SBAG: A Hybrid Deep Learning Model for Large Scale Traffic Speed Prediction

Adnan Riaz[1]
School of Computer Science and Technology
Dalian University of Technology
Dalian, China

Mehak Khan[3]
Department of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

Muhammad Nabeel[2]
School of Software Engineering
South China University of Technology
Guangzhou, China

Huma Jamil[4]
Department of Computer Science and Technology
PMAS Arid Agriculture University
Rawalpindi, Pakistan

***Abstract*—Intelligent Transportation System (ITS) is the fundamental requirement to an intelligent transport system. The proposed hybrid model Stacked Bidirectional LSTM and Attention-based GRU (SBAG) is used for predicting the large scale traffic speed. To capture bidirectional temporal dependencies and spatial features, BDLSTM and attention-based GRU are exploited. It is the first time in traffic speed prediction that bidirectional LSTM and attention-based GRU are exploited as a building block of network architecture to measure the backward dependencies of a network. We have also examined the behaviour of the attention layer in our proposed model. We compared the proposed model with state-of-the-art models e.g. Fully Convolutional Network, Gated Recurrent Unit, Long -short term Memory, Bidirectional Long-short term Memory and achieved superior performance in large scale traffic speed prediction.**

***Keywords*—Attention mechanism; large scale traffic prediction; Gated Recurrent Unit (GRU); Bidirectional Long-short term Memory (BiLSTM); Intelligent Transportation System (ITS)**

I. INTRODUCTION

The performance of Intelligent Transportation System (ITS) applications principally depends on the quality of traffic data. Lately, with the increment of both traffic volume and traffic data, traffic speed perdition has become very important in the ITS. In the past decades, short term traffic prediction is under the eyes of researchers. Many researchers have proposed different approaches and netwroks in the the past decades which shows it has a long history and this issue is yet to resolve in case of accuracy about traffic speed predcition. To overcome and enhance the efficiency and accuracy of the traffic prediction, several approaches were proposed [1]. Many traffic amenities and applications are dependent on prediction accuracy.

Existing models coarsely divided into two categories, i.e. computational intelligence (CI) approaches and classical statistical methods indicated by previous literature [2][3][4]. The statistical methods were introduced at earlier stages when traffic data is limited and less complex but later with the advancement in traffic sensing technologies, arise of traffic data and computational power most of later work centers around computational intelligence approaches for traffic forecasting.

In general, regarding taking care of complex traffic forecasting problems [5], the computational approaches shattered the statistical methods like autoregressive integrated moving average (ARIMA) [6] in terms of capacity to catch nonlinear relationship and to deal with complex data. By the ascent of neural systems (NN) based methods, the full protentional of artificial intelligence was not subjugated in any case but many neural network-based models like Feed Forward Neural Network [7], Fuzzy Neural Network [8], Recurrent Neural Network [9], Gaussian Process [10] and hybrid Neural Network [11][12] are adopted for traffic forecasting problems. Recently, some hybrid architectures are proposed for traffic speed forecasting. Many factors influence on traffic forecasting, so single-component models are not suitable to complete the traffic prediction task. To make progress in the accuracy of traffic prediction, hybrid models are used in traffic speed prediction. In complex road network CapsNet [13] architecture proposed which replaced the max polling operation of CNN. To cope with the temporal evolution of traffic status, Recurrent Neural Networks (RNNs) models are specifically very appropriate because of the dynamic nature of transportation.

Structure of RNNs has internal memory with loops [14] that sequence data by maintaining a chain-like structure. However, RNNs are challenging to train during the backpropagating process because of the vanishing gradient problems, owing to the depth of the loop and chain-like structure. LSTMs addressed the aforementioned difficulties successfully. A spatial-temporal LSTM network, MapLSTM [15] for fine-grained traffic conditions. To predict traffic flow with big data, hybrid Deep Neural networks (DNN) [16] was proposed. Structural RNN (SRNN) proposed to deal with graph data of a road network [17].

LSTMs have the ability to deal with long term dependencies. In recent days, they have been gaining popularity in traffic forecasting because of a representative

deep learning method handling sequence data. In the domain of transportation, the capability of LSTM is not fully utilized yet. To predict large-scaled transportation traffic, it is becoming a vital and challenging topic. In most of the existing studies, network-wide prediction achieved only, when for N nodes, the same number of N models were trained for a traffic network [18] because they use traffic data along with a corridor or sensor location. However, learning complex spatial-temporal features of network-wide traffic should be explored by only one model.

