Micromobility Flow Prediction: A Bike Sharing Station-level Study via Multi-level Spatial-Temporal Attention Neural Network

Xi Yang Jiachen Wang Song Han Suining He[†]
School of Computing, University of Connecticut [*]

ABSTRACT

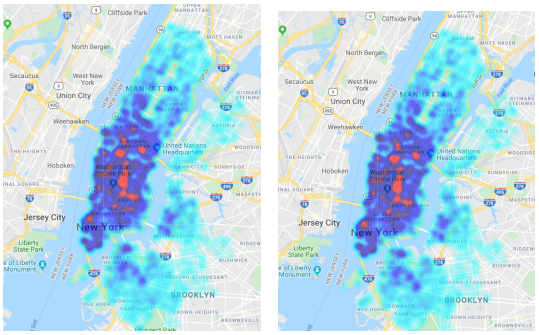
Efficient use of urban micromobility resources such as bike sharing is challenging due to the unbalanced station-level demand and supply, which causes the maintenance of the bike sharing systems painstaking. Prior efforts have been made on accurate prediction of bike traffics, i.e., demand/pick-up and return/drop-off, to achieve system efficiency. However, bike station-level traffic prediction is difficult because of the spatial-temporal complexity of bike sharing systems. Moreover, such level of prediction over entire bike sharing systems is also challenging due to the large number of bike stations. To fill this gap, we propose `BikeMAN`, a multi-level spatio-temporal attention neural network to predict station-level bike traffic for entire bike sharing systems. The proposed network consists of an encoder and a decoder with an attention mechanism representing the spatial correlation between features of bike stations in the system and another attention mechanism describing the temporal characteristic of bike station traffic. Through experimental study on over 10 millions trips of bike sharing systems (> 700 stations) of New York City, our network showed high accuracy in predicting the bike station traffic of all stations in the city.

1 INTRODUCTION

Bike sharing has become an essential micromobility option for urban residents worldwide due to its advantages in convenience and economy over other urban mobility means. Within a docked/station-based bike sharing system, a micromobility user picks up a bike from one station (demand), accomplishes her or his trip, and drops the bike off at another station (return).

Through data analysis of the bike sharing systems in several cities, a critical issue lies in that there were imbalanced demands and returns distributions in those bike sharing systems, i.e. a majority of demand and a majority of redistributed/returned bikes only occurred in a small portion of stations [14]. A typical imbalanced traffic map of the bike sharing system is shown in Figure 1, where the bike traffics spatially vary across the city map. Such imbalance not only degrades the user satisfaction, but also leads to operation inefficiency and resource waste to the bike sharing systems. How to predict the future bike traffic (pick-ups and drop-offs) and the subsequent rebalanced distribution of bikes in every station becomes essential and challenging. In order to solve the balancing problems, the general idea is propose an accurate pick-up and drop-off prediction model.

Bike sharing traffic (pick-up and drop-off) prediction can usually be defined as a time series prediction problem from multi-source and heterogeneous data [8]. In this work we propose a Bike sharing Multi-level Attention neural Network called `BikeMAN` for station-based pick-up and drop-off prediction with historical spatial and temporal traffic features and some external related factors (including weather conditions and points of interest (POIs)). Our work will not only give the prediction results for bike sharing system operators to efficiently rebalance their bike distribution, but also to provide information for citizens about the bike availability at the stations of their interest in future hours or minutes. Our main contributions are as follows:



(a) Demands on Fri, Jun 7, 2019 (b) Returns on Fri, Jun 7, 2019

Figure 1: Imbalanced bike usage heatmap of NYC. Red areas represent high demand, while light blue areas represent low demand.

