A Survey on Deep Learning in Traffic Forecasting and Traffic State Estimation

Zhengyou Han

Ito's Lab
University of Tokyo

I. Introduction

Transportation is an essential and ubiquitous element of our society in our daily lives. With advancements in digitization, new intelligent transportation applications such as car-hailing and autonomous vehicles (AV) are improving the quality of our lives. However, they also impose an increasing burden on transportation management in addition to the growth of cities. As traffic systems grow in complexity, there is a pressing need for more efficient methods to analyze and organize them. Fortunately, the advancements in sensors offer the access to data with higher fidelity, enhanced timeliness, and unprecedented abundance. Researchers, therefore, see significant potential in data-driven methods within the realm of transportation science.

Traffic forecasting has always been among the most extensively researched topics of intelligent transportation systems(ITS). The goal of forecasting is to accurately predict traffic parameters (such as density, speed, or flow) across different lengths of time steps. This ability in forecasting is essential for various ITS applications such as traffic management and pedestrian control systems. During the last decade, efforts in traffic forecasting have increasingly focused on data-driven methods, resulting in numerous attempts to build flexible, large-scale, and real-time forecasting models. Currently, the rise of deep learning is further reshaping the field of traffic forecasting. Traffic state estimation(TSE) is a very similar field as traffic forecasting. However, by focusing on reconstruction of the real-time traffic parameters, TSE obtains a very unique position in intelligent transportation system studies. In traffic forecasting, researchers tend to predict the future based on retrieved data with utilizing neural network to foresee the changes. But TSE study is more like a science attempting to find the currently missing elements in the traffic network. At the very beginning, scientists put their efforts in the network reconstruction based on traffic models, which are generally inspired by the hydrodynamic models. Although the traffic flow seems comply the law of liquid, these model-based methods failed to capture all the features of the network with balanced complexity. With the emerging data-driven methods in transportation science, modern TSE studies are categorized into three types: model-driven methods, data-driven methods, and physical-informed deep learning methods(PIDL) which is a combination of former two types, though some believe PIDL is much more data-driven fouced than it claims.

Despite the higher accuracy and robustness of recent deep learning (DL) works, traffic forecasting and traffic state estimation (TSE) still lack efficient solutions in practical applications. Firstly, owing to the fact that he large size of neural networks and the complexity of traffic forecasting problems require extensive computational resources, most works address the problem within 'tiny' traffic networks. Secondly, the low interpretability of MLP-based methods compromises the authenticity of results from a decision maker's perspective. In most prevailing methods, the importance of results' interpretability is often overlooked. Transportation science, which provides crucial advice with a wide impact on society, requires sufficient what-if analysis and values the solidity of conclusions, sometimes even over performance.

Thirdly, while transportation science aims to enhance the quality of traffic and life globally, many areas lack the qualified data needed to implement advanced learning models. Previous works often neglect the replicability of training models in different contexts, yet this capability is a high priority for practical implementations. Lastly, the generalizability of models has been largely overlooked in ITS studies. A model with good generalizability should perform well in unseen situations, which are often more critical for traffic analysis and operations.

In the following, I am going to introduce some iconic DL methods in a chronical manner as well as a novel framework of DL potentially capable of improving current methods. And then, a discussion of my future research direction will follow.

II. A glance at the deep learning topics in forecasting and TSE

A. Graph Neural Network in traffic forecasting

During the last decade, researchers have harnessed the power of deep learning in traffic forecasting to develop flexible, large-scale, and real-time estimation models. However, traditional deep learning models may neglect some unique properties of traffic networks. For instance, most recent studies based on traditional deep learning architectures use convolution operators to consider spatial dependencies among data points. However, due to the special characteristics of transportation networks, spatial correlations are not necessarily distributed in Euclidean space. For example, a

railway close to a highway parallel to it may be near in Euclidean distance but should be categorized into different transportation systems with little spatial correlation.

This is where graph neural networks (GNNs) change the game. Scientists are using GNNs to capture features that traditional methods fail to. One of the first attempts comes from Shahsavari [1], who proposed a graph-oriented model for considering the spatial-temporal correlations between data captured by sensors. In this framework, nodes correspond to sensor locations with features such as traffic flow, density, and speed, while edges represent spatial interrelations forced by network topology, such as length, capacity, and direction. A GNN model is trained in a supervised learning approach to predict short-term future traffic conditions, serving as a general framework for subsequent works.