In terms of dependency in prediction problem, the LSTM process the information in the forward direction, so LSTM only process forward dependency[5]. There is highly possible that some useful information may not efficiently filter or passed, so to consider the backward dependencies is very important. The other reason to consider the backward dependency is periodicity in traffic data, because traffic conditions have strong regularity and periodicity[19]. As per the literature review, a few studies utilized backward dependency. To cover this gap, bidirectional LSTMs (BDLSTMs) architecture is adopted as a network structure component because it can handle both forward and backward dependencies. In a traffic network, the impact of downstream and upstream speed on any location cannot be ignored while predicting the large-scale traffic speed. Along a corridor, future speed values of a location are affected by past speed values of upstream and downstream locations that only use forward dependencies in time series data, shown from previous studies[19][12][20]. In spatial-temporal data, the learned feature will be more inclusive with both backward and forward dependencies.

In this paper, we proposed a hybrid deep learning model known as stacked bidirectional LSTM with attention GRU (SBAG) neural network for large scale traffic speed prediction. Our model achieved better performance with the comparison of state-of-the-art methods. We consider the traffic forecasting to a large scale traffic network rather than several adjacent locations or specific location along a corridor. We proposed a hybrid model considering the backward dependencies using Bidirectional LSTM to improve feature learning. We examined the behaviour of attention mechanism to make improvements in the proposed model.

The remainder of the paper is described as Section II Methodology, Section III Performance Evaluation, Section IV Conclusion.

II. Methodology

The component of the proposed model SBAG is detailed explained in this section.

A. Input Data

In this study, the proposed and the compared model takes the large-scale speed data as input, to take network-wide influences into account. When traffic jam propagates, it not only affects the nearby location but also far away locations in a whole network. In traffic speed prediction, the input data use a sequence of speed values along n historical time step at one location [2][18][21], denoted by a vector,

$$Y_T = [Y_{T-n}, Y_{T-(n-1)}, \dots, Y_{T-2}, Y_{T-1}] \quad (1)$$

Suppose the traffic network consists of P locations, and we need to predict the traffic speeds at time T using n historical time frames (steps), the input can be characterized as a speed data matrix,

$$Y_T^P = \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^P \end{bmatrix} = \begin{bmatrix} Y_{T-n}^1 & Y_{T-n+1}^1 & \cdots & Y_{T-2}^1 & Y_{T-1}^1 \\ Y_{T-n}^2 & Y_{T-n+1}^2 & \ddots & Y_{T-2}^2 & Y_{T-1}^2 \\ \vdots & \vdots & & \vdots & \vdots \\ Y_{T-n}^P & Y_{T-n+1}^p & \cdots & Y_{T-2}^P & Y_{T-1}^P \end{bmatrix} \quad (2)$$

Where each element Y_T^P is the speed at 'p^{th}' location and 't^{th}' time steps. To signify temporal attributes of speed data and streamline the expression of the equation, vector $Y_T^P = [Y_{T-n}, Y_{T-(n-1)}, \dots, Y_{T-2}, Y_{T-1}]$ represents the speed matrix where each element signifies 'P' locations speed values.

B. Bidirectional Long Short Term Memory (BDLSTMs)

The idea of using Bidirectional LSTMs comes from bidirectional RNN. The bidirectional LSTMs join two hidden-layers to the same output layer. Bidirectional LSTMs showed superiority in different fields over unidirectional e.g. speech-recognition [22], phoneme-classification [23]. The structure of Bidirectional LSTMs is shown in Fig. 1.

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (3)$$

$$\hat{y}_t = \sigma_h(W_{hy}h_t + b_y) \quad (4)$$

$$\mathrm{f}_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$\mathrm{i}_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

$$\mathrm{o}_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (7)$$

$$\tilde{\mathrm{C}}_t = \tan h\,(W_C x_t + U_C h_{t-1} + b_C) \quad (8)$$

Where, W_f W_i, W_o, W_C U_f, U_i, U_o, U_C. b_f, b_i, b_o, and b_C, are the weight matrices and bias vector parameter which need to be learned during training. σ_g is the gate activation function and hyperbolic tangent function being $\tan h$.

