네트워크 트래픽 예측의 진화: 딥러닝 및 그 너머에 대한 조사

Van-Vi Vo[1], Thi Le Quyen Nguyen[2], Sardar Jaffar Ali[1], 추현승[2]
[1]성균관대학교 컨버전스연구소
[2]성균관대학교 전자전기컴퓨터공학과
{vovanvi, ntlequyen, jaffar.ali, choo}@skku.edu

Evolving Network Traffic Prediction: A Survey of Deep Learning and Beyond

Van-Vi Vo[1], Thi Le Quyen Nguyen[2], Sardar Jaffar Ali[1], Hyunseung Choo[2]
[1]Convergence Research Institute, Sungkyunkwan University
[2]Dept. of Electrical and Computer Engineering, Sungkyunkwan University

요 약

As telecommunications advance toward next-generation networks beyond 5G, they encounter the growing challenge of accommodating more users and devices, leading to increased traffic with limited resources. Accurate traffic analysis and demand forecasting are vital for creating intelligent networks, and Deep Learning (DL) harnesses vast network data to improve prediction accuracy and optimize service design and management. This survey explores recent breakthroughs in network traffic prediction (NTP), focusing on DL-based models to highlight popular techniques and categorize existing research into Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs)/ Temporal Convolutional Networks (TCNs), Graph Neural Networks (GNNs), and Large Language Models (LLMs). It offers a detailed, tutorial-style overview of these methods, supported by practical data analyses and experiments, and addresses their performance in real-world scenarios. The paper concludes with insights into current challenges and future opportunities, providing a roadmap for advancing NTP through DL innovations in evolving network environments.

1. Introduction

In recent years, the softwarization of networks has transformed how operators manage their infrastructures, providing greater flexibility and control to enhance network performance. This shift has paved the way for anticipatory decision-making, allowing proactive measures like traffic engineering, resource allocation, and service orchestration to adapt to fluctuating traffic demands. Unlike traditional reactive methods that rely on human intervention, this anticipatory approach holds significant promise for improving resource efficiency and boosting end-user quality of service. However, the success of these proactive strategies depends heavily on the precision of traffic predictions, highlighting the essential role of time-series forecasting in driving innovative network management solutions.

Network traffic prediction (NTP) has been a key area of study in networking for decades, with early explorations dating back to the 1970s and a recent boom driven by advancements in deep learning technologies. Various surveys have emerged, organizing the vast body of NTP research into different categories, focusing primarily on deep learning methods such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and emerging Large Language Models (LLMs). Recent reviews, such as those by Cao et al. [1] and Huang et al. [2] on deep neural networks, and Jiang and Luo [3] on GNNs, have concentrated on specific techniques, offering limited perspectives. In contrast, broader surveys like those by Joshi and Hadi [4] cover a range of methods, including preprocessing techniques like discretization and feature selection to improve data quality, alongside nonlinear prediction approaches. Similarly, Jiang [5] targets cellular traffic prediction, categorizing models into machine learning and deep learning frameworks, reflecting the diversity of current research.

While existing surveys offer useful classifications and insights into specific deep learning techniques, this paper takes a more expansive approach by synthesizing a broad spectrum of NTP research. It focuses on categorizing and evaluating deep learning methods – RNNs, CNNs/TCNs, GNNs, and LLMs – and this survey provides a comprehensive literature review. Unlike previous reviews, this study not only highlights recent advancements but also provides a foundation

for understanding their real-world performance, paving the way for deeper exploration of challenges and future innovations in the field.

The survey is structured as follows. Section 2 provides a literature review, categorizing deep learning methods into four types – RNNs, CNNs/TCNs, GNNs, and LLMs – and includes an overview of key approaches. Section 3 discusses challenges and future directions. Finally, Section 4 offers concluding remarks.