Following initial attempts towards graph-based learning, later works can be categorized into three genres: recurrent GNNs, convolutional GNNs, and graph autoencoders. In 2017, Li et al. [17] developed the diffusion convolutional recurrent neural network (DCRNN), incorporating recurrent units into GNNs. This model captures spatial dependencies as diffusion convolution on a directed graph while temporal dependencies are captured by diffusion convolutional gated recurrent units. Later, the spatial-temporal graph convolutional network (STGCN) [2] was proposed in 2018, incorporating the convolution operator in the GNN model. STGCN consists of two spatial-temporal convolutional blocks, followed by a fully connected layer, showing significant improvement compared to baselines such as Fully-connected LSTM(FC-LSTM), and Forward Neural Networks(FNN). And STGCN and DCRNN have become popular baselines for later forecasting studies.

Later, Zhao et al. [3] integrated gated recurrent units(GRU) and graph convolutional networks(GCN), proposing T-GCN to capture the topological structure of traffic networks while considering spatial dependencies using a graph convolution network. Cui et al. [7] combined graph convolutional methods with long-short term memory neural networks, proposing the state-of-the-art TGC-LSTM. In 2022, Shin and Yoon [4] proposed the multi-weight traffic graph convolutional network (MW-TGC), utilizing multi-weighted adjacency matrices to combine features such as speed limit, distance, and angles between road segments. Their model includes a Seq2Seq model with LSTM units to learn temporal relationships from the weighted graph convolution. Meanwhile, by defining the neighborhood matrices and the free flow reachability matrix, they captured the graph edge properties and high-order neighborhood in the traffic graph.

Recently, efforts in forecasting have shifted to attention-based methods to consider dynamic temporal dependencies. Guo et al. [8] developed an attention-based spatial-temporal graph convolutional network model (ASTGCN), which consists of three independent components for hourly, daily, and weekly time intervals. Each component is followed by graph convolutions for capturing spatial patterns and standard convolutions for describing temporal features. Inspired by the encoder-decoder framework, Pan et al. [5] proposed ST-MetaNet, a spatial-temporal meta graph attention network for multi-step traffic forecasting. ST-MetaNet consists of an RNN for embedding historical data sequences, a Meta-knowledge learner for learning node and edge attributes, Meta-GAT for capturing spatial correlations from meta-knowledge, and Meta-RNN for capturing temporal correlations from meta-knowledge.

In 2022, Chen et al. [9]argued that previous models only consider limited and static external factors. They proposed AARGNN, an attentive attributed recurrent GNN that considers multiple static and dynamic factors during the forecasting process. AARGNN takes into account road network topology, driving distance, points of interest, road physical properties, and incident data as link-level features; traffic state data as node-level features; and weather and event information as graph-level features. An attention mechanism is used to identify the contribution of each factor to the prediction tasks. AARGNN demonstrated better accuracy compared to state-of-the-art models such as DCRNN [17], TGC-LSTM [7], and GMAN [**?**]."

B. Deep learning in Traffic State Estimation(TSE)

Although focused on distinct objectives, TSE and traffic forecasting share many similar or even identical algorithms (such as STGCN and DCRNN). All methods in TSE studies can be generally categorized into three domains: model-driven approaches, data-driven approaches, and hybrid/physical-informed deep learning (PIDL) approaches. In model-driven methods, researchers utilize traffic flow models (usually based on fluid dynamics) to predict and estimate traffic states. Commonly used models include the Lightwhil-Whitham-Richards model [12] [13], Aw-Rascle-Zhang model [14] [15] . However, model-based methods are notorious for their complexity, tedious parameter calibration, and inability to comprehensively capture features. Data-driven approaches utilize large volumes of real-world traffic data for analysis and modeling. While data-driven methods overcome the drawbacks of model-based methods, they still require high-quality and sufficient data, which is rare in practical scenarios. Due to data sparsity, data-driven methods often produce results that are difficult to explain with domain knowledge, and sometimes resulting in unnatural outcomes like negative traffic flows. To address these issues, researchers have proposed physical-informed deep learning (PIDL) approaches, combining model-driven and data-driven methods. PIDL methods perform better by utilizing domain knowledge to improve accuracy. In 2021,Shi et al. [16]proposed a deep learning paradigm combining PIDL with a fundamental diagram (FD) learner, known as PIDL-FDL. They successfully applied PIDL-FDL to solve popular first-order and second-order traffic models, reconstructing the FD relation and modeling parameters outside FD terms. In 2022, Ji et al. [18] introduced the Spatio-Temporal Differential Equation Network (STDEN) to enhance the influence of physical laws in deep learning schemes. STDEN unifies the traffic potential energy field differential

