Urban Traffic Prediction from Mobility Data Using Deep Learning

Zhidan Liu, Zhenjiang Li, Kaishun Wu, and Mo Li

Abstract

Traffic information is of great importance for urban cities, and accurate prediction of urban traffics has been pursued for many years. Urban traffic prediction aims to exploit sophisticated models to capture hidden traffic characteristics from substantial historical mobility data and then makes use of trained models to predict traffic conditions in the future. Due to the powerful capabilities of representation learning and feature extraction, emerging deep learning becomes a potent alternative for such traffic modeling. In this article, we envision the potential and broard usage of deep learning in predictions of various traffic indicators, for example, traffic speed, traffic flow, and accident risk. In addition, we summarize and analyze some early attempts that have achieved notable performance. By discussing these existing advances, we propose two future research directions to improve the accuracy and efficiency of urban traffic prediction on a large scale.

Introduction

Comprehensive urban traffic information benefits urban citizens' daily life and improves urban transportation efficiency. Accurate predictions of such traffic information are of great importance for route planing, navigation, and other mobility services. Urban traffic prediction generally applies traffic models to analyze various historical and real-time traffic data to predict traffic conditions in the future. Traffic speed, traffic flow, and accident risk are representative indicators of traffic conditions, and tremendous efforts have been made to accurately predict such indicators as the traffic prediction targets in the past decades by leveraging types of mobility data and traffic models [1].

Traditionally, people are used to deploy various infrastructures, including loop detectors, traffic cameras, and radars, at some important road intersections to collect mobility data [2]. However, due to the high deployment and maintenance costs, it is prohibitive to widely adopt them on a city scale, which thus largely limits the coverage of traffic monitoring. Thanks to the popularity of ubiquitous sensing and Intelligent Transportation Systems (ITS) in recent years, we can gather unprecedented mobility data by exploiting a variety of mobile devices (e.g., smartphones and on-board GPS devices) and automatic fare collection (AFC) devices widely deployed by urban transit systems (e.g., subways, buses, and taxis). Such emerging big data substantially augment the data availability (coverage and fidelity) and also enriches data diversity, so that large-scale and reliable traffic predictions become viable.

To leverage such benefits, conventional methods utilize statistical models or machine learning models to predict traffic flows. They rely on human-crafted features to unveil and capture underlying traffic characteristics and further take instant traffic condition measurements as input, together with models built on the obtained features, to predict future traffic conditions. However, traffic flows can be influenced by various factors in practice, for example, transport regulations, weather conditions, and so on. These manually selected features have been shown to be inadequate to comprehensively describe traffic characteristics and thus cannot achieve accurate predictions [2].

Recently, unprecedented data availability and the ability to rapidly process these data together make possible the immense development of deep learning theory [3]. Deep learning has drawn much attention due to its remarkable capability to automatically extract features from large-scale raw data, and has already been successfully applied in various domains, for example, computer vision and speech recognition.

Compared to classic machine learning models, for example, SVM and ANN, which only have a shallow architecture to capture features, deep learning models inversely use multi-layer (i.e., "deep") architecture to discover intricate structures and complex patterns, where different layers capture features from different perspectives and finally together form a multi-level abstraction.

In view of the powerful capabilities of deep learning, we envision the potential and broad usage and impact of its integration with rich mobility data in future urban traffic prediction. In this article, we introduce the basic components involved in the procedure of urban traffic prediction, including the types of input mobility data, traffic modeling, and various target traffic indicators, for example, traffic speed, traffic flow, and accident risk. We investigate the possible approaches of applying deep learning to various kinds of traffic predictions, and meanwhile discuss those early attempts that have already exploited deep learning for accurate predictions of various traffic indicators. Based on discussing these existing advances, we analyze the inherent match between deep learning and data-driven traffic pre-

Zhidan Liu and Kaishun Wu (corresponding author) are with the College of Computer Science and Software Engineering, Shenzhen University; Zhenjiang Li is with the Department of Computer Science, City University of Hong Kong; Mo Li is with the School of Computer Science and Engineering, Nanyang Technological University.

Digital Object Identifier:
10.1109/MNET.2018.1700411

0890-8044/18/$25.00 © 2018 IEEE

diction. Moreover, we also point out two potential research directions, i.e., joint optimization of multi-source data and traffic modeling, and parallel computing promoted deep learning to accelerate traffic predictions, for future explorations. To the best of our knowledge, this is the first article that examines and summarizes deep learning based urban traffic predictions, and we believe this work could inspire a variety of follow-up work in this area.