1) *Comprehensive bike data analysis*: We have provided a comprehensive and detailed real-world data analysis on how the weather conditions, station locations and surrounding POIs impact the bike usage in the metropolitan area like New York City, and visualized them to validate our insights.
2) *Multi-level spatio-temporal attention model*: To the best of our knowledge, we propose the *first* novel multi-level attention neural network consisting of a spatial attention and a temporal attention mechanism to predict station-level bike usage (pick-ups and drop-offs), which accurately predicts the traffics of all the stations. The spatial attention mechanism developed in this work captures not only the correlation of features within a single station, but also the correlation spatial of features across all stations in the city. The developed temporal attention mechanism captures temporal correlation between features across different timestamps.
3) *Extensive experimental studies*: We have conducted extensive experimental studies upon over 10 millions trips. Our multi-level attention neural network demonstrates high accuracy compared with other basic and encoder-decoder models in multi-station predictions.

[*†]Corresponding Author

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Xi Yang Jiachen Wang Song Han Suining He†
School of Computing, University of Connecticut

The rest of the paper is organized as follows. We first review the related in work in Section 2. After that, we present the data analysis in Section 3. Given the features, we then present the core model of our approach in Section 4. We show the performance evaluation in Section 5, and conclude in Section 6.

2 RELATED WORK

The station-level prediction is challenging because of the spatial-temporal characteristics of the bike sharing demand pattern of a station and a large number of stations in a city. Recent studies have been focused on tackling this challenge. Hulot et al. [6] proposed a statistical learning model using temporal and weather features to predict the hourly pick-ups and drop-off at each station of the system. Chen et al. [4] used recurrent neural networks (RNNs) trying to capture the temporal characteristics in bike sharing demand. Li, et al. [8] proposed a representation learning method which encoded the spatial-temporal information of the bike sharing system, and the representation was then fed to Long-Short-Term-Memory (LSTM) to predict bike demand. Chai, et al. developed a multi-graph convolutional networks as the representation of the spatial-temporal data instead, but they also fed the learned representation to LSTM [3]. Lin et al. [10] proposed a novel graph convolutional neural network with data-driven graph filter model to predict hourly demand for a large number of stations in the bike sharing systems. Despite the accuracy shown upon the selected bike stations (say, only 272 selected out of all ~800 stations in [10]), they have not extensively, systematically and comprehensively studied the *entire* bike sharing system (like NYC) for station traffic prediction, and provided deployment insights like our work here.

Attention mechanisms have been widely used in natural language processing [1, 11]. The idea of implementing attention on neural networks extend to speech recognition [2, 5, 15], video summarization [7] and captioning [12], and human action recognition [13]. Liang, et al. employed a attention mechanisms in an encoder-decoder structure for station-level geo-sensory time series prediction [9]. Different from the prior studies, based on extensive data analysis, we design a novel multi-level attention mechanism for bike sharing system.

3 DATA ANALYSIS

We first overview the datasets and then present the data analysis of the bike sharing dataset.

3.1 Data Overview

We use three types of data sets in our project: the user trip data, the weather condition data and POIs data. Details of each are discussed in the following.

1) **User Trip Data**: It contains information of every single trip including the trip's duration, start/end time , start/end stations and their longitude/latitude, and user information. The user trip data of NYC bike sharing system is provided on a monthly base with over 2 million records each month. In this study, we use the user trip data from June 2019 to October 2019 collected online[1], with a total of over 10 millions trips.
2) **Weather Condition Data**: Characterizing the hourly weather historical conditions, it contains hourly temperature, precipitation, wind speed, humidity and other 6 weather observations. We adopt the hourly weather data of New York City from June 2019 to October 2019[2].
3) **POI Data**: POIs data, obtained from NYC Open Data Portal[3], includes the GPS coordinates, facility domain and some other details (say, facility names) of over 20 thousand POI facilities in New York City. There are totally 13 major types of POIs (e.g. residential, transportation and commercial etc.), each of which contains a number of minor types.

3.2 Analysis

We present the following insights regarding effects of weather and POIs upon the station traffics based on the data and visualization.

