Multi-Task Spatial-Temporal Graph Attention Network for Taxi Demand Prediction

Mingming Wu
School of Software Engineering, University of Science and Technology of China
Hefei, China
SA517402@mail.ustc.edu.cn

Chaochao Zhu
School of Software Engineering, University of Science and Technology of China
Hefei, China
cczhu@mail.ustc.edu.cn

Lianliang Chen
Department of Computer Science, University of Science and Technology of China
Hefei, China
cll006@mail.ustc.edu.cn

ABSTRACT
Taxi demand prediction is of much importance, which enables the building of intelligent systems and smart city. It is necessary to predict taxi demand accurately to schedule taxi fleet in a reasonable and efficient way and to reduce the pressure of traffic jam. However, the taxi demand involves complex and non-linear spatial-temporal impacts. The superiority of deep learning makes people explore the possibility to apply it to traffic prediction. State-of-the-art methods on taxi demand prediction only capture static spatial correlations between regions (e.g., Using static graph embedding) and only take taxi demand data into consideration. We propose a Multi-Task Spatial-Temporal Graph Attention Network (MSTGAT-Net) framework which models the correlations between regions dynamically with graph-attention network and captures the correlation between taxi pick up and taxi drop off with multi-task training. To the best of our knowledge, it is the first paper to address the taxi demand prediction problem with graph attention network and multi-task learning. Experiments on real-world taxi data show that our model is superior to state-of-the-art methods.

CCS Concepts
• **Applied computing→Forecasting** • **Computing methodologies→Temporal reasoning**

Keywords
Deep learning; Taxi demand prediction; Multi-task learning;

1. INTRODUCTION
Traffic reflects the mobility of human in an urban city. One of the most popular trends is to build a smart city. To achieve this, the intelligent transportation system will play a important role in the future. Hence, It is necessary to design traffic prediction models. As is known to us, the more accurate the prediction of traffic, the better we can schedule taxi fleet and reduce the pressure of traffic jam. To this day, the online taxi requesting services such as Uber and Didi Chuxing (in China) are getting more and more popular. A large amount of taxi data generates every day continuously. How to make full use of these big data to improve the performance of taxi demand prediction is a critical problem, which has drawn the attention of many researchers.

In this paper, we focus on the taxi demand prediction problem. Specifically, we want to predict the number of taxi demand for a given region in a given future timestamp based on the historical taxi demand data. Great efforts have been made in the field of traffic data prediction, most of them utilize the time series techniques to model the problem. For example, auto regressive integrated moving average (ARIMA) and its variants are applied to predict traffic data [1]; More impacts are taken into consideration based on the time series methods recently, such as spatial relations [2] and external data (e.g., POI, weather, and events). While these traditional methods perform better with additional factors, the complex nonlinear sptial-temporal correlations can not be easily captured.

With the advancement of deep learning in computer vision [3] and natural language process [4], more and more researchers begin utilize these techniques to solve traffic prediction problems. Recent studies [5] treat the traffic in a city as an image and the traffic volume for a given time period as pixel value. With the image series, the output of the model is the image which represents traffic volume at the next timestamp. The complex spatial correlation is modeled by convolutional nerual network (CNN) [6]. Long Short Term Memory networks (LSTM) [7] is applied to predict loop sensor reading, which shows the ability of LSTM to capture the complex temporal dependency. The state-of-the-art model to predict taxi demand combines LSTM and local CNN to capture the complex nonlinear relations of both space and time and extract the semantic relations (e.g., traffic pattern similarity) between regions with the static region representations by the pre-training of graph embedding [8]. However, the static graph embedding is independent from the main prediction model, which may lead to the sub optimal results. Moreover, none of them consider the correlation between taxi pick up and taxi drop off.

In this paper, we harness the power of graph attention network (GAT) [9] and multi-task learning in an end-to-end network that models the semantic relations between regions dynamically and capture the correlation between taxi demand and taxi drop off.