The rest of this article is organized as follows. First we introduce the concepts involved in urban traffic prediction. Then we discuss the potential of deep learning in traffic prediction and analyze existing attempts. Next we discuss possible directions to improve the accuracy and efficiency of large-scale traffic prediction. Finally, we conclude this article.

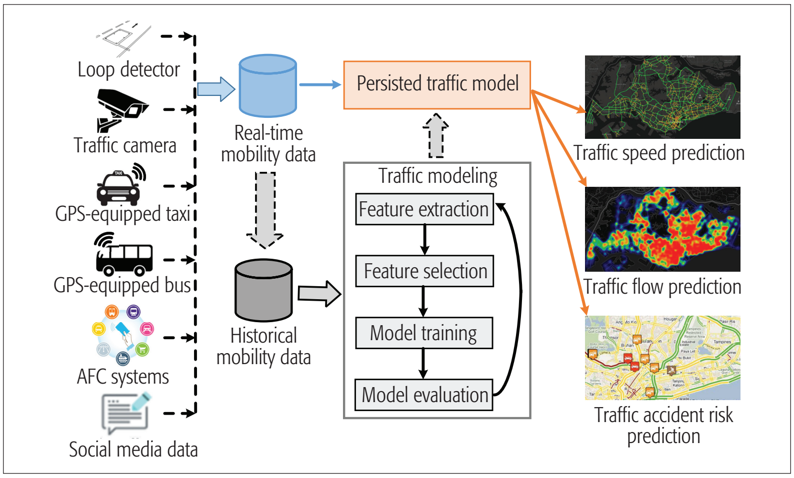

FIGURE 1. The basic components of urban traffic prediction.

Concepts of Urban Traffic Prediction

Urban traffic prediction concerns the prediction of traffic conditions made from a few seconds to a few hours into the future based on current and historical traffic information [1]. Many research efforts have been made to accurately model traffic indicators such as traffic speed, traffic flow, and accident risk, and produce anticipated traffic conditions. Figure 1 demonstrates the high-level procedure of urban traffic prediction, including mobility data collection, advanced traffic modeling, and targets of traffic predictions.

Mobility Data Collection

The mobility data involved in traffic predictions can be classified into the following categories.

Traffic Data from Infrastructures: Many infrastructure devices, e.g., loop detectors and traffic cameras, have been deployed in cities to continuously collect traffic data. The loop detectors are buried under traffic lanes of some important roads, and can detect vehicles passing by. Such measurements are used to calculate the traveling speed of each individual vehicle and also count the total number of vehicles passing by (i.e., traffic flow) within a period. Similarly, cameras are placed above road intersections and used to capture images of vehicles passing by. Based on computer vision techniques, traveling speeds of vehicles and traffic flows can also be derived.

Trajectory Data from Vehicles: In urban cities, a large number of public vehicles (e.g., taxis and buses) have been equipped with GPS devices, and thus can periodically report their status, including current location, traveling speed, direction, and so on Those reports indicate the trajectories of vehicles that contain traffic condition measurements of the roads.

AFC Records from Transit Systems: Modern public transportation networks rely heavily on AFC devices to automatically collect transit fees from bus and subway passengers, who need to tap their smartcards to AFC readers when they get on and off buses or subways. Thus, AFC systems record the boarding/alighting (bus or subway) stations/time of passengers, and all such records can be used to construct a trip origin-destination (OD) matrix that reveals mobility flows.

Other Data Sources: There are other data sources useful for traffic predictions. For example, accident reports, which contain location, severity, and event of each accident, provide helpful information to assess potential accident risk of each location within a city. Social networking services can treat humans as sensors to probe the dynamics of a city, and thus social media data can help infer traffic anomalies (e.g., accidents) as well. Cellphone data indicate users' movements within a city at cell-tower levels, and provide hints for inferences of traffic conditions. In addition, sensing data from crowdsourcing systems also serve as an important data source for traffic prediction. All such data measure urban traffic from a complimentary perspective.

Advanced Traffic Modeling

Urban traffic is complicated and usually non-linear, and thus some advanced traffic models are preferred, for example, statistical models or machine learning models, to capture the hidden traffic characteristics from mobility data and then facilitate the predictions based on input of real-time data.

