationships [12], counterfactual cash-flow forecasting and strategy simulation [13], memory-enhanced transformer anomaly detection [14], multi-scale attention for dynamic user-interest evolution [15], log semantic graph construction for system-event detection [16], lightweight compressed representation learning at the edge [17], and collaborative scheduling for multi-tenant distributed inference services [18], provide methodological support for combining multi-scale sequence representation, structural correlation modeling, and proactive load balancing decisions.

2. Proposed Approach

This study proposes a network traffic prediction method for load balancing optimization. By constructing a high-precision time series modeling framework, it aims to provide forward-looking traffic trend references for network nodes, thereby enabling intelligent resource allocation and scheduling. The model architecture is shown in Figure 1.

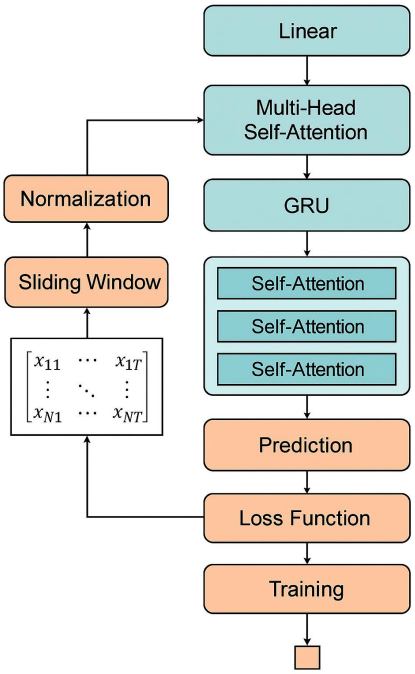

Figure 1. Overall Model Architecture

Specifically, we construct the historical traffic data of each node in the network as a multivariate time series matrix. Assuming the total number of nodes is N and the time step is T, the input features can be represented as a matrix $X \in R^{N \times T}$. During the modeling process, the input data is first normalized using the min-max normalization method:

$$x_{i,t} = \frac{x_{i,t} - \min(x_i)}{\max(x_i) - \min(x_i)}, \forall i \in \{1,\ldots, N\}, t \in \{1,\ldots, T\}$$

Where $x_{i,t}$ represents the original flow value of the i-th node at the t-th time point, and $\hat{x}_{i,t}$ is the normalized input.

In terms of model architecture design, we adopted a hybrid framework that combines a gated recurrent unit (GRU) with a multi-head self-attention mechanism to simultaneously capture local temporal dependencies and global traffic correlations. The GRU module is responsible for modeling short-term dependencies, and its state update rule is as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \otimes h_{t-1}))$$
$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t$$

Among them, z_t and r_t are the update gate and reset gate, respectively, h_t represents the hidden state at the current moment, and $\otimes$ is the element multiplication.

To enhance the model's ability to model long-term dependencies, we introduced a multi-head self-attention module based on positional encoding. This module establishes explicit connections between different periods in the traffic sequence by calculating the attention weights between each time step. Given a query Q, key K, and value V, the multi-head self-attention calculation process is:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Where d_k represents the dimension of each attention head. By concatenating and linearly mapping the results of multiple heads, an embedding representation containing rich temporal dependency information can be obtained.

During the model training phase, a sliding window method is used to generate supervised samples, and a weighted mean square error is introduced as the objective function to enhance the model's sensitivity to areas with large traffic changes. The loss function is defined as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot (y_i - y_i)^2$$

Here, y_i and y_i represent the actual value and predicted value, respectively, and C is the sample weight dynamically adjusted based on historical volatility. Ultimately, the model's predicted output serves as the prior input for the load balancing scheduling module, guiding the generation and updating of subsequent resource optimization strategies and ensuring efficient and stable system operation under complex network conditions.

3. Performance Evaluation

3.1 Dataset

The network traffic dataset used in this study is the MAWI Working Group Traffic Archive. This dataset is provided by the WIDE project in Japan and is collected from real Internet backbone environments. It is widely used in network modeling, traffic analysis, and anomaly detection. The dataset contains detailed packet-level information, including timestamps, source and destination IP addresses, protocol types, packet lengths, and transmission directions. It features high frequency, strong authenticity, and strong representativeness, making it suitable for capturing dynamic changes in complex network traffic.

The dataset is organized by date, allowing for comparisons across different periods and enabling trend modeling. This study selects traffic segments spanning several consecutive days and constructs time series inputs using fixed-interval sampling. Preprocessing steps include denoising, aggregation, and normalization to meet the continuity and stability requirements of time series prediction models. The diversity and scalability of the data make it an ideal foundation for traffic prediction in load balancing optimization tasks.