$$\hat{y}_t = \sigma(\overleftarrow{h}, \overrightarrow{h}) \quad (9)$$

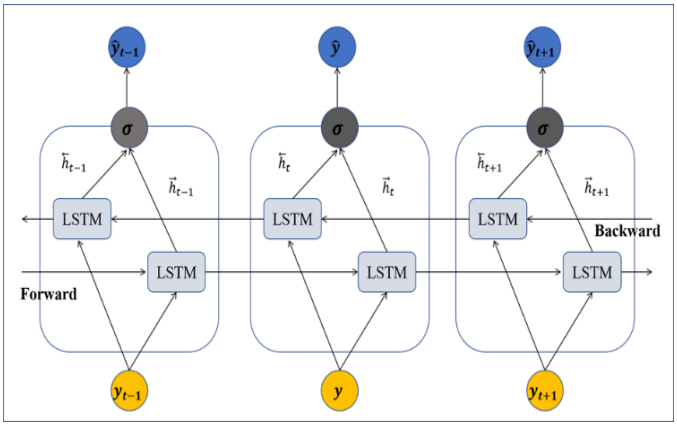

Fig. 1. Unfold Architecture of BDLSTM.

Where $\vec{h}$ and, $\overleftarrow{h}$ are the forward and backward layer output that iteratively calculated by using positive sequence inputs from time $T-n$ *to* $T-1$ and vice versa, backward and forward layers outputs are calculated by eq. 3-8, $\hat{Y}_T$ is an output vector that can be generated by Bidirectional LSTM, where each element is calculated from the eq. 9. Where σ is an average function used to join the two output sequences.

C. GRU

GRU is a well-known variant based on LSTM proposed by Cho et al. [24]. GRU is simpler than LSTM because it has fewer parameters than LSTM and its performance is significant in some tasks. It consists of forget gate and input gate. It combines forget-gate and input-gate to an update-gate. In Fig. 2 GRU block diagram is shown. The memory cell of a GRU has four components that allow cells to access and save information for a longer time period. GRU calculate the hidden states by following equations:

$$z_t = \sigma\left(W^{(z)}.[h_{t-1}, x_t]\right) \quad (10)$$

$$r_t = \sigma\left(W^{(r)}.[h_{t-1}, x_t]\right) \quad (11)$$

$$\tilde{h}_t = tanh(W.[r_t * h_{t-1}, x_t]) \quad (12)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (13)$$

In the above equations,z_t and r_t are update and reset gate. σ is an activation function. $\tilde{h}_t$ is candidate activation function, h_t is the actual activation-function of the proposed GRU at time t.

D. Attention Mechanism

The hidden unit of GRU consists of an update and reset gate that captures dependencies of different timescales. In time series sequence, spatial-temporal dependency has not contributed equally, so attention mechanism is adopted in this paper with GRU to solve this problem [25]. The magnitude of the weights $\widehat{\propto_i}$ learned by the network signifies the importance of hidden states. We compute $\hat{r}$ as combination of all $\widehat{h_i}$, after using the attention mechanism.

$$\hat{r} = \sum_{i=1}^{N=1} \widehat{\propto}_i \, \hat{h}_i \quad (14)$$

At each time step, the hidden-state vector $\widehat{h_i}$ is the input of the attention layer. For this time step the attention weight $\widehat{w_i}$ can be calculated as

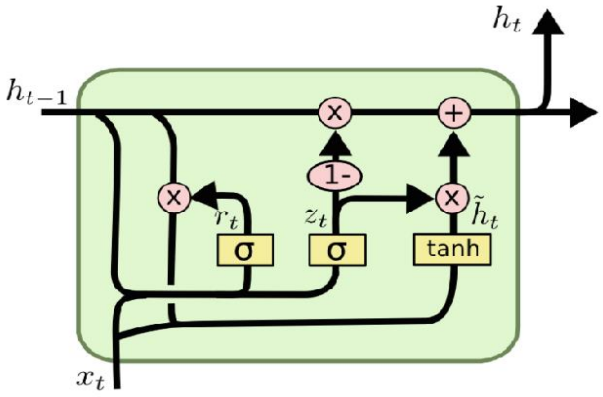

Fig. 2. A typical GRU Block Diagram.

$$\widehat{w}_i = \tan h\left(\widehat{h}_i\right) \quad (15)$$

$$\hat{b} = a^T m_i + b \quad (16)$$

$$\widehat{\propto_i} = \frac{\exp(\hat{b})}{\sum_k \exp(\hat{b})} \quad (17)$$

Where parameters of attention layer are a and b. At i_{th} time-step the attention layer output is formulated as:

$$\hat{r}_i = \widehat{\propto}_i \, \hat{h}_i \quad (18)$$

E. BDLSTM-GRU Module (with Attention Mechanism) for Spatial-Temporal Correlation Features Learning

Existing studies demonstrated that LSTMs work effectively in sequence tasks. BDLSTMS has the power to process data in both ways backward and forward direction so we adopted BDLTSM as the first layer in our proposed model to capture spatial-temporal information while feeding input to the model during the feature learning process. The top layer of the model only required learned feature when predicting future speed values. We used GRU as the last layer of the architecture the output of the BDLSTM is fed into GRU as input We also utilized attention mechanism to enhance the capability of GRU to process large scale traffic speed prediction.