3.2.1 Station Traffic and weather. The bike usage is highly correlated with weather conditions. The two dominating effects are *precipitation* and *wind speed*. In general, wind speed and precipitation negatively impact the number of bike usage. Figure 2 shows the relationship between the hourly user trip amount of entire city and temperature, precipitation and wind speed in June, 2019 in NYC. Precipitation has the greatest influence on bike usage. On June 10, 0.46 inch of precipitation reduced the total amount of cycling from an average of 8000 to less than 4000.

Wind speed is another important factor. Usually when the wind speed is higher than 15mph, a notable decrease in bike usage will happen. While the temperature's impact on hourly bike demand is not as significant as the others, but its long-term influence cannot be ignored. Given above, we take those three factors into account and feed them to the deep learning model as input features.

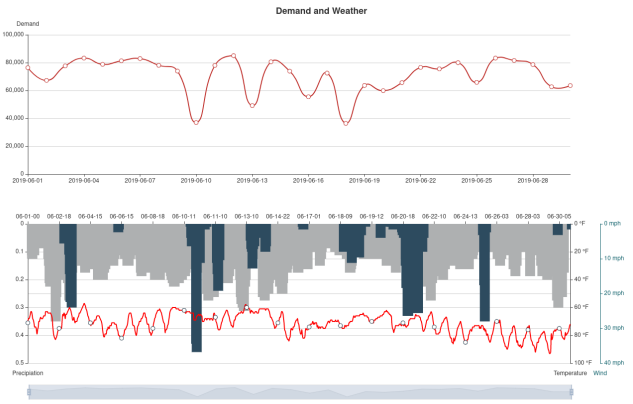


Figure 2: Bike demand and weather correlation in June, 2019.

3.2.2 Station Traffic and POIs. From the previous demand heatmap, Figure 1, we can also see that the bike usage of stations is highly location dependent. One of the major reasons behind this is that different locations have different types of POIs. Figure 3 illustrate this clearly. We use the distribution of commercial and residential POIs from the 13 categories as an example for visualization. The

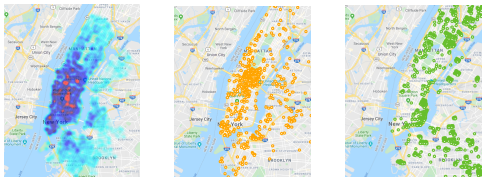
[1]https://www.citibikenyc.com/system-data

[2]https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA/

[3]https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj

commercial POIs are mainly located on the center of Manhattan, surrounded by residential facilities. The different distributions of each types of POIs result in the dependency of bike usage on locations. Therefore, we take into account the distribution of the 13 types of POIs within <150m of each station as input features.

In order to avoid the inclusion of another bike station when counting the nearby POIs around one station, we have to choose the distance to the station as small as possible. However, during our study, we found that the closest distance between two bike stations is less than 100m. If we use this closest distance, a large number of stations will have no POIs within this distance. Therefore, we set the distance of counting surrounding POIs to be 150 m. In this way, there are 91.25% of stations of which distance between each other is larger than 150m, and there are only 6.92% of stations having no POIs within this range, which suffices to characterize POI features.



(a) Demand heatmap on Thu, Jun 6, 2019 (b) Distribution of commercial POIs (c) Distribution of residential POIs

Figure 3: Demand heatmap and distribution of commercial and residential POIs.

In summary, based on the analysis above, besides the bike demand we use weather conditions including precipitation, wind speed, and temperature, nearby (< 150 m) POIs, and station longitude and latitude as the input features of each bike stations.

4 BIKEMAN FOR BIKE STATION TRAFFIC PREDICTION

Our system's framework is shown in Figure 4. Bike demand data set and two external features: hourly weather conditions data set and point of interest data set are processed respectively and then integrated together as the neural networks input. The model then will output the future demand prediction of one station.