The openness and structured design of this dataset provide high practical value for research on network traffic. Since the data is captured from real operational networks, it reflects temporal variability and sudden changes under realistic conditions. This helps models learn complex and realistic traffic patterns during training. As a result, the prediction performance gains better generalization and robustness in real-world load balancing scenarios.

3.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

As shown in Table 1, the performance of different models on the network traffic prediction task varies significantly. Traditional architectures such as TransMUSE exhibit relatively high error metrics. This indicates that their modeling capacity is still insufficient for handling complex and time-varying network

traffic. In particular, the poor results on MSE and RMSE suggest a lack of accuracy in predicting large-scale traffic fluctuations. Such cumulative prediction errors can lead to delayed resource scheduling in load balancing strategies, which negatively impacts the overall system's response speed and stability.

Table1: Comparative experimental results

Model	MAE	MSE	RMSE	MAPE
TransMUSE[19]	4.27	36.15	6.01	12.83%
DGCRN[20]	3.91	30.42	5.52	11.47%
Trafficformer[21]	3.66	28.10	5.30	10.92%
Ours	3.12	21.85	4.67	9.35%

In comparison, models with stronger graph or temporal modeling capabilities, such as DGCRN and Trafficformer, achieve better prediction accuracy. Key metrics such as MAE and MAPE are reduced, showing that the introduction of structure-aware or attention-based mechanisms improves the model's ability to capture traffic trends. This enhancement supports more proactive load balancing. However, these models still struggle with capturing long-term dependencies and integrating multi-scale information. These limitations reduce their generalization performance in complex network environments.

The proposed method in this study achieves the best results across all metrics. Specifically, it lowers the MSE to 21.85 and the MAPE to 9.35 percent. These results demonstrate a clear advantage in accurately modeling traffic fluctuations. By integrating gated recurrent structures with multi-head attention mechanisms, the model captures both short-term and long-term dependencies. It also adapts to heterogeneous traffic interaction patterns across multiple nodes. This design improves both the stability and sensitivity of the predictions. For load balancing systems, such accurate forecasting enables more timely and precise resource pre-allocation and hotspot migration, which significantly reduces the risk of congestion and resource waste.

Overall, the experimental results confirm the robustness and adaptability of the proposed model in dynamic network scenarios. In the context of load balancing optimization, reduced prediction errors lead directly to more refined scheduling strategies and improved service performance. The model not only enhances the system's responsiveness to sudden traffic surges but also provides a practical algorithmic foundation for building more efficient and intelligent network management systems.

This paper also analyzes the impact of different time window lengths on prediction performance, and the experimental results are shown in Figure 2.

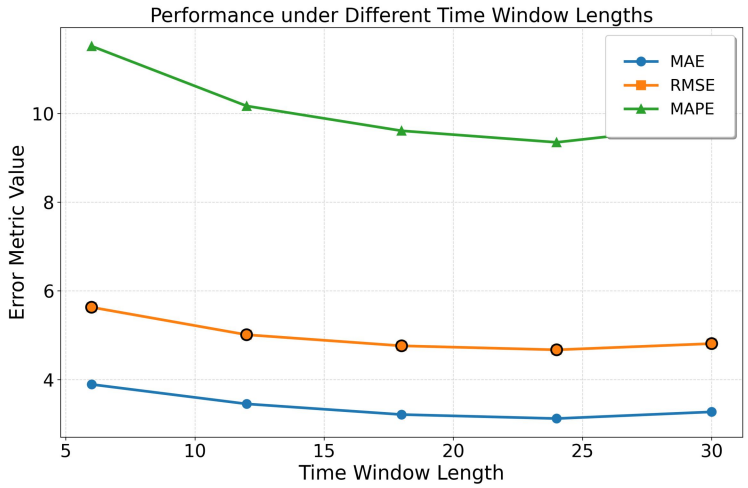

Figure 2. Analysis of the impact of different time window lengths on prediction performance

As shown in Figure 2, the length of the time window has a significant impact on the performance of the network traffic prediction model. The three evaluation metrics—MAE, RMSE, and MAPE—exhibit clear variation trends under different window settings. Overall, as the time window length increases, the error metrics generally show a downward trend. This suggests that longer windows provide richer historical context, improving the model's ability to capture temporal features in traffic.

Specifically, the MAE decreases steadily as the window grows from 6 to 24, reaching its optimal point before slightly increasing again. This indicates that excessively long time windows may introduce noise or redundant information, which can degrade prediction accuracy. In load balancing scenarios, selecting an appropriate window length is therefore essential to balance information richness and noise interference. Similarly, RMSE, which emphasizes large fluctuations, follows a trend similar to MAE. This further confirms the model's stability and robustness under medium-length windows.