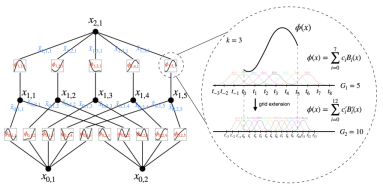

Fig. 1. KANs framework. Left: Notations of activations that flow through the network. Right: an activation function is parameterized as a B-spline, which allows switching between coarse-grained and fine-grained grids. [19]

equation with neural networks and was tested on real-world datasets, demonstrating significant superiority in accuracy over contemporary baseline methods.

C. Kolmogorov-Arnold Networks(KAN)

KAN is a recent study introduced by Liu et al. [18], presenting a novel neural network architecture designed to be compared with traditional MLP frameworks. Mathematically, KAN is based on the Kolmogorov-Arnold Representation Theorem (KART), which claims that all functions can be approximated by the sum of a finite number of univariate functions. Based on KART, Liu et al. proposed a neural network framework that substitutes linear weights with spline-based univariate functions along the edges of the network. These functions are specifically structured as learnable activation functions. As shown in Figure 1, the authors believe KANs, especially due to its structure and parametrication, offer solutions to the limitations of accuracy and, particularly, the poor interpretability in MLPs while handling scientific computations. Liu et al. highlighted that the network can produce a parametric explanation of results based on its topological structure. Therefore, KANs allow users to interact with the network by manually adjusting each learnable function to tune or test the neural network. These two features combined can provide better interpretatbility.

Inspired by KANs, on June 19th, F. Zhang and X. Zhang [19] proposed a novel scheme for graph neural networks by incorporating the message-passing framework into KAN. The resulting GraphKAN showed higher accuracy compared to graph convolution networks, although the computation time was 25% more. Nevertheless, GraphKAN successfully demonstrated a new option for GNN frameworks in the transportation field.

Furthermore, the DL community subsequently introduced Temporal-KAN (TKAN) [20] and Time-series KAN [21], focusing on forecasting with time-series data. Reflecting on GNN methods in traffic forecasting and TSE, KAN is likely to inspire new transportation deep learning frameworks in the near future.

However, KAN is still controversial in many aspects. Researchers claim that KAN generally suffers from a larger loss than MLP-based neural networks. Others argue that the accuracy, at least at this point in time, is not satisfactory in variants such as GraphKAN and Time-series KAN and the reduction in sizes of the network is not impressive as promised. Therefore, it is safe to say that KAN has a long way to go before it can replace MLPs. But, since it has only been two months since the release of KANs, it is still too early to draw definitive conclusions about their potential.

III. Future research directions

My research direction is currently focused on deep learning in traffic forecasting and traffic state estimation. However, it will be under slight adjustments as I refine my approach. Since I am relatively new to this field, my research in the near future will mainly involve implementing previous works using Japan domestic transportation data. Meanwhile, I will stay updated with the latest advancements in deep learning.

My recent research interest lies in Kolmogorov-Arnold Neural Networks (KANs), which is a risky yet attractive option. I am fascinated by their potential to enhance interpretability. Although the interpretability of neural networks is usually domain-specific, I generally believe the type of interpretability offered by KANs meets universal needs in transportation science. As I finish this abstract, the KANs community has announced a new seemingly feasible variant for physical simulation. At this point in time, I think it is prudent to refrain from drawing any conclusions about KANs until I have thoroughly inspected their potential. So far, there is no concrete proof of KANs' capacity to replace MLPs, but I have not seen any criticism on the interpretatbility and efficiency under limited paramters offered. Thus, this is also the reason that I believe KANs remain promising for some topics under certain domain. For the next step, I will attempt to capture spatial-temporal dependencies with KANs, an area where little work exists.