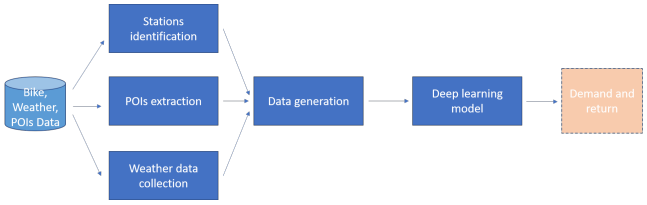


Figure 4: System architecture of BikeMAN.

4.1 Preliminaries

We consider that for totally N stations in a city, each station i has s features at timestamp t : $\boldsymbol{F}_{i,t} = (F_{i1,t}, ..., F_{ij,t}, ..., F_{is,t})$, where $i \in [1, ..., N]$ and $j \in [1, ..., s]$, and let $F^{(p)}_{i1,t}$ and $F^{(d)}_{i1,t}$ be the bike traffic (pick-ups and drop-offs) of station i at time t.

The problem of this work is defined as: given the features of each station of previous $\mathcal{T}$ timestamps,

$$\mathbb{F}_{t-\mathcal{T}:t} = (\boldsymbol{F}_{1,t-\mathcal{T}:t}, ..., \boldsymbol{F}_{i,t-\mathcal{T}:t}, ..., \boldsymbol{F}_{N,t-\mathcal{T}:t}), \quad i \in [1, ..., N], \quad (1)$$

we aim at predicting the bike traffic of all the stations $\hat{y}^{(P)}_{i,t:t+\tau}$ and $\hat{y}^{(D)}_{i,t:t+\tau}$, for all $i \in [1, ..., N]$ at the next τ timestamps in the bike sharing system such that

$$\min \sqrt{\sum_{k=0}^{\tau} ||\hat{y}^{(P)}_{i,t+k} - F^{(P)}_{i1,t+k}||^2}, \quad \text{and} \quad \min \sqrt{\sum_{k=0}^{\tau} ||\hat{y}^{(D)}_{i,t+k} - F^{(D)}_{i1,t+k}||^2}. \quad (2)$$

4.2 Design of Multi-level Attention Neural Networks

The structure of our network follows the encoder-decoder architecture with two attention mechanisms on the top of it. The *encoder* and *decoder* are two separate RNN components each of which is a stack of two LSTMs. In encoder part, a *spatial attention* mechanism is implemented on the input features to generate the inputs of the RNN with the previous RNN states. In the decoder part, a *temporal attention* mechanism is implemented on the output hidden states of the other RNN to compute the decoder outputs based on their correlation with encoder hidden states. We detail the two attention mechanisms as follows.

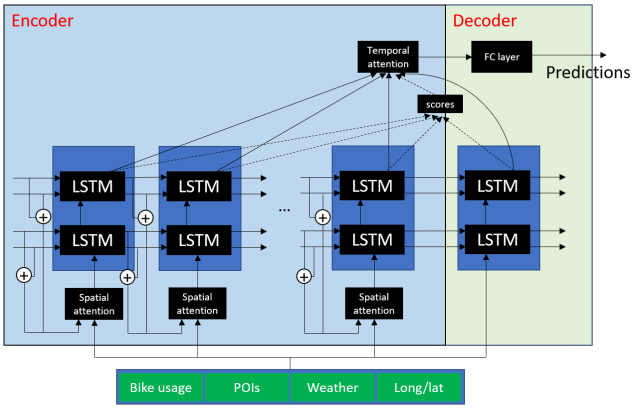


Figure 5: Core model architecture of BikeMAN.

4.2.1 Spatial Attention. The spatial attention mechanism is developed to describe the spatial correlation between the features of all stations. The intuition behind the spatial attention is that the traffic at each station is largely location-dependent as shown in Fig. 3, which can be characterized by the neighborhood POI distributions. For example, on a weekday morning demand in residential areas would be higher than commercial areas because people need to ride to work. Inspired by the work of [1, 2, 7, 9, 12, 13], we designed the following spatial attention mechanism.