The MAPE metric shows more noticeable fluctuations. It performs worst under short time windows, indicating the model's sensitivity to proportional error when contextual information is lacking. This high error may cause the load balancing system to respond slowly to traffic peaks, leading to node congestion or service disruptions. In contrast, with window lengths between 20 and 24, MAPE reaches its lowest values. This shows that the model can better capture local variations and trends and adapt to traffic dynamics across different time scales.

Overall, the experimental results highlight the critical role of time window length as a key hyperparameter. In load balancing optimization tasks, accurately determining the depth of historical information not only affects prediction accuracy but also directly impacts the system's ability to anticipate traffic trends. This, in turn, determines the efficiency of resource allocation and the robustness of system operations. Therefore, selecting a suitable time window is a fundamental requirement for achieving high-quality network traffic prediction.

This paper also gives the impact of data sampling frequency on prediction accuracy, and the experimental results are shown in Figure 3.

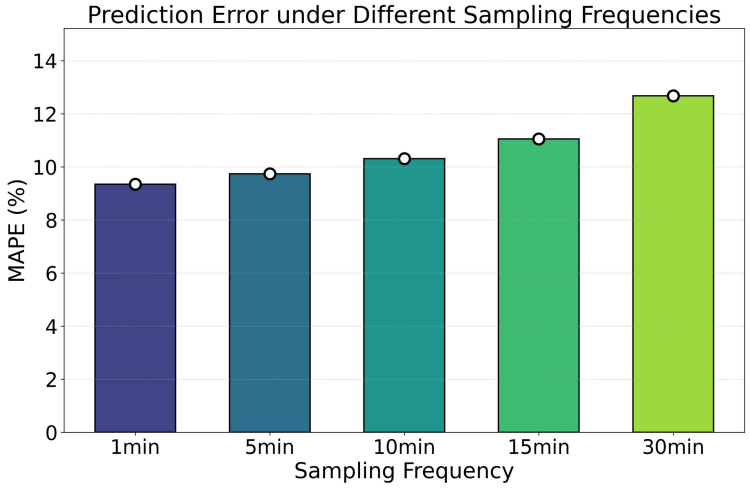

Figure 3. The impact of data sampling frequency on prediction accuracy

As shown in Figure 3, the data sampling frequency has a significant impact on the accuracy of the network traffic prediction model. The overall trend indicates that higher sampling frequency, or shorter sampling intervals, results in lower prediction errors. In particular, when sampling is performed every minute, the MAPE reaches its lowest value, approximately 9.35 percent. This shows that the model can effectively capture fine-grained traffic variations, which enhances its ability to predict future trends.

When the sampling interval increases to 10 minutes and 15 minutes, prediction errors rise significantly, and the model's performance declines. This suggests that sparse data may cause the loss of important fluctuation patterns, limiting the model's capacity to capture time-varying characteristics. In load balancing systems, such a decrease in accuracy may cause the scheduler to misjudge system status. This can lead to node congestion or unbalanced resource allocation, reducing the system's overall stability and service quality.

With a 30-minute sampling interval, the error becomes the highest. The MAPE exceeds 12 percent, indicating that low-frequency sampling fails to reflect real traffic dynamics. This information loss leads to enlarged prediction deviations. It also confirms the critical role of temporal data density in load forecasting tasks. In real-world deployment, relying solely on low-frequency data can limit the effectiveness of prediction algorithms and increase the delay in strategic decision-making.

This experiment further highlights the importance of high-frequency data in network load prediction tasks. Fine-grained sampling improves the model's sensitivity to traffic fluctuations and provides more timely and accurate guidance for load balancing decisions. Especially in multi-node and dynamic environments, the choice of sampling frequency directly affects the model's practical value and scalability. Therefore, high-resolution data collection should be prioritized during system design to ensure the effectiveness of the prediction module.

4. Conclusion

This study addresses the problem of load balancing optimization in complex network environments. A network traffic prediction model is proposed, which combines a gated recurrent structure with a multi-head self-attention mechanism. By integrating temporal modeling with global correlation learning, the model effectively captures the dynamic evolution of traffic across multiple nodes. It enables accurate forecasting of future traffic trends. Evaluation under multiple metrics shows that the proposed method offers strong robustness and high prediction accuracy. It provides more reliable data support for load balancing strategies and improves overall resource allocation efficiency in the system.

In terms of model design, this work focuses on challenges such as data diversity, structural heterogeneity, and strong temporal dependencies in network traffic prediction. The use of a time-window mechanism and attention structures enhances the model's ability to detect both short-term fluctuations and long-term trends. It also improves responsiveness to sudden traffic changes. In addition, this study conducts systematic experiments on sensitive hyperparameters such as sampling frequency and window size. The results confirm the model's adaptability and stability across different environmental settings. These findings are valuable for advancing network traffic modeling toward higher accuracy and stronger generalization.

