TSANTP: 공간 코딩 주의 메커니즘을 통합한 새로운 네트워크 트래픽 예측 모델

이용비 [1], 김경백 [2]
[1] 전남대학교 인공지능융합학과 석사과정
[2] 전남대학교 인공지능융합학과 교수

lilongfei0712@gmail.com, kyungbaekkim@jnu.ac.kr

TSANTP: A Novel Network Traffic Prediction Model Integrating Space Coding and Attention Mechanisms

LongFei Li [1], Kyungbaek Kim [2]
Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

With the widespread application of 5G technology, network traffic has increased unprecedentedly, which has a significant impact on network traffic management. Traditional network traffic prediction methods rely on time series analysis of seasonal patterns, ignoring the inherent spatial correlation of network traffic. Graph convolutional networks (GCN) learn spatial correlations. By combining GCN with time series models, spatiotemporal features can be captured simultaneously, thereby improving prediction accuracy. This paper introduces a new network traffic prediction model TSA-NTP based on the attention mechanism, which aims to more effectively capture spatiotemporal features in complex network environments.

1. Introduction

With the rapid increase in the number of mobile devices and connections globally, particularly with the widespread adoption of 5G technology, global network traffic is exhibiting unprecedented growth. Recent data indicates that by 2023, 5G devices and connections will constitute 15% of global mobile devices and connections, an increase from the previously projected 10%. This trend has resulted in a sharp rise in traffic between network devices, significantly complicating network traffic management. Traditional network traffic prediction methods predominantly rely on the analysis of seasonal patterns and time-series characteristics. [1]However, the nonlinearity and dynamic variability exhibited by real-world network traffic present substantial challenges to achieving high-precision predictions. [2] [3]

Network traffic prediction is typically regarded as a time-series forecasting task, wherein historical network traffic data is analyzed to develop a time-series model that predicts future network traffic. The accuracy of network traffic predictions has been enhanced with the development of local statistical algorithms for time-series forecasting, such as ARIMA. These models reduce prediction errors by capturing the seasonal patterns inherent in the time series. However, as the volume of data continues to grow and deep learning methods are increasingly applied in various fields, methods such as Recurrent Neural Networks (RNN) have shown greater efficacy in extracting temporal features from traffic flow sequences. [4]For instance, Convolutional Neural Networks can automatically extract features from network traffic data, resulting in improved predictive performance.

Despite these advancements, existing prediction methods primarily emphasize the temporal dimension and often overlook the spatial correlations present in network traffic. Network traffic is exchanged between nodes at multiple sites and traverses network links. Due to the adjacency inherent in network topology, there is significant correlation in traffic behavior between these links. For example, links adjacent to congested links are more likely to be affected, leading to the propagation of congestion. [5] As a result, traditional time-series prediction models struggle to effectively leverage network topology to capture spatial information when forecasting traffic at network nodes.

To address this challenge and effectively capture the spatial features of network traffic, GCN provide a powerful solution. [6] GCN extend convolution operations to graph structures, enabling the learning of correlations within non-Euclidean spaces derived from network topology, thereby

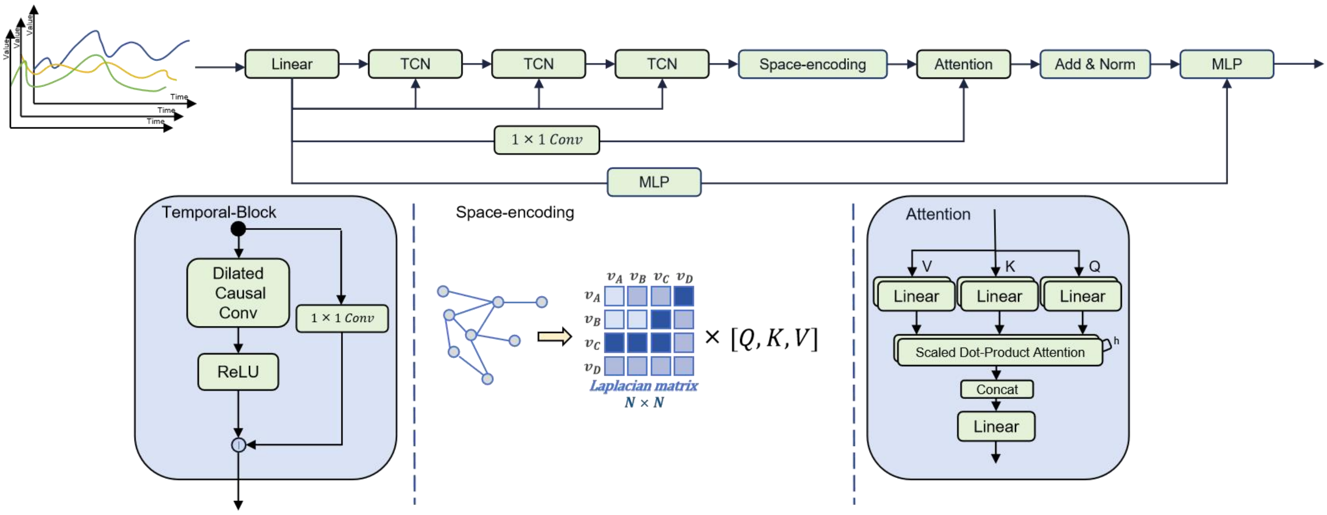

Figure 1: TSANTP model structure

capturing the complex relationships between network links. While time-series models excel at extracting temporal and spatial features, combining these techniques with GCN allows for the simultaneous capture of spatiotemporal features in network traffic, thus enhancing both prediction accuracy and robustness.