As shown in Fig. 1, advanced traffic modeling is an iterative process that consists of several phases. To construct a traffic model, we first need to extract some desired values (i.e., *features*) from the raw mobility data. Such a set of features are correlated with the target traffic conditions. Taking the traffic condition c_i of a road segment s_i as an example, c_i is not only influenced by traffic conditions of s_i's neighboring road segments in the spatial dimension, but also impacted by time of the day (e.g., peak hours and non-peak hours) and day of the week (e.g., workday and weekend) in the temporal dimension. Those spatial-temporal factors together determine the evolution of c_i and play an important role in accurately predicting its future status. After the feature extraction phase, a small set of the most relevant features are further selected based on some criteria, for example, information entropy, to simplify the modeling and enhance the generalization capability of a model. After constructing the traffic model only using the most informative and non-redundant features, we can tune the parameters through massive training data and evaluate the derived model with testing data. The whole process of traffic modeling can be repeated until target prediction performances (e.g., accuracy) are achieved. The persisted traffic model is the one that encodes the traffic char-

Category	Involved data sources	Desired output
Traffic speed prediction	Infrastructures, GPS-equipped vehicles	Average traffic speed (or congestion level)
Traffic flow prediction	Infrastructures, AFC systems	Total number of objects passing through a road/region
Traffic accident risk prediction	Infrastructures, AFC systems, social media data, historical accident reports	Accident risk probability for each road/region

TABLE 1. A summary of different urban traffic predictions.

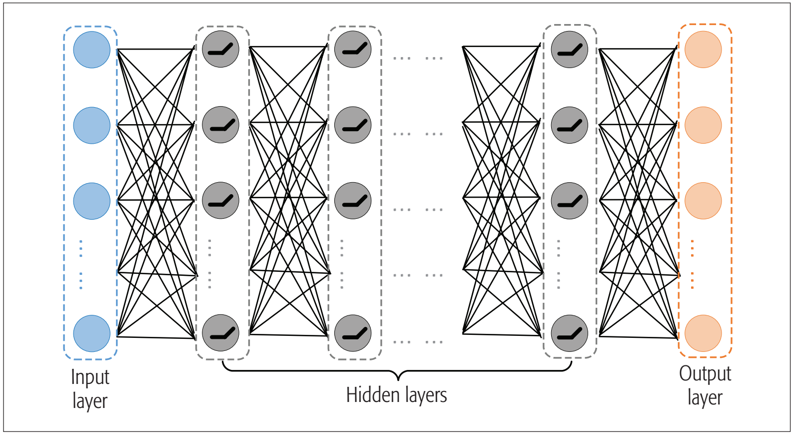

FIGURE 2. A deep neural network with fully-connected layers. It contains an input layer, an output layer, and many hidden layers. Each hidden layer contains a number of units that use an activation function (i.e., ReLU) to calculate the state based on units from the immediately previous layer.

acteristics and can be used for traffic prediction given the real-time input mobility data.

Existing works mainly rely on models like ARIMA, ANN, and SVM to capture the complex traffic [1]. When building such traffic models, feature extraction and selection are significantly important as they will determine the final performance of a traffic model. These procedures, however, are heavily dependent on man-crafted feature engineering, which calls for rich experiences and expert knowledge.

Targets of Traffic Prediction

According to the prediction targets of interest, urban traffic predictions can be further subdivided into *traffic speed prediction*, *traffic flow prediction*, and *traffic accident risk prediction*. Table 1 summarizes these types of traffic predictions, as well as their involved data sources and desired output.

Traffic Speed Prediction: Traffic speed is a widely adopted indicator to measure the traffic condition of one road segment, which is generally calculated as the average traveling speed of all sampling vehicles on a given road segment. Existing works derive such vehicular speed measurements either indirectly from data collected by loop detectors and cameras [2] or directly from GPS-equipped vehicles [4]. They construct a traffic speed model from historical data by adopting classic machine learning models, and take real-time sampling speeds as the input to predict future traffic speeds. The predicted traffic speeds can be translated to certain congestion levels (e.g., slow, normal, and fast) according to some mapping rules.

Traffic Flow Prediction: In general, traffic flow is defined as the total number of target objects (i.e., vehicles or humans) that pass through an area during a period. The area can be a road segment or a region in the city. Different from traditional works that hold many assumptions on human mobility, more recent approaches model and predict the traffic flow based on the realistic human mobility data collected from infrastructures and AFC systems. Traffic flows reveal the movements of crowds and potentially determine the traffic distributions [5].

Traffic Accident Risk Prediction: Traffic accidents, although rare, have serious impacts on urban traffic. Therefore, it is necessary to assess traffic accident risks for each specific road and region, which can be measured as likelihoods, meaning how likely is it that traffic accidents might occur on a road/region. Recent practices mainly associate accident risks with current traffic conditions and human mobility, and thus they develop models to mine relations between mobility data and historical accident reports for traffic accident risk prediction [6].