From an application perspective, the proposed method has broad potential in various intelligent network scenarios. It is particularly applicable to resource scheduling in data centers, optimization in content delivery networks, and traffic-aware management in edge computing systems. Accurate traffic prediction helps avoid resource waste and node overload, thereby improving service quality and operational reliability. The model's structural generality and extensible inputs also allow it to adapt to different types of network architectures and business models. This makes it suitable for building unified, efficient, and intelligent network decision-support platforms in the future.

Several directions remain open for future research. First, improving the model's robustness against extreme and anomalous traffic events is a key challenge for stable forecasting. Second, incorporating multimodal data such as network logs and behavioral sequences could enhance the model's ability to understand complex contexts. Finally, integrating traffic prediction results with real-time load balancing strategies can help build an end-to-end closed-loop optimization system. This will support the shift from predictive assistance to automated decision-making in intelligent network management. As networks continue to grow and services become more complex, intelligent traffic modeling and scheduling mechanisms will play an increasingly central role in the future information infrastructure.

References

[1] Liu Y, Meng Q, Chen K, et al. ALB-TP: adaptive load balancing based on traffic prediction using GRU-Attention for Software-Defined DCNs[J]. Journal of Network and Computer Applications, 2025, 236: 104103.

[2] Owusu E T, Agyekum K A P, Benneh M, et al. A transformer-based deep q learning approach for dynamic load balancing in software-defined networks[J]. arXiv preprint arXiv:2501.12829, 2025.

[3] Chauhan N S, Kumar N. Confined attention mechanism enabled Recurrent Neural Network framework to improve traffic flow prediction[J]. Engineering Applications of Artificial Intelligence, 2024, 136: 108791.

[4] Zeb S, Rathore M A, Mahmood A, et al. Edge intelligence in softwarized 6G: Deep learning-enabled network traffic predictions[C]//2021 IEEE Globecom Workshops (GC Wkshps). IEEE, 2021: 1-6.

[5] Zhang C, Dang S, Shihada B, et al. Dual attention-based federated learning for wireless traffic prediction[C]//IEEE INFOCOM 2021-IEEE conference on computer communications. IEEE, 2021: 1-10.

[6] Tan M, Ba R, Li G. Data Center Traffic Prediction Algorithms and Resource Scheduling[J]. Sensors, 2022, 22(20): 7893.

[7] Zhang Y, Wang Y, Cao H, et al. Self‐similar traffic prediction for LEO satellite networks based on LSTM[J]. IET communications, 2025, 19(1): e12863.

[8] Li X, Song Z. Reinforcement Learning-Based Elastic Scaling Policy Optimization for Microservices[J]. 2024, 4(4).

[9] Li X, Yin R, Dang S. Structure-Guided Attention and Multi-Scale Behavior Modeling for Microservice Anomaly Detection[J]. 2025, 4(9).

[10]Liu Y, Zhang Z, Liang S. Intelligent Backend Failure Detection with Uncertainty Quantification and Asymmetric Risk Optimization[J]. 2024, 4(3).

[11]Ren L, Liu Y, Zhao Y. Learning-Based Intelligent Agents for Backend Resource Scheduling and Operational Decision Making[J]. 2025, 5(3).

[12]Yan R, Guo M. Enterprise Audit Risk Identification Through Temporal Context and Entity Relationship Anomaly Modeling[J]. 2024, 4(2).

[13]Nie Y. Corporate Cash Flow Forecasting Based on Counterfactual Time Series and Strategy Simulation[J]. 2024, 3(1).

[14]Yang Y. Memory-Enhanced Transformer for Backend Microservice Anomaly Detection[J]. 2024, 4(1).

[15]Huang Z. Dynamic User Interest Evolution Modeling for Personalized Recommendation Systems Based on Multi-Scale Temporal Awareness and Attention Mechanisms[J]. 2024, 4(1).

[16]Yang Y, Xu A. Log Semantic Graph Construction and System Event Anomaly Detection via Convolutional Neural Networks[J]. 2024, 4(2).

[17]Zhang X, He A. Lightweight Anomaly Detection for Edge Intelligence through Compressed Representation Learning[J]. 2025, 5(12).

[18]Huang Z, Luo J, Chen C. Learning-Driven Collaborative Scheduling for Multi-Tenant Distributed Inference Services[J]. 2025, 5(4).

[19]Li Y, Shao Z, Xu Y, et al. Dynamic frequency domain graph convolutional network for traffic forecasting[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 5245-5249.

[20]Qin Y, Fang Y, Luo H, et al. DMGCRN: Dynamic multi-graph convolution recurrent network for traffic forecasting[J]. arXiv preprint arXiv:2112.02264, 2021.

[21]Chai W, Luo Q, Lin Z, et al. Spatiotemporal Dynamic Multi-Hop Network for Traffic Flow Forecasting[J]. Sustainability (2071-1050), 2024 (14).