2. Literature Review

2.1 Recurrent neural networks-based approaches

RNNs and their improved versions are great at spotting patterns that change over time in data sequences, making them a good fit for NTP tasks where traffic trends follow time-based rhythms. Li et al. [6] used a special type of RNN called Long Short-Term Memory (LSTM) networks to predict traffic in cellular networks and transportation systems, respectively. LSTMs, first created by Hochreiter and Schmidhuber [7], are designed to handle time series data effectively. They work with a unique setup that includes three key steps: (1) A "forget gate" decides what old information to let go of by looking at the current data and past results, using a simple process to filter out unimportant details. (2) An "input gate" figures out what new information to keep, combining it with a fresh set of data to update the memory. (3) An "output gate" then decides what to share next based on the updated memory. This design helps LSTMs remember important details over long periods, solving issues where regular RNNs forget earlier patterns.

A simpler version of RNNs is the GRU, introduced by Cho et al. [8], which uses just two steps (update and reset) to make calculations easier and faster. Patil et al. [9] applied GRUs to predict IoT traffic and found they worked better than the older ARIMA method. Likewise, Fu et al. [10] tested GRUs and LSTMs for traffic flow prediction in California, showing both outperformed traditional autoregressive moving average models. In GRUs, the update step balances old and new information, while the reset step can ignore past data when needed, starting fresh with the latest input.

While RNN-based methods are effective for tracking sequential traffic patterns in NTP, they can be heavy on computing power and struggle with very long sequences. This often makes it necessary to combine them with other models to get the best results.

2.2 Image-based approaches

Image-based approaches transform network traffic data into matrix or tensor representations, leveraging CNNs to extract spatial patterns. These methods treat traffic matrices as images, enabling the detection of local correlations.

CNNs, as described by LeCun et al. [11], consist of convolutional layers where kernels slide over input matrices to produce feature maps. Bega et al. [12] constructed a distance matrix based on time series similarities among base stations and applied CNNs for forecasting, emphasizing local patterns. Chen et al. [13] used CNNs for traffic flow prediction, capitalizing on their ability to process image-like data. Key operations include valid convolution (reducing output dimensions by sliding kernels within borders) and padding techniques (zero or symmetric) to preserve or expand dimensions. CNNs reduce parameters through sparse connections and shared weights, making them efficient for high-dimensional data. In NTP, tensors are often used, as in Ong et al. [14] and Deng et al. [15], where rank-3 tensors store spatiotemporal information.

TCNs, an extension of CNNs for sequential data, utilize dilated convolutions to capture long-range dependencies with fewer layers. Introduced by Bai et al. [16], TCNs employ causal convolutions – ensuring predictions rely solely on past data – and residual connections to enhance stability and gradient flow. This architecture offers significant advantages over RNNs, including parallelizability and computational efficiency, making TCNs well-suited for large-scale NTP datasets. However, careful tuning of dilation rates is often required to balance model complexity and performance. In the context of NTP, Wang et al. [17] integrated TCNs with Transformers to address limitations in capturing both short- and long-term traffic patterns. This hybrid approach leverages TCNs' temporal modeling strengths and Transformers' ability to handle complex dependencies, improving prediction accuracy across diverse network scenarios. While TCNs excel in handling sequential data efficiently, their effectiveness in NTP may depend on dataset characteristics and the specific integration with other models.

Overall, image-based methods enhance spatial feature extraction but often need integration with temporal models for comprehensive STP.

2.3 Graph Neural Networks-based approaches

GNNs model networks as graphs, with nodes representing devices (e.g., routers, base stations) and edges denoting connections, enabling the capture of spatiotemporal dependencies. Introduced by Scarselli et al. [18], GNNs learn node representations by aggregating neighbor features. Wu et al. [19] provided a taxonomy: (1) Recurrent GNNs (RecGNNs) use RNNs for iterative neighbor information exchange until convergence. (2) Convolutional GNNs (ConvGNNs) generalize convolutions to graphs, stacking layers for feature learning. (3) Graph Autoencoders encode graphs into latent spaces for unsupervised reconstruction. (4) Spatiotemporal GNNs combine graph convolutions with CNNs/RNNs for dynamic data.