In this paper, we proposed a novel hybrid model stacked Bidirectional LSTM and attention GRU (SBAG) for the large scale traffic speed prediction. The proposed model takes spatial input and predicts traffic speed value for the next time step. Fig. 3 illustrates the architecture of the proposed model.

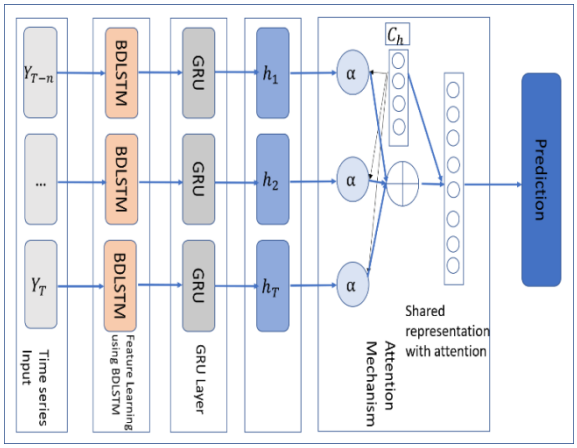

Fig. 3. Block Diagram of Proposed Model.

III. Performance Evaluation

A. Dataset Description

In this study, we used a publicly available dataset knows as a loop detector used by authors [26]. The dataset covers I-5, I-405, I-90, and SR-520 connected freeways; it has 5 minutes time step interval and 323 sensor stations and covers 5 minutes intervals over the entirety of 2015. Fig. 4 is the diagram of the loop detector dataset.

B. Experimental Setup

In input data Y_T^P, each sample is a 2-dimensional vector. The dimensions of input data are [n, P] = [10,323], based on model description. The time lag is set as 10.

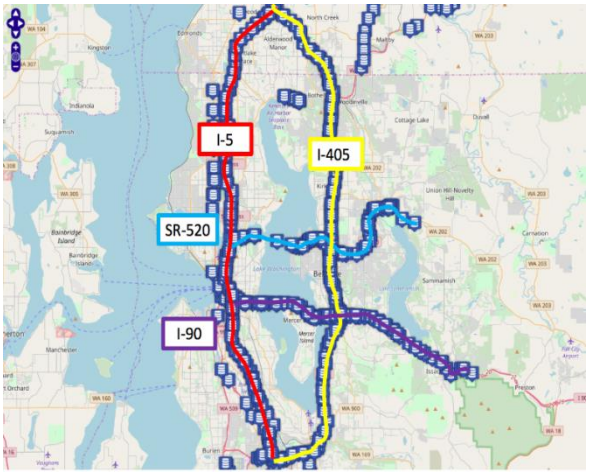

Fig. 4. Loop Detector Dataset.

C. Model Optimization

In training, mini-batch gradient descent is used. MSE is used as a loss function and RMSProp optimizer. To avoid overfitting, the early stopping mechanism is used.

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left(h_{pred} - h_{true}\right)^2 \quad (19)$$

Where h_{pred} is the predicted results, h_{true} is the ground truth value and N denotes the number of training samples.

D. Evaluation Criteria

To evaluate the effectiveness of state-of-the-art models, Mean Squared Errors (MSE), Root Mean Squared Errors (RMSE) and R2 are calculated using the following equations:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(h_{pred} - h_{true}\right)^2 \quad (20)$$

$$RMSE = \sqrt{\sum_{N}^{1}\frac{\left(h_{pred} - h_{true}\right)^2}{N}} \quad (21)$$

$$R^2 = 1 - \frac{Explanied\ Variation}{Total\ Variation} \quad (22)$$

Where h_{pred} is the predicted results, h_{true} is the ground truth value and N denotes the number of training samples. Fig. 5 and Fig. 6 show the training and validation loss of different models compared in this study.