Meanwhile, the research on existing models' generalizability and replicability require the participation of the experiments on the real-world data. As essential parts of practical implementations requiring rigid testification, these two types of research will be carefully conducted in long-term research during my master career. As for the research on the model's efficiency, I think it is the most challenging yet crucial topic for me. Requiring knowledge of the bottlenecks in both transportation science and deep learning, this research field needs a insightful and efficient plan to explore. And I am still carefully designing it.

References

[1] B. Shahsavari and P. Abbeel, "Short-term traffic forecasting: Modeling and learning spatio-temporal relations in transportation networks using graph neural networks," Univ. California at Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2015-243, 2015.

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

[3] L. Zhao et al., "T-GCN: A temporal graph convolutional network for traffic prediction," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[4] Y. Shin and Y. Yoon, "Incorporating dynamicity of transportation network with multi-weight traffic graph convolutional network for traffic forecasting," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 3, pp. 2082–2092, Mar. 2022.

[5] K. Elissa, “Title of paper if known,” unpublished.
[6] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, “Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 11, pp. 4883–4894, Nov. 2020.
[7] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, “Attention based spatial–temporal graph convolutional networks for traffic flow forecasting,” in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, no. 1, pp. 922–929.
[8] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, “Urban traffic prediction from spatio-temporal data using deep meta learning,” in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2019, pp. 1720–1730.
[9] L. Chen, W. Shao, M. Lv, W. Chen, Y. Zhang, and C. Yang, “AARGNN: An attentive attributed recurrent graph neural network for traffic flow prediction considering multiple dynamic factors,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 17201–17211, Oct. 2022.
[10] C. Zheng, X. Fan, C. Wang, and J. Qi, “GMAN: A graph multi- attention network for traffic prediction,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 1, pp. 1234–1241.
[11] L. Chen, W. Shao, M. Lv, W. Chen, Y. Zhang, and C. Yang, “AARGNN: An attentive attributed recurrent graph neural network for traffic flow prediction considering multiple dynamic factors,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 17201–17211, Oct. 2022.
[12] M. J. Lighthill and G. B. Whitham, ”On kinematic waves II. A theory of traffic flow on long crowded roads,” Proceeding of the Royal Society of London. Series A. Mathematical and Physical Sciences, vol 229, no. 1178, pp. 317-345, 1955.
[13] P. I. Richards, ”Shock waves on the highway,” Operation Research, vol. 5, no.1, pp. 42-51, 1956.
[14] A. Aw and M. Rascle, ”Resurrection of ”second-order” models for traffic flow,” SIAM Journal on Applied Mathematics, vol. 60, no. 3, pp. 915-938, 2000.
[15] H. M. Zhang, ”A non-equilibrium traffic model devoid of gas-like behavior,” Transportation Research Part B: Methodological, vol. 36, no. 3, pp. 275-290, 2002.
[16] R. Shi, Z. Mo, K. Huang, X. Di, and Q. Du, ”A physical-informed deep learning paradigm for traffic state and fundamental diagram estimation,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 11 688–11 698, Aug., 2022.
[17] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” 2017, arXiv:1707.01926
[18] J. Ji, J. Wang, Z. Jiang, J. JIang, H. Zhang, ”STDEN: Towards physics-guided neural networks for traffic flow prediction”
[19] Z. Liu, Y. Wang, S. Vaidya, F. Ruehe, J. Halverson,M. Soljaci ˇ c,T. Y. Hou, and M. Tegmark, “Kan: Kolmogorov-arnold net- works,” Preprint arXiv:2404.19756, 2024.
[20] F. Zhang, and X. Zhang ”GraphKAN: Enhancing Feature Extraction with Graph Kolmogorov Arnold Networks” arXiv:2406.13597v1, Preprint, 2024
[21] R. Genet, and H. Inzirillo, ”TKAN: Temporal Kolmogorov-Arnold Networks,” Preprint, arXiv: 2405.07344v2, 2024
[22] C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus, ”Kolmogorov-Arnold Networks(KANs) for time series analysis,” Preprint, arXiv: 2405.08790v1, 2024