It is important to emphasize that the GNNs above have different mathematical formulations and applications, where the details of each case are also discussed above. Wang et al. [20] proposed the Time-Series Graph Attention Network (TSGAN), using dynamic time warping (DTW) for cellular traffic forecasting, outperforming standard GNNs and GRUs across short-, mid-, and long-term horizons. Zhou et al. [21] introduced a Spatiotemporal Graph Convolutional Network, surpassing baselines like LSTM, ConvLSTM, and diffusion CNN-RNN in accuracy.

GNNs excel in topology-aware predictions but face challenges in scalability for large graphs and require domain-specific graph construction.

2.4 Large Language Models-based approaches

LLMs, pretrained on extensive textual datasets, have recently emerged as a novel paradigm for NTP by leveraging their proficiency in processing sequential data and generating context-aware forecasts. These models reinterpret traffic time series as textual sequences, facilitating zero-shot or few-shot

predictions through prompting mechanisms. This approach capitalizes on LLMs' pre-existing language understanding capabilities, adapting them to the spatiotemporal dynamics of network traffic.

Liu et al. [22] developed ST-LLM+, a graph-enhanced spatiotemporal LLM that integrates graph structures with LLM architectures like GPT to model spatial dependencies, achieving superior performance on urban traffic prediction tasks. Chen et al. [23] introduced UrbanGPT, a spatiotemporal LLM framework which focuses on traffic flow forecasting and provides interpretable outputs through natural language explanations, enhancing model transparency. For mobile network applications, Zhang et al. [24] proposed an LLM-based framework that employs efficient in-context learning to predict mobile traffic, reducing computational demands while surpassing traditional deep learning models in energy-efficient scenarios. Meanwhile, Ma et al. [25] presented TPLLM, a pretrained LLM fine-tuned for traffic prediction, emphasizing the role of embedding modules in adapting sequential traffic data, thereby improving forecast accuracy. Together, these studies underscore the versatility of LLMs in addressing diverse NTP challenges.

LLMs offer significant advantages, including the ability to handle multimodal data (e.g., incorporating external factors like events) and support transfer learning across domains. However, their deployment is constrained by high resource requirements and the risk of overfitting without optimized prompting strategies. Future research directions may focus on developing hybrid LLM-deep learning architectures to enable real-time NTP, particularly in the context of emerging 6G networks.

3. Challenges and Future Directions

3.1 Challenges

Computational Complexity and Performance Trade-offs: A major challenge in NTP is achieving high accuracy while keeping computational demands manageable, as complex deep learning models increase training and inference times. Many studies neglect to evaluate model complexity or runtime, limiting their suitability for resource-constrained settings like edge networks. We urge the research community to include comparative analyses against benchmarks, assessing both accuracy and efficiency metrics (e.g., FLOPs, latency), to clarify trade-offs and guide real-world adoption.

Benchmarking Against Baselines: Advances in time series forecasting, including NTP, reveal that sophisticated deep learning models often lag behind simple baselines, as seen in events like the M4 Forecasting Competition. NTP literature frequently skips direct comparisons with baseline deep learning models (e.g., vanilla LSTMs) or uses inconsistent hyperparameters, skewing evaluations. We suggest a shared repository of standardized baselines to ensure fair and reproducible assessments, with the experimental code in this survey serving as an initial step toward this goal.

Practical Deployment and Optimization: NTP research highlights its importance for resource allocation and anticipatory decision-making, yet it falls short in showing how models apply to real-world optimization tasks like energy-efficient scaling or 6G closed-loop control. Current evaluations focus on error metrics (MAE, MAPE) but ignore live deployment impacts. Future efforts should test models in simulated or real settings to measure operational benefits, such as lower latency or costs, compared to alternatives.

Lack of Standardized Datasets: Unlike image processing with datasets like ImageNet, NTP lacks a universal benchmark due to operators' reluctance to share sensitive traffic data under NDAs, resulting in sparse, outdated public datasets that hinder reproducibility. A promising solution is synthetic data generation using techniques like STAN's generative models or NetDiffus' diffusion methods to create realistic spatiotemporal traffic traces, potentially forming a shared community dataset to overcome privacy issues and improve comparisons.