Specifically, given the set of features of all the station of previous $\mathcal{T}$ timestamps, $\mathbb{F}_{t-\mathcal{T}:t}$, we first flatten it to create a vector $\mathbf{f}_{t-\mathcal{T}:t}$ so that the length at each timestamp is $N \times s$. For the k-th feature of the vector $\mathbf{f}_{t-\mathcal{T}:t}$, denoted as $\mathbf{f}^k_{t-\mathcal{T}:t}$, the score of this feature is

calculated by Eq. (3):

$$\epsilon_t^k = \mathbf{v}_s^\intercal \tanh\left(\mathbf{W}_s[\mathbf{h}_{t-1}; \mathbf{c}_{t-1}] + \mathbf{U}_s \mathbf{f}_{t-\mathcal{T}:t}^k + \mathbf{b}_s\right), \tag{3}$$

where $\mathbf{v}_s$, $\mathbf{b}_s$, $\mathbf{W}_s$, and $\mathbf{U}_s$ are the learnable model parameters, $\mathbf{h}_{t-1}$ and $\mathbf{c}_{t-1}$ are the hidden state and cell state of the RNN, respectively, at the previous timestamp. Then the attention weight of this feature is a softmax function of the score of this feature as in Eq. (4):

$$\alpha_t^k = \frac{\exp\left(\epsilon_t^k\right)}{\sum_{j=1}^{N\times s} \exp\left(\epsilon_t^j\right)}, \tag{4}$$

This way, the spatial attention captures the correlation between features of a station as well as the correlation between features across all of them. The input vectors of the RNN at timestamp t is then computed by the multiplication of the attention weights and original input features of the model:

$$\tilde{\mathbf{f}}_t = \left(\alpha_t^1 \mathbf{f}_t^1, \quad ..., \quad \alpha_t^k \mathbf{f}_t^k, \quad ..., \quad \alpha_t^{N\times s} \mathbf{f}_t^{N\times s}\right). \tag{5}$$

4.2.2 Temporal Attention. The temporal attention mechanism captures correlation between features across different timestamps. The intuition behind this is that there is a strong time dependency of the demand of a station at time t on previous demand. For example, on a sunny day morning if a large number of people choose to ride bikes to work, then in the evening the demand of stations around business areas is likely to be huge due to the return flows. The temporal attention scores between the current decoder RNN hidden state and one of the previous encoder RNN hidden states, $\lambda_{t,t'}$, are calculated as Eq. (6) by a concatenation manner as [11]:

$$\lambda_{t,t'} = \mathbf{v}_l^\intercal \tanh\left(\mathbf{W}_l[\mathbf{h}_t; \mathbf{h}_{t'}]\right), \tag{6}$$

where $\mathbf{v}_l$ and $\mathbf{W}_l$ are learnable parameters, $\mathcal{T}$ is the number of timestamps of encoder, $\mathbf{h}_t$ is hidden state of encoder at timestamp t which is in range of $[1, \mathcal{T}]$, and $\mathbf{h}_{t'}$ is hidden state of decoder at timestamp t'. The attention weight, denoted as $\gamma_{t,t'}$, is then a softmax function of $\lambda_{t,t'}$ as in Eq. (7):

$$\gamma_{t,t'} = \frac{\exp\left(\lambda_{t,t'}\right)}{\sum_{t=1}^{\mathcal{T}} \exp\left(\lambda_{t,t'}\right)}, \tag{7}$$

Finally, the output of the attention mechanism, $\mathbf{d}_{t'}$, is then a weighted sum of the encoder RNN hidden states shown in Eq. (8):

$$\mathbf{d}_{t'} = \sum_{t=1}^{\mathcal{T}} \gamma_{t,t'}[\mathbf{h}_t; \mathbf{h}_{t'}]. \tag{8}$$