Deep Learning Based Traffic Prediction

A Primer on Deep Learning

Although there exist various forms of deep learning models, they share a common architecture as shown in Fig. 2, which contains an input layer, an output layer, and from several to more than a thousand hidden layers in between. Raw data initialize the values of the input layer while the output layer emits the desired inferences. All hidden layers are responsible for transforming states of the input layer into the expected inferences of the output layer by capturing the high-level abstractions. Each layer in the network contains a number of units, and the sizes could vary among different layers. Links exist between units of any two neighboring layers and each link is associated with a *weight*. Every unit has an *activation function* that determines how to calculate its own state based on units from the immediately previous layer and in turn exposes its state to the next layer. One of the most popular activation functions recently is the rectified linear unit (*ReLU*), which is a half-wave rectifier $f(x) = \max(x, 0)$.

Next we will introduce several popular models that have already been well exploited [3].

Convolutional Neural Network (CNN): The CNN model is primarily designed to process 2-dimensional data, for example, images. As shown in Fig. 3a, a CNN model is composed of an input layer and an output layer, as well as multiple hidden layers, which could be the convolutional, pooling, or fully connected layers. The convolutional layers adopt convolutional filters, which apply certain transformations on the input data to capture their properties. Next follow pooling layers that combine the output of unit clusters at a previous layer into a single unit in the next layer by employing the max or min filter. A pooling layer learns more abstract representations of the data, and meanwhile acts as a form of dimensionality reduction to simplify the whole model. A fully connected layer is used to complete the inference.

Recurrent Neural Network (RNN): The RNN model is mainly used for tasks that are involved with sequential inputs, for example, speech and language, due to its "memory" design in the form of a loop as shown in Fig. 3b. A loop allows information to be passed from one step to the next (Fig. 3b left). RNNs process an input sequence one element at a time, maintaining output results in the hidden units that implicitly persist information about the history of all past elements. When unfolding the loop, an RNN can be viewed as a stack of separate neural networks with some parameters of each network fed from the previous one (Fig. 3b middle). Such parameters act as the memory of RNN models. Inside the repeating neural networks of an RNN (Fig. 3b right), the input element x_t at time step t is concatenated with the output y_{t-1} of previous time step and then are together fed into an activation function (e.g., `tanh`) to derive output y_t of the current time step. Such an architecture allows RNNs to capture temporal dynamics, but practices show that RNNs cannot support long-term dependency [3]. Thus, an improved RNN called a Long Short Term Memory network (*LSTM*) is proposed, which uses special hidden units (i.e., memory cells) to remember inputs for a long time. LSTM models are able to learn long sequences and automatically determine the optimal time lags for prediction.

Stacked Autoencoder (SAE): An autoencoder is a three-layer neural network with an input layer, an output layer, and a hidden layer, as shown in the left part of Fig. 3c. The target output is intentionally set as the input of the model, and thus the hidden layer aims to learn the representations of the input data, which can be viewed as a dimensionality reduction or encoding of input data. Due to this function, the hidden layer of an autoencoder is also called the *feature layer*. The SAE model links such feature layers in a stacked fashion to create higher-level abstractions of input data, which forms a deep architecture, as shown in the right part of Fig. 3c. One of the most popular variant autoencoders is a denoising autoencoder, which takes deliberately corrupted samples as the inputs while is forced to recover the original uncorrupted data. When stacking multiple denoising autoencoders, we thus derive a variant of SAE called a stacked denoising autoencoder (SdAE). Compared to the SAE model, SdAE is able to discover relatively stable features, which makes it robust against noisy inputs and thus perform much better.

There are other deep learning models, such as the Restricted Boltzmann Machine (RBM) and Deep Belief Network (DBN). Table 2 presents a summary of the above models and their early adoptions in traffic predictions to be discussed later.