3.2 Future directions

Our survey of NTP models highlights the profound influence of deep learning across both scientific research and practical applications, driven by notable improvements in prediction accuracy. This analysis also uncovers several promising avenues for refining existing deep learning architectures, including RNNs, CNNs/ TCNs, GNNs, and LLMs, to further elevate forecasting performance. These pathways are designed to meet the dynamic needs of next-generation networks, capitalizing on the distinct capabilities of each method to enhance network management.

One promising direction is the development of hybrid strategies that integrate deep learning, notably LLMs, with alternative modeling approaches, optimized through a cohesive training framework. These combinations hold the potential to outperform standalone models by merging the contextual awareness of LLMs with the structured insights of other techniques, though current efforts are still in early stages and require more extensive research to realize their full capabilities. Another critical focus is the reevaluation of loss functions used in NTP. Conventional metrics like MAE and MSE are effective for predicting traffic trends but may not align with the strategic decision-making needs of network operations. Ongoing innovations are exploring custom loss functions, shaped by expert input, to better match model outputs to user needs, while emerging meta-learning techniques aim to automate the design of loss functions tailored to specific goals, boosting model flexibility.

A major area of future development is the transition to online forecasting frameworks, moving away from current practices that rely on offline training with historical data, testing with replayed datasets, and minimal consideration of inference latency. Future systems should support real-time operation to meet stringent latency requirements, such as millisecond-level needs in radio access networks, necessitating research into accuracy degradation over time, the necessity for retraining or continual learning, and the adaptation of evaluation methods to handle live streaming data effectively. For LLMs, this could mean dynamically adjusting prompts to keep predictions relevant amid shifting traffic patterns. Lastly, enhancing the generalization of NTP models remains a pressing issue, as most are validated in specific contexts like urban regions, limiting their broader applicability. The transferability of models across different regions remains uncertain, underscoring the need for wider testing. Progress in transfer learning, potentially extended to LLMs' pretraining advantages, offers a pathway to develop more versatile and widely applicable forecasting tools.

4. Conclusion

This survey provides a comprehensive examination of NTP, showcasing the evolution from RNNs to advanced deep learning techniques, including CNNs/TCNs, GNNs, and the emerging role of LLMs. By categorizing diverse approaches and offering a tutorial-style explanation of their mechanics, the study equips researchers and practitioners with the knowledge to apply these models effectively. The practical experiments and data analyses underscore the potential of DL to enhance prediction accuracy, while the identified challenges – such as computational complexity, benchmarking gaps, and the need for standardized datasets – highlight areas for future focus. Looking ahead, the integration of hybrid models, online forecasting frameworks, and improved generalization through transfer learning, including LLM adaptations, promises to shape the future of NTP, particularly as networks transition to 6G technologies.

Acknowledgement. This work was partly supported by the Korea government (MSIT), IITP, Korea, under the ICT Creative Consilience program (IITP-2025-RS-2020-II201821, 40%) and Development of 6G Network Integrated Intelligence Plane Technologies (RS-2024-00392332, 25%); and by the National Research Foundation of Korea(NRF) under grant (RS-2024-00343255, 35%).

참고문헌

[1] P. Cao, F. Dai, G. Liu, J. Yang, and B. Huang, ‘‘A survey of traffic prediction based on deep neural network: Data, methods and challenges,’’ in Cloud Computing. Cham, Switzerland: Springer, 2022, pp. 17–29.

[2] C.-W. Huang, C.-T. Chiang, and Q. Li, ‘‘A study of deep learning networks on mobile traffic forecasting,’’ in Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC), Oct. 2017, pp. 1–6.

[3] W. Jiang and J. Luo, ‘‘Graph neural network for traffic forecasting: A survey,’’ Exp. Syst. Appl., vol. 207, Nov. 2022, Art. no. 117921.

[4] M. Joshi and T. H. Hadi, ‘‘A review of network traffic analysis and prediction techniques,’’ 2015, arXiv:1507.05722.

[5] W. Jiang, ‘‘Cellular traffic prediction with machine learning: A survey,’’ Exp. Syst. Appl., vol. 201, Sep. 2022, Art. no. 117163.