2. PROBLEM DESCRIPTION
In this section, we fix some notations and define taxi demand prediction problem. The set of non-overlapping regions $R = \{R_1, R_2, ..., R_i, ..., R_N\}$ as gird partitions of an area. The

ICMAI 2020, April 10–13, 2020, Chengdu, China

ACM ISBN 978-1-4503-7707-2/20/04…$15.00
https://doi.org/10.1145/3395260.3395266

set of time intervals as $I=\{I_1,I_2,...,I_t,...,I_T\}$. The length of the time interval is 30 minutes.

2.1 Taxi Pick Up and Taxi Drop Off

A taxi pick up p is defined as a tuple $(p.t,p.r)$, where $p.t$ is the timestamp, $p.r$ represents the region. Similarly, a taxi drop off is defined as $(d.t,d.r)$.

2.2 Taxi Demand

The taxi demand is defined as the number of taxi pick up at one region per time interval, e.g., $y_t^i=\left|\{p:p.t\in I_t\wedge p.r\in R_i\}\right|$, where $|\cdot|$ denotes the cardinality of the set. The I_t and R_i are abbreviated as t and i respectively for simplicity in the rest of the paper.

2.3 Taxi Demand Prediction Problem

The taxi demand prediction problem is to predict the demand at time interval $t+1$ given the historical taxi demand data until time interval t. Notice that we can also merge additional external context features such as holiday, weekend, weather features. We denote the context features for a region i and a time interval t as $e_t^i\in \mathrm{R}^n$, where n is the number of features. The final goal is to predict

$$y_{t+1}^i=F(Y_{t-h,...,t}^R,\varepsilon_{t-h,...,t}^R)$$

For $i\in R$, where $Y_{t-h,...,t}^R$ are historical demands and $\varepsilon_{t\text{-}h,...,t}^R$ are context features for all regions R for time interval from t-h to t, where t-h represents the starting time interval. The prediction function $F(\cdot)$ denotes the taxi demand prediction model. To capture the correlation between taxi demand and drop off and improve the generalization of our model, we can also make prediction for y'^i_{t+1}, which denotes the drop off number of given region i at time interval $t+1$.

3. PROPOSED MODEL

In this section, we will illustrate details of our Multi-Task Spatial-Temporal Graph Attention Network (MSTGAT-Net). Figure 1 shows the framework of our model which includes four major components: local CNN, GAT, external context features, LSTM and multi-task learning.

3.1 Local CNN

The local CNN [10] is a simple variant of CNN, which only uses the spatially nearby regions to predict the target region. As illustrated in the part A of Figure 1, the target region i with its neighbors are treated as a single channel $S\times S$ image at time interval t. Notice that i is at the center of the image and pixel values represent the taxi demand. We denotes the image tensor as $\mathbf{Y}_t^i\in \mathrm{R}^{S\times S\times 1}$, for each region i and time interval t. The forward propagation of local CNN can be formulated as

$$\mathbf{Y}_t^{i,k}=f\left(\mathbf{Y}_t^{i,k-1}*\mathbf{W}_t^k+\mathbf{b}_t^k\right)$$

where $*$ denotes the convolutional operation and $f(\cdot)$ is activation function, such as ReLU [11]. $\mathbf{W}_t^k$ and $\mathbf{b}_t^k$ are the parameters of the k^{th} convolution layer at the time interval t. A flatten layer is applied after the last convolution layer to transform $\mathbf{Y}_t^{i,k}\in \mathrm{R}^{S\times S\times\lambda}$ to a vector $\mathrm{s}_t^i\in \mathrm{R}^{S^2\lambda}$. Finally, we feed s_t^i into a fully connected layer to reduce the dimension

$$\hat{\mathrm{s}}_t^i=f\left(W_t^{fc}s_t^i+b_t^{fc}\right)$$

where W_t^{fc} and b_t^{fc} are parameters of fully connected work. Finally, the extracted representation of region i at time interval t by local CNN is $\hat{\mathrm{s}}_t^i\in \mathrm{R}^d$.