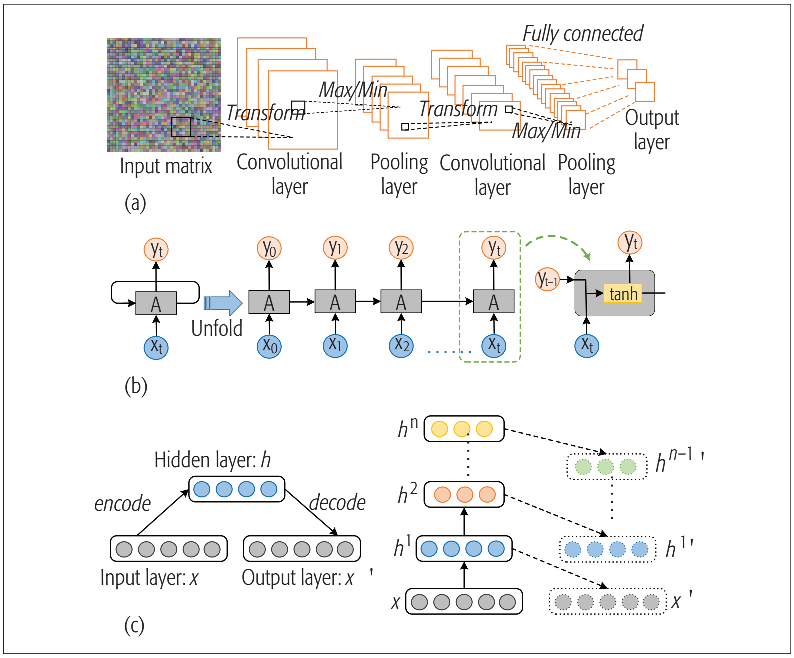

FIGURE 3. The architectures of various deep learning models: a) typical architecture of CNN model; b) typical architecture of RNN model; c) typical architecture of autoencoder and SAE model.

Model	Application scenarios	Referred works
CNN	2-dimensional data (e.g., images, videos)	Speed prediction [7, 8]; flow prediction [5]
RNN	Sequential data (e.g., speech, language)	Speed prediction [8, 9]
LSTM	Long sequential data (e.g., speech, language)	Speed prediction [10]; flow prediction [11]
SAE (SdAE)	Representation learning	Flow prediction [2]; accident risk prediction [12]
RBM DBN	Representation learning	Speed prediction [9, 13, 14]

TABLE 2. Summary of different deep learning models.

Traffic Speed Prediction

The traffic speed of one road segment is influenced by many factors in both the temporal and spatial dimensions, for example, time of day and traffic conditions of neighboring road segments.

We have two ways to apply deep learning to extract such temporal-spatial features at either the individual road segment level or the whole road network level. On one hand, we can mine detailed traffic features for each individual road segment and then make use of the derived features to construct classic machine learning models for traffic speed prediction. An early attempt follows this idea and has proposed DeepSense [14], which exploits the RBM model to extract high-level features for building an SVM model to predict traffic speed. Specifically, for each target road segment *s*, a number of correlated segments are selected and their traffic speeds, certain states, time intervals, and geographical distances between them and *s* are fed into an RBM model to automatically discover helpful features for constructing the SVM model. Substantial taxi traces are used to train DeepSense, and the experiment results show that DeepSense achieves higher prediction accuracy than its competitors.

The following works have explored the possible applications of other deep learning models in this direction, e.g the DBN model [13], the hybrid model of RBM and RNN [9], and the LSTM model [10]. Those works, however, primarily apply deep learning on temporal speed sequences of individual road segments for traffic prediction at a small network region.

On the other hand, we can consider traffic speed prediction at the road network scale and long time range so that we have essentially transformed the temporal-spatial traffic speeds into one 2-dimensional data matrix, which is the favorable input of CNN models. CNN is good at

It is highly expected that a deep learning job can be partitioned into a series of tasks running at different machines in parallel. However, how to achieve the best modeling performances while maintaining minimum costs on both communications and computations is quite difficult.

capturing spatial features of 2-dimensional data and has been widely applied in image recognition tasks with prominent performances achieved. Inspired by such successes, Ma *et al*. [7] propose a CNN based method to learn urban traffic as images. They convert road network traffic dynamics into an image that represents the temporal and spatial relations of traffic as a matrix. Each row of the matrix describes the evolution of one road segment along the time, while each column describes traffic conditions of the whole road network at a specific time step. They apply CNN to such images to extract network-scale features and use those features to build a fully connected neural network for network-wide traffic speed prediction. Experiments show that the CNN based method indeed has remarkable capability to process 2-dimensional data and outperforms the compared methods building on either conventional models (e.g., ANN) or other deep learning models (e.g., RNN and LSTM). By considering prediction errors, Wang *et al*. [8] further improve conventional CNN models with an additional error-feedback recurrent layer, which takes the output of CNN as the input and compensates the prediction errors using predicting results of previous periods.