[6] X. Luo, D. Li, Y. Yang, and S. Zhang, ‘‘Spatiotemporal traffic flow prediction with KNN and LSTM,’’ J. Adv. Transp., vol. 2019, pp. 1–10, Feb. 2019.

[7] S. Hochreiter and J. J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 80–1735, 1997

[8] H. Schwenk, and Y. Bengio, ‘‘Learning phrase representations using RNN encoder–decoder for statistical machine translation,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[9] S. A. Patil, L. A. Raj, and B. K. Singh, ‘‘Prediction of IoT traffic using the gated recurrent unit neural network-(GRU-NN-) based predictive model,’’ Secur. Commun. Netw., vol. 2021, pp. 1–7, Oct. 2021.

[10] R. Fu, Z. Zhang, and L. Li, ‘‘Using LSTM and GRU neural network methods for traffic flow prediction,’’ in Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC), Nov. 2016, pp. 324–328.

[11] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521, no. 7553 (2015): 436-444.

[12] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, ‘‘DeepCog: Cognitive network management in sliced 5G networks with deep learning,’’ in Proc. IEEE Conf. Comput. Commun. (INFOCOM), Apr. 2019, pp. 280–288.

[13] C. Chen, K. Li, S. G. Teo, X. Zou, K. Li, and Z. Zeng, ‘‘Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks,’’ ACM Trans. Knowl. Discovery Data, vol. 14, no. 4, pp. 1–23, Aug. 2020.

[14] Y. J. Ong, M. Qiao, and D. Jadav, ‘‘Temporal tensor transformation network for multivariate time series prediction,’’ in Proc. IEEE Int. Conf. Big Data, Dec. 2020, pp. 1594–1603.

[15] T. Deng, M. Wan, K. Shi, L. Zhu, X. Wang, and X. Jiang, ‘‘Short term prediction of wireless traffic based on tensor decomposition and recurrent neural network,’’ Social Netw. Appl. Sci., vol. 3, no. 9, pp. 1–14, Sep. 2021.

[16] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." *arXiv preprint arXiv:1803.01271* (2018).

[17] Ren, Qianqian, Yang Li, and Yong Liu. "Transformer-enhanced periodic temporal convolution network for long short-term traffic flow forecasting." *Expert Systems with Applications* 227 (2023): 120203.

[18] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, ‘‘The graph neural network model,’’ IEEE Trans. Neural Netw., vol. 20, no. 1, pp. 61–80, Jan. 2016

[19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, ‘‘A comprehensive survey on graph neural networks,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4–24, Apr. 2021.

[20] Z. Wang, J. Hu, G. Min, Z. Zhao, Z. Chang, and Z. Wang, ‘‘Spatial-temporal cellular traffic prediction for 5G and beyond: A graph neural networks-based approach,’’ IEEE Trans. Ind. Informat., early access, Jun. 20, 2022

[21] X. Zhou, Y. Zhang, Z. Li, X. Wang, J. Zhao, and Z. Zhang, ‘‘Large-scale cellular traffic prediction based on graph convolutional networks with transfer learning,’’ Neural Comput. Appl., vol. 34, no. 7, pp. 5549–5559, Apr. 2022.

[22] Liu, Chenxi, Kethmi Hirushini Hettige, Qianxiong Xu, Cheng Long, Shili Xiang, Gao Cong, Ziyue Li, and Rui Zhao. "ST-LLM+: Graph Enhanced Spatio-Temporal Large Language Models for Traffic Prediction." *IEEE Transactions on Knowledge and Data Engineering* (2025).

[23] Li, Zhonghang, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. "Urbangpt: Spatio-temporal large language models." In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5351-5362. 2024.

[24] Zhang, Han, Akram Bin Sediq, Ali Afana, and Melike Erol-Kantarci. "Mobile Traffic Prediction using LLMs with Efficient In-context Demonstration Selection." *IEEE Transactions on Communications* (2025).

[25] Ma, Tian, Yixuan Zhao, Minda Li, Yue Chen, Fangshu Lei, Yanan Zhao, and Maazen Alsabaan. "TPLLM: A traffic prediction framework based on pretrained Large Language Models." *Applied Soft Computing* (2025): 113840.