3.2 GAT

As we discussed above, there are semantic correlations (e.g., traffic pattern similarity) between regions. The common way to capture these correlations is to represent regions as embeddings. For example, [10] takes regions as nodes, then calculates Dynamic Time Warping (DTW) similarity of taxi demand time series which is used as the weight of edges. Graph embedding methods is applied to get the embedings of regions. However, there may exist complex correlations between regions which can not be measured by DTW. Moreover, the static graph embedding may be no longer effective as time goes on. We propose that using graph attention network (GAT) [9] to address these problems as illustrated in the part B of Figure 1. GAT can capture the complex correlations between regions by multi-head attention mechanism [12], so we do not have to design the function (e.g., DTW) to measure the correlations in advance. We further apply GAT in each time interval to get the dynamic representations of regions instead of static graph embedding. We can formulate the representation of region i at time interval t extracted by graph attention as

$$\mathrm{g}_t'^i=\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in R}\alpha_{ij}^kW^k\mathrm{g}_t^j\right)$$

where K is the hyper parameter of GAT, which represents the number of head of multi-head attention. R is the regions set. α_{ij}^k is the normalized attention coefficients computed by the k^{th} attention mechanism. W^k is the parameter of k^{th} attention. g_t^j is the feature vector of region i at time interval t.

3.3 LSTM

LSTM has been proved to be efficient to capture the long-term dependency of time series. It learns the sequential correlations by recursively applying a transition function to the hidden state of the input, which can be formulated as

$$h_t^i=\mathrm{argmaxP}\left(h\mid c_{t\text{-}1},x_t^i\right)$$

where $c_{t\text{-}1}$ is the memory cell at time step t-1, x_t^i and h_t^i are the input and the hidden state for the region i at time step t, respectively. As illustrated in the part D of Figure 1, the LSTM component concatenates the outputs of local CNN, GAT and

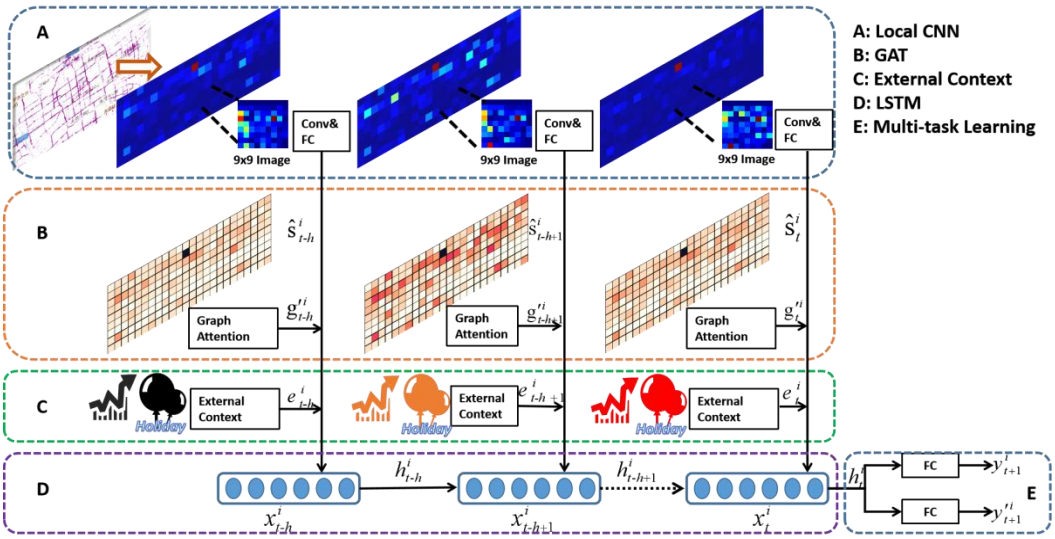

Figure 1. The Architecture of MSTGAT-Net.

external context features as x_t^i . Finally, we feed h_t^i into a two-layer fully connected network to get the prediction value.