Traffic Flow Prediction

Similar with traffic speed, traffic flow in a specific area is also affected by temporal and spatial factors. Different from traffic speed, traffic flow should be considered on a large scale since human mobility usually covers a large area. Therefore, we divide the road network or the whole city into grids and place traffic flows into such a 2-dimensional gridded space to form instant traffic flow snapshots. Some deep learning, especially CNN, models could be used to discover latent traffic flow features from such snapshots to build the flow predictor. A notable attempt is made by Zhang *et al*. [5], where they consider predictions of traffic inflow/outflow in each region of a city by exploiting historical mobility data, weather conditions and holiday events. In this work, traffic inflow/outflow can be measured as the number of pedestrians, the number of vehicles driving by near roads, and any other measurements related to human mobility. To capture the complex temporal-spatial dependencies, the authors transform historical and current inflow/outflow data into image-like matrices, and separate them into three groups, denoting *recent* time, *near* history, and *distant* history. Each group is applied with a CNN model retaining only convolution layers to hierarchically capture spatial structure information. A residual unit sequence is used to allow a CNN model to be appended with many layers. In addition, external factors like weather conditions and holiday events are considered through a fully connected neural network. The four components individually predict traffic inflow/outflow, and these predictions are then aggregatively fused to derive the final result. In addition to the high-level flow statistics, Song *et al*. has proposed *DeepTransport* [11], which exploits LSTM models to predict and simulate an individual's future movements and transportation modes. Lv *et al*. [2] exploit SAE models to predict traffic flows on specific road segments. Such detailed information will better facilitate the management and planning of urban traffic.

Traffic Accident Risk Prediction

There are many factors related to traffic accidents, for example, traffic congestion, driver behavior, and road and weather conditions, and thus accident risk prediction is much more challenging. In general, traffic accident risks are highly correlated with human mobility, land usage, and historical traffic accidents. Thus, we can divide a city into grids and assign traffic flow and historical traffic accident data into these grids to form a mobility matrix and an accident matrix. Taking these matrices as inputs, certain deep learning models could be used to extract complex features for building a traffic accident risk predictor using a traditional machine learning model.

The only attempt we found is made by Chen *et al*. [12]. Their proposed method divides a city into regions and the time of day into intervals. For each time interval t and each region r, it calculates risk level $g_{r,t}$ from historical accident data, and average human mobility density $d_{r,t}$ from historical GPS records. The derived data form two kinds of matrices and are fed into SdAE to extract robust and stable features to construct a logistic regression model. Given real-time human mobility data, the method outputs a risk assessment map that can be used to provide early warning for people of possible traffic accidents. This method, however, only considers human mobility but does not take other factors into account, e.g., weather and land usage, for a comprehensive accident risk prediction.

Discussion and Future Directions

Why Deep Learning Fits Traffic Prediction

Urban traffic can be influenced by many factors, such as transport regulations, road conditions, whether conditions, stochastic events, land usage, and so on, which together make traffic patterns extremely complex. The hand-crafted features from prior statistical or machine learning models are essentially a series of hypotheses proposed to approximate the unknown relation about how such factors impact traffic status. Due to the inherent complexity and hardness of such relations, however, the manually selected features have been shown to be inadequate to comprehensively describe traffic characteristics and thus cannot achieve accurate prediction results.

Thanks to the *deep* architecture of multiple processing layers, deep learning is capable of automatically discovering the most representative features from a massive amount of mobility data, which is impossible for prior methods with shallow architectures. By inspecting pioneering studies, we highlight the general workflow to apply deep learning for traffic predictions as shown in Fig. 4. Instead of directly inputting mobility data into classic machine learning models, the raw data are first fed into deep learning models to learn abstractions by many hidden layers. In general, low-level abstractions are first extracted from the input data and in turn fed to following layers to form higher-level abstractions. Finally, such a hierarchy of abstractions automatically selects some high-level features that are simultaneously sensitive to subtle details, e.g., different times of day, and insensitive to irrelevant variations, e.g., the types of passed vehicles on roads. Building on such features, the derived traffic models will be more informative, stable, and robust, and thus they can achieve much better prediction performance. In essence, deep learning can be viewed as an excellent feature extractor, which avoids burdensome feature engineering while automatically learning good features using a general-purpose learning procedure.