4.2.3 Predictions and Model Training. Through a fully connected (fc) layer, the concatenation of decoder RNN hidden state, $\mathbf{h}_{t'}$, and weighted sum of encoder RNN hidden states, $\mathbf{d}_{t'}$, is mapped to final predictions of traffics ($\hat{\mathbf{y}}_{t'} \in \{\hat{\mathbf{y}}_{t'}^{(P)}, \hat{\mathbf{y}}_{t'}^{(D)}\}$) at all stations of time t':

$$\hat{\mathbf{y}}_{t'} = \mathbf{W}_p[\mathbf{d}_{t'}; \mathbf{h}_{t'}] + \mathbf{b}_p, \tag{9}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are learnable parameters. The entire model is trained using the Adam optimizer which minimizes the mean square error between the predicted values and the ground truths as the loss function:

$$\mathcal{L}(\theta) = \sqrt{(\mathbf{y}_{t'} - \hat{\mathbf{y}}_{t'})^2}, \tag{10}$$

which is fast and efficient in practice. Based on the machines used, the training time for the New York City only takes 4 hours.

5 EXPERIMENTAL EVALUATIONS

We first present the evaluation setup, followed by the experimental results.

5.1 Evaluation Setup

5.1.1 Data Pre-processing. : We evaluate the BikeMAN and related schemes with the datasets presented in Section 3.1. All experimental evaluations are conducted upon a desktop with Intel i5-8700, 16GB RAM, an Nvidia GeForce GTX 1080Ti and Windows 10. We first go through all the bike sharing datasets and delete the stations which do not appear in all the datasets. We find 766 stations in common among all the datasets from Jun, 2019 to Oct, 2019. Then we extract the nearby (< 150 m) POIs around each of those common bike stations. The time interval is set to be 1 hour. The time stamps of encoder inputs are set 12, and the time stamps of decoder inputs are set 1. This means that we want to use this model to predict the return of all the stations in the next following hours based on their demand of previous 12 hours.

Since some of the weather conditions have multiple data points between two continuous whole points, we take mean values of these data over an hour as the hourly data of this time. Most of the precipitation data is 0 as it is usually sunny in NYC. However, when it is rainy, the hourly precipitation is not high, typically within 1 inch, which is too small compared to the bike demand/return. Therefore, we normalize the precipitation data by min-max normalization such that they fall in the range of [0, 10]. For longitude and latitude data, the difference between two stations is even smaller, typically at the magnitude of 10^{-2}. Therefore, we normalize these two features again by min-max normalization to make them fall in the range of [0, 100]. We use the trips in June, July, and August 2019 for training and validating our model, and use the ones in October, 2019 to test our trained model.

5.1.2 Parameter Setups. : The parameters are set as followed. The learning rate is 0.001. The rate of gradients clipping is 2.5. The dropout rate is 0.3. The number of layers stacked in LSTMs is 2. The total number of stations is 766, each of which has 19 features. The number of encoder time stamps is 12 and that of decoder is 1. Since the output of BikeMAN is a vector of length 766 with each element being the demand for a station, the dimensions of hidden states h in every LSTM of encoder and decoder are 1,024. The batch size is set to be 64. Total number of training epochs is 100 with 2,300 iterations.

5.1.3 Comparison Metrics. : Root mean square error (RMSE) and mean absolute error (MAE) are selected as the metrics for evaluations:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i}^{M}(y_i - \hat{y}_i)^2}, \tag{11}$$

$$MAE = \frac{1}{M}\sum_{i}^{M}(y_i - \hat{y}_i), \tag{12}$$

where M, y_i and $\hat{y}_i$ are the total number of predictions made by the model, the ground-truth of the bike pick-ups/drop-offs and the predicted bike pick-ups/drop-offs, respectively.