Moreover, with the growing success of Transformers in the time-series domain, they have been incorporated into network traffic prediction models to further improve predictive accuracy. [7]

In this paper, we propose a novel attention-based network traffic prediction model, TSA-NTP, which effectively captures the spatiotemporal features of complex network environments. We provide a detailed discussion on the construction of a graph Laplacian matrix to characterize the correlation between adjacent links within the network. We also generate a network traffic matrix using NSFNET and validate the performance of the proposed method through experiments conducted under varying network conditions. Our research not only offers a new perspective on traffic prediction in complex network environments but also provides theoretical support for the future deployment of AI-based predictive automation technologies aimed at managing and simplifying network operations across all domains.

2. Methods

In this study, we consider a network adjacency matrix $G \in R^{N \times N}$, where N represents the number of nodes in the network. The historical network traffic data is represented as $\mathrm{X} \in \mathrm{R}^{\mathrm{T} \times \mathrm{N} \times \mathrm{F}}$, where T is the number of time steps in the historical network traffic, and F is the feature dimension. The task of network traffic prediction is to predict future traffic X^{P}based on the historical traffic X^{T}.

First, we apply a linear layer to the historical traffic data to project the time steps into a higher dimension. This step is implemented as follows:

$$H_L = Xw_1 + b_1$$

where W_1 is the weight matrix, and b_1 is the bias term. This step enhances the model's capability to represent the temporal features.

Next, we employ a multi-layer Temporal Convolutional Network (TCN) to further model the processed data. Each layer is represented as:

$$H_{i+1} = Re\,LU(Conv1D(H_i, W_i, d) + b_i + H_L)$$

where ReLU is the activation function, Conv1D represents the one-dimensional convolution operation, W_i is the convolution kernel, and d is the dilation rate. After passing through multiple convolutional layers, we obtain the final TCN output:

$$H_{TCN} = H_n$$

The TCN module is designed to capture the temporal dependencies in the historical traffic data, particularly in long sequences.

To further leverage the network structure information, we compute the graph Laplacian matrix L, which is defined as:

$$\mathrm{L} = \mathrm{D} - \mathrm{A}$$

where A is the adjacency matrix representing the network structure, and D is the degree matrix. To prevent gradient explosion due to uneven degrees, we adopt the symmetrically normalized graph Laplacian matrix, expressed as:

$$\tilde{L} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

Next, we apply the Laplacian matrix to the feature matrix and compute the query (Q), key (K), and value (V) matrices for the attention mechanism:

$$H_{Q,K,V} = ReLU\ \ (H_{TCN}\tilde{L}w_{Q,K,V} + b_{Q,K,V} + H_{1\times1})$$

Where $\mathrm{W}_{\mathrm{Q,K,V}}$ are the parameter matrices, and $H^t_{1\times1}$ represents the result of H_L after a 1x1 convolution:

$$H^t_{1\times1} = b + \sum_{i=1}^{C} w_i \cdot H_i(t)$$

The attention mechanism is computed as follows:

$$H_{ATT} = \mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Finally, we perform residual connection and normalization on the output of the attention mechanism and the TCN output:

$$H_{ATT} = LayerNorm(H_{ATT} + H_{TCN})$$

After the above processing, the resulting feature vector is passed through a Multi-Layer Perceptron (MLP) to produce the final prediction:

$$Y = MLP(H_{norm} + H_{MLP})$$

where the MLP is defined as:

$$MLP(H) = W_f ReLU(HW_e + b_e) + b_f$$

Thus, we obtain the final network traffic prediction result Y.

To ensure the stability and generalization ability of the model, we introduce residual connections and LayerNorm layers in the network, which effectively prevent gradient vanishing or explosion. Additionally, we employ appropriate weight initialization and regularization techniques (such as Dropout) to further improve the model's robustness.

3. Experimentation and Evaluation

The model is implemented using Pytorch-GPU 2.01 based on Python 3.11 and is trained on a PC running Windows 11 Education WSL, equipped with an Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, an Nvidia GeForce RTX 2060super GPU, and 64 GB of memory.

We created a simulator using OMNeT++ and applied 200 different routing schemes on NSFNET. We then generated 50,000 traffic matrix samples to reflect various data flows within the network. NSFNET was operational from the 1980s to the early 1990s, connecting supercomputing centers and universities in the United States. It consisted of 14 nodes and 42 directed links.We used MSE (Mean Squared Error) as the evaluation metric, with the formula:

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

BIGRU: BIGRU is a bidirectional GRU model that can capture temporal dependency information from both the forward and backward directions of the sequence.