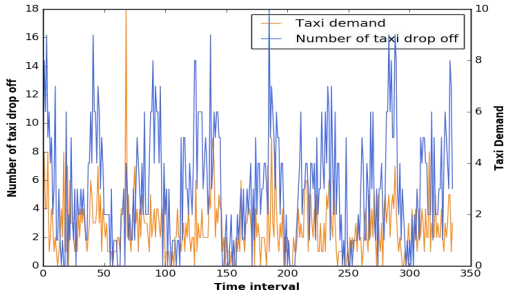

Figure 2. Correlation between taxi demand and number of taxi drop off

3.4 Multi-task Learning

Previous studies on taxi demand prediction mainly focus on the taxi demand itself and consider little about the correlation between taxi demand and drop off. As illustrated in Figure 2, we select a grid and a week randomly to show the change of them over the time. It is obvious that there are some correlations between taxi demand and drop off. Recently, multi-task learning has attracted more and more attention for its ability to improve the generalization of deep learning models [13]. Therefore, we adopt it to our model by predicting the taxi demand and the number of drop off simultaneously as illustrated in part E of Figure 1. We can formulate the multi-task learning by the following loss function

$$L(\theta)=\sum_{i=1}^{N}\ \alpha\times\left(y_{t+1}^{i}-\hat{y}_{t+1}^{i}\right)^2+\beta\times\left(y_{t+1}^{\prime i}-\hat{y}_{t+1}^{\prime i}\right)$$

where θ are all learnable parameters in the MSTGAT-Net. α and β are the weight for the taxi demand and drop off prediction loss, respectively. We use Adam [14] as the method of stochastic gradient descent.

4. EXPERIMENT RESULTS

4.1 Dataset Description

In this paper, we use a large-scale taxi dataset collected from Suzhou Industrial Park (SIP) , which is a leading city in China, by the traffic police department. The dataset contains taxi gps with pick up and drop off status from 01/01/2017 to 03/31/2017. There are 24×9 regions in our data. The size of each region is $0.7km\times0.7km$. The average taxi demand is 96425 each day. For fair, we only use holiday, weekend and the average demand value in the last four time intervals as external context features.

In the experiment, we train our model from 01/01/2017 to 02/28/2017 (59 days) and evaluate from 03/01/2017 to 03/31/2017 (31 days). The length of time interval is 30 minutes. The number of time intervals to predict the target value is 8. S for local CNN is 9 , which follows the setting of [10].

4.2 Evaluation Metric

We use Mean Average Percentage Error (MAPE) and Rooted Mean Square Error (RMSE) to evaluate our model, which are defined as

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}-y|}{y}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}-y)^2}$$

where $\hat{y}$ and y are prediction value and ground truth for the taxi demand. Notice that the drop off data is only used for the model training and we do not evaluate on it.

4.3 Baselines

We compared our model with the following methods.

Historical average (HA): HA predicts the taxi demand of each region by using average values of previous demands at the given region in the same relative time interval (e.g., the same time of the day).

Auto regressive integrate moving average (ARIMA) [15]: As is known to us, ARIMA combines moving average and auto regressive components to model time series.

Deep multi-view spatial temporal network (DMVST-Net) [10]: DMVST predicts the taxi demand with local CNN , LSTM and static graph embedding.

4.4 Performance Comparison

4.4.1 Comparison with state-of-the-art

Table 1 shows the performance of the proposed method as compared to all other competing methods. MSTGAT-Net achieves the lowest MAPE (0.1501) and the lowest RMSE (4.066) among all the methods, which is 12.88% (MAPE) and 25.50% (RMSE) relative improvement compared to DMVST-Net. As illustrated in Table 1, HA and ARIMA predict relative poorly (e.g., have a MAPE of 0.2792 and 0.2540, respectively), because they can not capture the complex correlations between regions and the long-term dependency in the historical demand data. DMVST-Net combines spatial-temporal features and static semantic correlations between regions with graph embedding to achieve better performance. As for our model, we further utilize GAT to model the deep dynamic latent semantic correlations between regions, while take multi-task learning to capture the correlation between the taxi demand and the number of taxi drop off. Notice that we neither need to design the function (e.g., DTW) to measure the correlations between regions in advance nor get the pre-training embedding for each region compared to DMVST-Net and we can train the model end-to-end.

Table 1. Comparison with different baselines

Method	MAPE	RMSE
Historical average	0.2792	14.161
ARIMA	0.2540	13.587
DMVST-Net	0.1723	5.458
MSTGAT-Net	**0.1501**	**4.066**

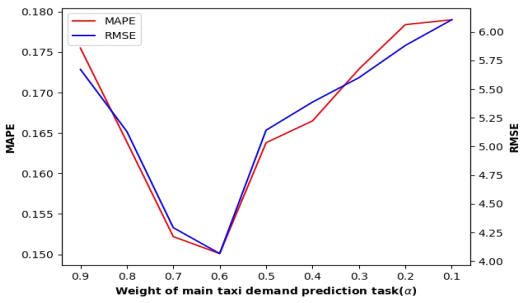

Figure 3. Impacts of the weights of multi-task learning.