E. Comparison with State-of-the-Art Models

We compared the proposed model with state-of-the-art models to check the efficiency and effectiveness of the model. We compared the proposed model with Fully Convolutional Networks (FCN), Gated Recurrent Unit (GRU), Long-short Term Memory (LSTM), Long-short Term Memory with Deep Neural Network (LSTM-DNN) and Bidirectional Long-short Term Memory (BDLSTM). The performance comparison of different algorithms are demonstrated in Table I. The Means Squared Error and Root mean squared error of Fully convolutional network is 0.68 and 8.22 respectively. The performance of simple GRU is better than Fully Convolutional Network and the MSE and RMSE are reduced to 0.25 and 5.01 respectively. Because GRU cannot process long sequences and simple GRU is not a suitable choice for this problem. So, we compared the performance of LSTMs in this study. LSTM can address the short comings of GRU because of the gated structure of LSTM and the results are significantly better than GRU and FCN and error reduced to 0.1541 and 3.92. Furthermore, we added a DNN layer with LSTM and demonstrated that results are comparatively better, and error reduced to 0.1463 and 3.83. We also compared the performance of BDLSTM, and we concluded that the performance of BDLTSM is better than LSTM because BDLSTM process the sequence data both backward and forward and error reduces to 0.1408 and 3.75, respectively. The proposed model composed of BDLSTM and Attention-based GRU and its performance is superior over state-of-the-art models compared in this study and error reduces to 0.1371 and 3.70, respectively. Additionally, we also calculated the R^2 factor, which showed the performance of every model compared in this study and R^2 values of the proposed model are higher than other state of the art models compared which is 0.8620. In summary, we can come to know that the proposed model SBAG outperformed over FCN, GRU, LSTMS and BDLSTMs, as shown in Table I, respectively. Fig. 7 shows the error of models compared in this paper.

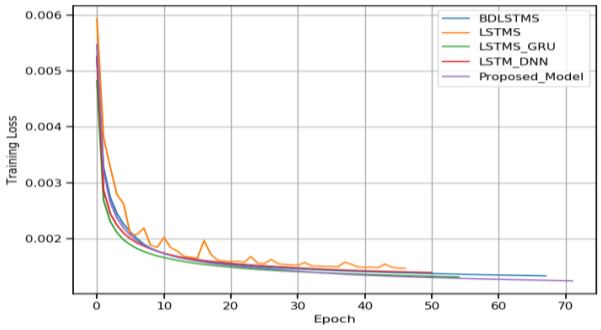

Fig. 5. Training Loss of different Models.

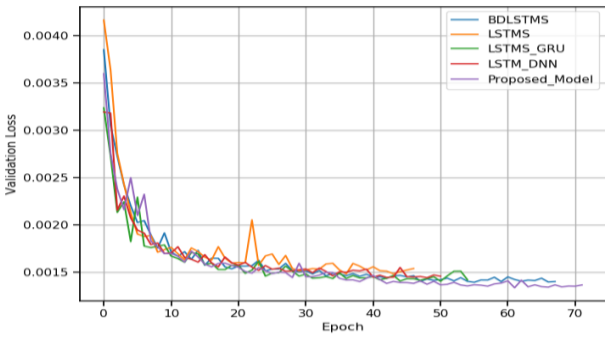

Fig. 6. Validation Loss of different Models.

TABLE. I. COMPARISON WITH STATE OF THE ART

Model	MSE	RMSE	R^2
FCN	0.6772	8.22	0.48
GRU	0.2512	5.01	0.7165
LSTM	0.1541	3.92	0.8550
LSTM-DNN	0.1463	3.83	0.8530
BiLSTM	0.1408	3.75	0.8548
(SBAG) Proposed Model	**0.1371**	**3.70**	**0.8620**

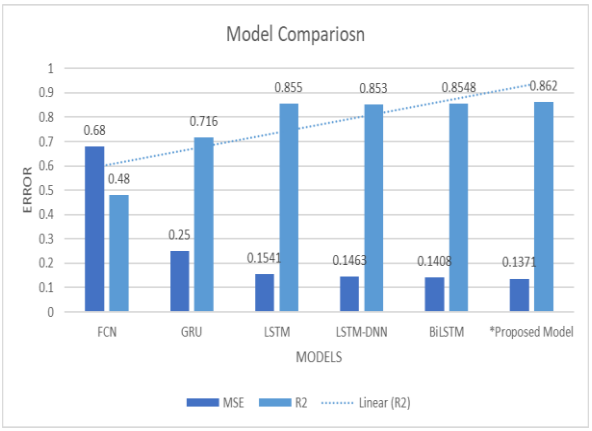

Fig. 7. Model Comparision Results.