5.2 Evaluation Results

5.2.1 Predictions by Varying Number of Involved Stations. As a starting point, we first evaluate BikeMAN on bike traffic of a single station. We choose three stations for evaluation according to their demands sorted from high to low. Station 519 is the most popular bike station in NYC with the highest monthly traffic among all datasets. We then extend to carry out multi-station prediction, that is, training and predicting multiple station using a single model at the same time. Besides the 766 stations which we find commonly existing among all the datasets, we choose here 190 stations out of the 766 stations that are common among all the datasets. The 190 stations we choose is based on the total demand from June to October 2019. We first sort all the stations by the total demand of and then pick 190 out of them that equally split the stations array. The accuracies of our model are compared with the encoder-decoder of the same structure as BikeMAN but no attention mechanisms on it. The comparison accuracies are shown in Table 1.

Table 1: Predictions of BikeMAN compared with basic encoder-decoder

Number of Stations	BikeMAN		LSTM Enc-Dec	
	RMSE	MAE	RMSE	MAE
1 (station 519)	11.70	7.29	11.78	7.16
190	3.86	1.95	5.95	3.26
766	3.79	2.08	5.64	3.26

The results are very interesting as shown in Table 1. It turns out that the more stations we test, the more effective the attention mechanism is in improving the model accuracy. Only with multi-station predictions will BikeMAN outperform the basic encoder-decoder. The reasons probably lie in the spatial attention mechanism we design. In this work, the spatial attention is able to capture not only the correlation between the bike demand and other features for a single station, but also the correlation between the features of stations across the entire city. If there is only one station's features fed as input, the model will lose the information of others. Therefore, our model works better for predictions with multiple stations as inputs rather than single-station predicitons.

5.2.2 Entire system predictions. For operators of a bike sharing system, it is rather important to forecast the pick-ups and drop-offs for every station in the system than just a portion of stations. Therefore, we carry out extensive experimental test on the demand/pick-up and return/drop-off predictions for all 766 bike stations in NYC and compare our model with the baselines. We vary the RNN unit in BikeMAN from LSTM to Gated Recurrent Unit (GRU) (BikeMAN-GRU). Our models are compared with basic LSTM and GRU encoder-decoder. The results of station traffic predictions are shown in Table 2.

From Table 2, we can see that our model BikeMAN with LSTM as the RNN units has the best performance among all the models

Table 2: Predictions of BikeMAN compared with basic encoder-decoder

Models	Demand		Return	
	RMSE	MAE	RMSE	MAE
GRU Dec-Enc	5.661	3.312	5.729	3.386
LSTM Dec-Enc	5.678	3.350	5.693	3.338
BikeMAN-GRU	3.461	1.884	3.441	1.852
BikeMAN	3.366	1.818	3.369	1.797

for both demand and return predictions. The RMSE and MAE of BikeMAN drop by over 40% from LSTM Dec-Enc. This proves the success of the spatial and temporal attention mechanisms proposed in Sec. 4, which are able to capture more information in the heterogeneous input data. The model accuracy decreases slightly when LSTM cells are replaced by GRU cells. For each of the models tested, the prediction accuracy of demand is almost the same as that of return. From Fig. 1 it can be seen that the demand and return have similar pattern over 24 hour period, which explains the similar performance of models on demand and return predictions. The predicted station traffics for the first week of October, 2019 by BikeMAN and BikeMAN-GRU compared to the ground truth are shown in the figures below. We can see that BikeMAN highly resembles the ground-truths, validating the accuracy and effectiveness.

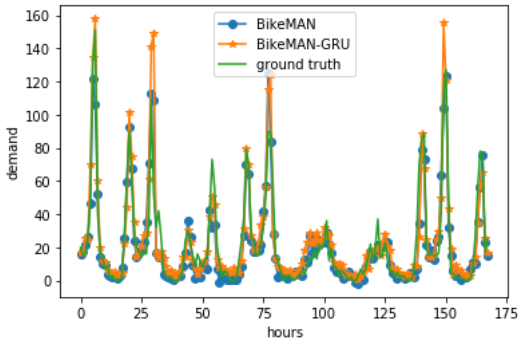


Figure 6: Prediction of BikeMAN and BikeMAN-GRU on the bike demand of station 519 for the first week of October, 2019 (red line) compared with ground truth.