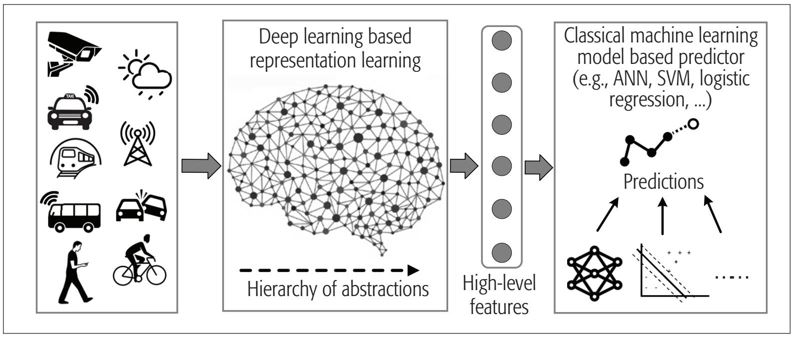

FIGURE 4. Deep learning models hierarchically learn representations of mobility data and output high-level features to support classical machine learning models for better traffic predictions.

Future Research Directions

In this article, we propose two potential and crucial research directions for this research topic.

Joint Optimization of Multi-Source Data and Traffic Modeling: As introduced earlier, various types of mobility data can serve as deep learning's input, and multiple traffic condition indicators need to be predicted as well. Of course, each indicator may not be reliably inferred from any individual single-source mobility data, while how to select the most appropriate mobility data sources to satisfy each indicator's prediction requirement is so far unknown yet. In addition, even if such a selection could be eventually figured out, how to further determine suitable deep learning model details, e.g., the number of models, model types, layers, and so on, to fuse these mobility data sources and meanwhile link the input and output is non-trivial. Thus, applying deep learning in traffic prediction encounters a joint optimization of data modality, model structure, and fusion methodology.

One possible solution we propose is to exploit all available mobility data sources for prediction based on a multi-model strategy. For each data source, we apply deep learning to capture respective features and then produce one prediction. All predictions from multiple traffic models can be carefully fused to obtain the comprehensive result.

Such a multi-model based traffic prediction is feasible and attractive, where we can exploit the ensemble learning theory to integrate those models and their predictions for a better result. In practice, we may apply different deep learning models for different mobility data sources to obtain diverse traffic models, and adopt the weighted average strategy to compute the final prediction. The weights of different models are determined through a training procedure. We thus omit the data selection issue and design the deep learning model for each data source independently.

Parallel Computing Promoted Deep Learning to Accelerate Traffic Predictions: To fully extract abstractions from mobility data, deep learning models are usually designed to contain hundreds to thousands of layers, and thus numerous parameters need to be tuned. Conventional computing systems are thus inadequate to such computationally intensive tasks. It becomes even more serious when multi-source mobility data are involved, where the storage and computation overheads will significantly increase. Therefore, scalable and efficient parallel computing systems (e.g., computer clusters) are preferable to store such big data and accelerate data processing, traffic modeling and the prediction.

It is attractive yet challenging to handle deep learning based traffic predictions in a distributed manner. It is highly expected that a deep learning job can be partitioned into a series of tasks running at different machines in parallel. However, how to achieve the best modeling performances while maintaining minimum costs on both communications and computations is quite difficult. First, how to parallelize this modeling job is unclear, and even if possible, how to merge pieces of parameters learned at different machines into the complete final model for quick prediction remains to be explored. In addition, since mobility data are used as training data for traffic modeling, the wise placements of those data among machines are of great importance to reduce unnecessary data exchanges (i.e., training data and intermediate parameters) between machines. To address these challenges, we can build our deep learning models by exploiting the parameter server (PS) architecture [15] to manage and synchronize the model parameters among machines. In the PS architecture, server nodes maintain the latest model parameters and make them available to worker nodes, while worker nodes update the model parameters using the assigned training data. Also, since regions nearby are correlated in traffic flows, we can place the mobility data among machines according to their geographical information to significantly reduce data transfer among machines when training the deep learning models. In practice, we can embed more domain knowledge of transportation into our model and system design to further improve the accuracy and efficiency of large scale traffic predictions.

Conclusions

In this article, we envision the potential of rich mobility data and deep learning on urban traffic prediction, and discuss some pioneering attempts. Deep learning will advance traffic predictions through powerful representation learning and has shown initial successes. By discussing the existing advances, we proposed two research directions to further improve the accuracy and efficiency of traffic prediction on a large scale.