BiLSTM: BiLSTM is a bidirectional LSTM model that processes sequence data in both forward and backward directions to capture global temporal dependencies.

STGCN: STGCN combines graph convolutional networks and temporal convolutional networks to simultaneously capture spatial and temporal dependencies in data. [8]

TGCN: TGCN combines graph convolutional networks with time series models to capture spatial and temporal correlations in data. [9]

		Model				
Strength	length	BIGRU	BiLSTM	STGCN	TGCN	TSANTP
9	3	0.111	0.118	0.132	0.154	0.101
	6	0.113	0.118	0.129	0.139	0.102
12	3	0.105	0.109	0.109	0.143	0.088
	6	0.106	0.109	0.112	0.143	0.084
15	3	0.121	0.124	0.131	0.162	0.107
	6	0.121	0.124	0.133	0.165	0.109

Table 1:Performance comparison of TSANTP and other models on NSFNET

Based on the comparative analysis of the data in the table, TSANTP outperformed other models under different network intensities and prediction lengths. When the network intensity was 9 and the prediction length was 3, my model achieved the lowest error of 0.101, whereas the errors for BIGRU, BiLSTM, STGCN, and TGCN were 0.111, 0.118, 0.132, and 0.154, respectively. Similarly, under a network intensity of 12 and a prediction length of 3, my model again performed excellently with an error of 0.088, significantly lower than other models. Additionally, regardless of whether the network intensity was 15 or the prediction length was 6, my model consistently maintained the lowest error, demonstrating strong robustness and predictive capability. This indicates that my model possesses superior generalization ability and accuracy when handling tasks with varying network intensities and prediction requirements, making it suitable for practical applications.

We compared our model with four benchmark methods on the NSFNET dataset. Table 1 shows the results of prediction performance at network intensities of 9, 12, and 15, with prediction steps of 3 and 6. As seen in Table 1, our TSANTP achieved the best performance in terms of MAE (Mean Absolute Error). When the network intensity was 9 and the prediction length was 3, my model achieved the lowest error of 0.101, whereas the errors for BIGRU, BiLSTM, STGCN, and TGCN were 0.111, 0.118, 0.132, and 0.154, respectively. Similarly, under a network intensity of 12 and a prediction length of 3, my model again performed excellently with an error of 0.088, significantly lower than other models. Furthermore, regardless of whether the network intensity was 15 or the prediction length was 6, my model consistently maintained the lowest error, demonstrating strong robustness and predictive capability. It can also be observed that traditional time series analysis methods yielded better results, while models considering both temporal and spatial correlations, such as STGCN and TGCN, were less effective. This suggests that these methods have limited capacity in modeling nonlinear and complex traffic data. Our TSANTP employs an attention mechanism and has outperformed previous state-of-the-art models, proving the advantages of our model in combining spatiotemporal features in network traffic prediction.

4. Conclusion

This study proposes a network traffic prediction model (TSA-NTP) based on the attention mechanism, which can effectively capture spatiotemporal characteristics in a complex network environment, thereby improving the accuracy and robustness of the prediction. Experimental results show that TSA-NTP performs well under different network strengths and prediction lengths, especially under high network strength and long prediction length, the model has the lowest error, showing superior generalization ability and prediction performance. In contrast, traditional time series analysis methods and models that only consider spatiotemporal correlations do not perform as well as TSA-NTP in processing nonlinear and complex traffic data. This shows that the model proposed in this study not only provides theoretical support for artificial intelligence-based prediction automation technology, but also has the potential to be widely used in actual network operation and management.

Acknowledgement

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2022-00156287, 50%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, 50%) grant funded by the Korea government(MSIT).

참고문헌

[1] Joshi M, Hadi T H. A review of network traffic analysis and prediction techniques[J]. arXiv preprint arXiv:1507.05722, 2015.

[2] Li L, Kim K. GTT-NTP: A Graph Convolutional Networks-Based Network Traffic Prediction model[C]//NOMS 2024-2024 IEEE Network Operations and Management Symposium. IEEE, 2024: 1-7.

[3]Mukherjee S, Ray R, Samanta R, et al. Nonlinearity and chaos in wireless network traffic[J]. Chaos, Solitons & Fractals, 2017, 96: 23-29.

[4]Lazaris A, Prasanna V K. Deep learning models for aggregated network traffic prediction[C]//2019 15th International Conference on Network and Service Management (CNSM). IEEE, 2019: 1-5.

[5]Yeom S, Choi C, Kolekar S S, et al. Graph convolutional network based link state prediction[C]//2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2021: 246-249.

[6]Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications[J]. AI open, 2020, 1: 57-81.

[7]Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[8]Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting[J]. arXiv preprint arXiv:1709.04875, 2017.

[9]Zhao L, Song Y, Zhang C, et al. T-GCN: A temporal graph convolutional network for traffic prediction[J]. IEEE transactions on intelligent transportation systems, 2019, 21(9): 3848-3858.