6 CONCLUSION

In order to realize accurate and practical bike sharing station traffic prediction, we propose BikeMAN, a novel multi-level attention neural network. We design a novel encoder-decoder architecture, and leverage the spatial and temporal attentions to capture the correlations between the station features. BikeMAN further takes into account the external features like weather and points of interest to enhance the performance. We have conducted extensive data analysis and experimental studies upon the dataset from the New York City, demonstrating the accuracy of our proposed scheme.

7 ACKNOWLEDGEMENT

This project is supported, in part, by the National Science Foundation (NSF) under Grant 2303575, the Google Research Scholar

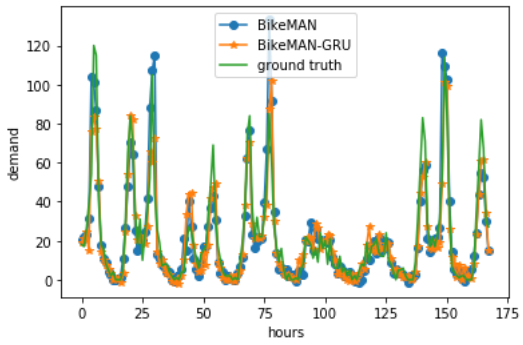


Figure 7: Prediction of `BikeMAN` and `BikeMAN-GRU` on the bike return of station 519 for the first week of October, 2019 (red line) compared with ground truth.

Award Program (2021–2022), the NVIDIA Applied Research Accelerator Program Award (2021–2022), and the University of Connecticut (UConn) Research Excellence Program Awards (2020–2021, 2022–2024).

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *Proc. IEEE ICASSP*. IEEE, 4945–4949.

[3] Di Chai, Leye Wang, and Qiang Yang. 2018. Bike flow prediction with multi-graph convolutional networks. In *Proc. ACM SIGSPATIAL*. ACM, 397–400.

[4] Po-Chuan Chen, He-Yen Hsieh, Xanno Kharis Sigalingging, Yan-Ru Chen, and Jenq-Shiou Leu. 2017. Prediction of station level demand in a bike sharing system using recurrent neural networks. In *Proc. IEEE VTC Spring*. IEEE, 1–5.

[5] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proc. NIPS*. 577–585.

[6] Pierre Hulot, Daniel Aloise, and Sanjay Dominik Jena. 2018. Towards Station-Level Demand Prediction for Effective Rebalancing in Bike-Sharing Systems. In *Proc. ACM SIGKDD*. ACM, 378–386.

[7] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).

[8] Youru Li, Zhenfeng Zhu, Deqiang Kong, Meixiang Xu, and Yao Zhao. 2019. Learning Heterogeneous Spatial-Temporal Representation for Bike-Sharing Demand Prediction. In *Proc. AAAI*, Vol. 33. 1004–1011.

[9] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. 2018. GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction.. In *Proc. IJCAI*. 3428–3434.

[10] Lei Lin, Zhengbing He, and Srinivas Peeta. 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies* 97 (2018), 258–276.

[11] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[12] Jingkuan Song, Zhao Guo, Lianli Gao, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. 2017. Hierarchical LSTM with adjusted temporal attention for video captioning. *arXiv preprint arXiv:1706.01231* (2017).

[13] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proc. AAAI*.

[14] Shuai Wang, Tian He, Desheng Zhang, Yunhuai Liu, and Sang H Son. 2019. Towards Efficient Sharing: A Usage Balancing Mechanism for Bike Sharing Systems. In *Proc. WWW*. ACM, 2011–2021.

[15] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018. Improved training of end-to-end attention models for speech recognition. *arXiv preprint arXiv:1805.03294* (2018).