Acknowledgment

This research was supported by the NSF of SZU (No. 2018061); China NSFC Grant (No. 61601308 and No. 61472259); Guangdong NSF Grant (No. 2017A030312008); Guangdong Provincial Science and Technology Development Special Foundation (No. 2017A010101033); Shenzhen Science and Technology Foundation (No. JCYJ20170302140946299 and No. JCYJ20170412110753954); Fok Ying-Tong Education Foundation for Young Teachers in the Higher Education Institutions of China (No. 161064); Guangdong Talent Project (No. 2014TQ01X238 and No. 2015TX01X111); GDUPS Grant (2015); the ECS grant from the Research Grants Council of Hong Kong (No. CityU 21203516); and a GRF grant from the Research Grants Council of Hong Kong (No. CityU 11217817); Singapore MOE Tier 2 grant MOE2016-T2-2-023; and NTU CoE grant M4081879.

References

[1] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-Term Traffic Forecasting: Where We Are and Where We're Going," *Transportation Research Part C: Emerging Technologies*, vol. 43, part 1, 2014, pp. 3–19.

[2] Y. Lv *et al.*, "Traffic Flow Prediction with Big Data: A Deep Learning Approach," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 2, 2015, pp. 865–73.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, 2015, pp. 436–44.

[4] Z. Liu *et al.*, "Mining Road Network Correlation for Traffic Estimation via Compressive Sensing," *IEEE Trans. Intelligent Transportation Systems*, vol. 17, no. 7, 2016, pp. 1880–93.

[5] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," *Proc. of AAAI*, 2017, pp. 1655–61.

[6] J. Sun *et al.*, "A Dynamic Bayesian Network Model for Real-Time Crash Prediction using Traffic Speed Conditions Data," *Transportation Research Part C: Emerging Technologies*, vol. 54, 2015, pp. 176–86.

[7] X. Ma *et al.*, "Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction," *Sensors*, vol. 17, no. 4, 2017, Article No. 818.

[8] J. Wang *et al.*, "Traffic Speed Prediction and Congestion Source Exploration: A Deep Learning Method," *Proc. IEEE ICDM*, 2016, pp. 499–508.

[9] X. Mao *et al.*, "Large-Scale Transportation Network Congestion Evolution Prediction using Deep Learning Theory," *PloS one*, vol. 10, no. 3, 2015.

[10] R. Yu *et al.*, "Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting," *Proc. SIAM ICDM*, 2017, pp. 777–85.

[11] X. Song, H. Kanasugi, and R. Shibasaki, "DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level," *Proc. IJCAI*, 2016, pp. 2618–24.

[12] Q. Chen *et al.*, "Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference," *Proc. AAAI*, 2016, pp. 338–44.

[13] Y. Jia, J. Wu, and Y. Du, "Traffic Speed Prediction using Deep Learning Method," *Proc. IEEE ITSC*, 2016, pp. 1217–22.

[14] X. Niu, Y. Zhu, and X. Zhang, "DeepSense: A Novel Learning Mechanism for Traffic Prediction with Taxi GPS Traces," *Proc. IEEE GLOBECOM*, 2014, pp. 2745–50.

[15] M. Li *et al.*, "Scaling Distributed Machine Learning with the Parameter Server," *Proc. USENIX OSDI*, 2014, pp. 583–98.

Biographies

Zhidan Liu (liuzhidan@szu.edu.cn) received the B.E. degree in computer science and technology from Northeastern University, China, in 2009, and the Ph.D. degree in computer science and technology from Zhejiang University, China, in 2014. He is currently an assistant professor at Shenzhen University (SZU), China. His research interests include distributed sensing and mobile computing, big data analytics, and urban computing.

Zhenjiang Li (zhenjiang.li@cityu.edu.hk) received the B.E. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2007, the M.Phil. degree in electronic and computer engineering and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2009 and 2012, respectively. He is currently an assistant professor at City University of Hong Kong, Hong Kong. His research interests include wearable and mobile sensing, deep learning and data mining, distributed and edge computing.

Kaishun Wu (wu@szu.edu.cn) received his Ph.D. degree in computer science and engineering from HKUST, Hong Kong, in 2011. After that, he worked as a research assistant professor at HKUST. In 2013, he joined SZU as a professor. He is the inventor of six U.S. and 43 Chinese pending patents (13 are issued). He received the 2014 IEEE ComSoc Asia-Pacic Outstanding Young Researcher Award.

Mo Li (limo@ntu.edu.sg) received the B.S. degree in computer science and technology from Tsinghua University, China, in 2004 and the Ph.D. degree in computer science and engineering from HKUST, Hong Kong, in 2009. He is currently an associate professor at Nanyang Technological University, Singapore. His research interests include networked and distributed sensing, wireless and mobile, cyber-physical systems, smart city, and urban computing.