4.4.2 Impacts of the weights of multi-task learning

Our intuition was that applying multi-task learning can capture the latent correlation between taxi demand and taxi drop off. We verified that intuition by vary the weights of multi-task loss. Specifically, we fixed $\alpha+\beta=1$ and change α from 1 to 0.1. A higher value of α indicates that we pay more attention to the taxi demand prediction task, which is the main task in our model. In Figure 3, we show the performance of method with respect to the weights of multi-task learning. We can see that when the α is 0.6 and β is 0.4, the model achieves the best performance. The taxi demand prediction increases as the α decrease to 0.6. Because the loss of auxiliary task (e.g., taxi drop off prediction) dominates the total loss when the α is too small, which leads to the over-fitting.

4.4.3 Case visualization of graph attention network

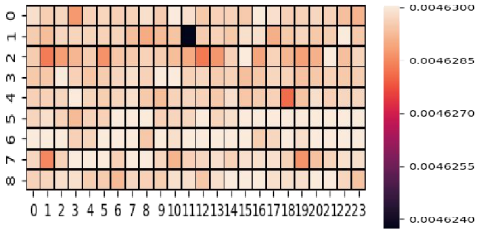

Figure 4. The visualization of graph attention network.

In this section, we visualize the attention learned by graph attention network. We pick a region and a timestamp randomly in the first head of multi-head attention. The value of grid (i,j) represents the attention weight α_{ij} , which can be considered as the intensity of the correlation between target region and region (i,j) . As illustrated in Figure 4, the orange grid means that there are strong latent correlations between regions. It is important to interpret the model to understand the deep learning better.

5. CONCLUSIONS

In this paper, we propose a novel Multi-Task Spatial-Temporal Graph Attention Network for taxi demand prediction (MSTGAT-Net). Our model combines spatial-temporal features and latent dynamic semantc correlations between regions, which are modeled by local CNN, LSTM and GAT,respectively. We further take multi-task learning to capture the correlation between the taxi demand and taxi drop off, which boosts the generalization of our model. The experiments on a large scale real dataset show that our proposed method performs significantly better compared to state-of-the-arts methods.

6. ACKNOWLEDGMENTS

Thanks to the great efforts and sincere suggestions of anonymous reviewers.

7. REFERENCES

[1] Moreira-Matias, L. , Gama, J. , Ferreira, M. , Mendes-Moreira, J. , & Damas, L. 2013. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3), 1393-1402.

[2] D. Wang, W. Cao, J. Li and J. Ye, DeepSD: Supply-Demand Prediction for Online Car-Hailing Services Using Deep Neural Networks, 2017. *IEEE 33rd International Conference on Data Engineering*, San Diego, CA, 2017, 243-254.

[3] Ren, S. , He, K. , Girshick, R. , & Sun, J. 2015. Faster r-cnn: towards real-time object detection with region proposal

networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6), 1137-1149.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Zhang, J., Zheng, Y., & Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. *In Thirty-First AAAI Conference on Artificial Intelligence*.

[6] LeCun, Y. 2015. LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 20, 5.

[7] Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[8] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. 2015. Line: Large-scale information network embedding. *In Proceedings of the 24th international conference on world wide web*, 1067-1077.

[9] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

[10] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. *In Thirty-Second AAAI Conference on Artificial Intelligence.*

[11] Nair, V., & Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. *In Proceedings of the 27th international conference on machine learning*, 807-814.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. 2017. Attention is all you need. *In Advances in neural information processing systems*, 5998-6008.

[13] Li, Y., Fu, K., Wang, Z., Shahabi, C., Ye, J., & Liu, Y. 2018. Multi-task representation learning for travel time estimation. *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1695-1704.

[14] Kingma, D. P., & Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[15] Box, G. E., & Pierce, D. A. 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332), 1509-1526.