IV. Conclusion

In this study, Stacked Bidirectional Long-Short Term Memory Attention-based Gated Recurrent Unit (SBAG) neural network is proposed for the large scale traffic speed prediction. The contributions and improvements focus on the three aspects: 1) Along a corridor, we consider the traffic forecasting to a large scale traffic network instead of specific location/several adjacent locations; 2) we proposed a hybrid model considering both backward and forward dependencies of large scale traffic data; 3) We considered the significance of attention layer that by adding attention layer with GRU, the performance of the proposed model significantly enhanced. Experiment results show that the SBAG is the best model for predicting large scale traffic speed. In comparison with GRU, LSTMs and BDLSTMs methods, Bidirectional LSTM and Attention-based GRU proven to be more competent to learn spatial-temporal features and the best model for large scale traffic speed prediction.

In the future, further improvements can be made based on the proposed study and will improve more significant towards the graph structure to interpret and learn the spatial features or to improve by hybridizing with GANs [27] to make further amendments.

References

[1] H. van Lint and C. van Hinsbergen, "Short-Term Traffic and Travel Time Prediction Models," Transp. Res. E-Circular, 2012.

[2] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," Transp. Res. Part C Emerg. Technol., vol. 54, pp. 187–197, May 2015.

[3] S. D. Khan, F. Porta, G. Vizzari, and S. Bandini, "Estimating speeds of pedestrians in real-world using computer vision," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014.

[4] S. M. B. Muhammad Saqib, Sultan Daud Khan, "Vehicle Speed Estimation using Wireless Sensor Network," in INFOCOMP 2011 : The First International Conference on Advanced Communications and Computation, 2011.

[5] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," Transp. Res. Part C Emerg. Technol., 2011.

[6] Q. Ye, W. Y. Szeto, and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," IEEE Trans. Intell. Transp. Syst., 2012.

[7] D. Park and L. R. Rilett, "Forecasting freeway link travel times with a multilayer feedforward neural network," Comput. Civ. Infrastruct. Eng., 1999.

[8] H. Yin, S. C. Wong, J. Xu, and C. K. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," Transp. Res. Part C Emerg. Technol., 2002.

[9] J. W. C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks," Transp. Res. Rec. J. Transp. Res. Board, 2007.

[10] P. Wang, Y. Kim, L. Vaci, H. Yang, and L. Mihaylova, "Short-Term Traffic Prediction with Vicinity Gaussian Process in the Presence of Missing Data," in 2018 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2018, 2018.

[11] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting," in Proceedings of the 2017 SIAM International Conference on Data Mining, 2017.

[12] Z. Cui, R. Ke, and Y. Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," arXiv Prepr. arXiv1801.02143, 2018.

[13] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, "A Capsule Network for Traffic Speed Prediction in Complex Road Networks," in 2018 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2018, 2018.

[14] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," JMLR, 2015.

[15] X. Wei, J. Li, Q. Yuan, K. Chen, A. Zhou, and F. Yang, "Predicting Fine-Grained Traffic Conditions via Spatio-Temporal LSTM," Wirel. Commun. Mob. Comput., 2019.

[16] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," Transp. Res. Part C Emerg. Technol., 2018.

[17] Y. Kim, P. Wang, and L. Mihaylova, "Structural Recurrent Neural Network for Traffic Speed Prediction," arXiv Prepr. arXiv1902.06506, 2019.

[18] Y. Duan, Y. Lv, and F. Y. Wang, "Travel time prediction with LSTM neural network," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2016.

[19] X. Jiang and H. Adeli, "Wavelet packet-autocorrelation function method for traffic flow pattern analysis," Comput. Civ. Infrastruct. Eng., 2004.

[20] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting & Control. 2015.

[21] Y. Y. Chen, Y. Lv, Z. Li, and F. Y. Wang, "Long short-Term memory model for traffic congestion prediction with online open data," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2016.

[22] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings, 2013.

[23] S. Xuan, H. Kanasugi, and R. Shibasaki, "DeepTransport: Prediction and simulation of human mobility and transportation mode at a citywide level," in IJCAI International Joint Conference on Artificial Intelligence, 2016.

[24] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," 2015.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Microbes Infect., vol. 11, no. 3, pp. 367–373, Sep. 2014.

[26] 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," 2018.

[27] M. Nabeel, A. Riaz, and W. Zhenyu, "Cas-GANs: An approach of dialogue policy learning based on GAN and RL techniques," Int. J. Adv. Comput. Sci. Appl., 2019.