Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220726

Multi-Head Attention Mechanism for Multi-Station Passenger Flow Prediction

Long Chen[a], Lijuan Liu[a, b, 1] and Lei Yuan[a]
[a] *College of Computer and Information Engineering, Xiamen University of Technology, Xiamen, 361024, China*
[b] *Fujian Key Laboratory of Pattern Recognition and Image Understanding, Xiamen, 361024, China*

Abstract. Passenger flow prediction is of great significance for public transportation. Most of the existing studies mainly predict the flow for a single station only extracting temporal features without considering spatial features. Passenger flow predicting for multiple stations or even the whole network is beneficial to grasp the overall situation, which has more research value in practical application. Thus, a passenger flow prediction model for multiple stations has been proposed based on spatio-temporal attention mechanism. This model is applied to Xiamen bus rapid transit (BRT) using the time granularity of 5-min. The experimental results demonstrate that our model improves the prediction accuracy compared to the baselines.

Keywords. attention mechanism, passenger flow, prediction, multi-station, spatio-temporal

1. Introduction

Passenger flow prediction is of great significance to the design and planning, operation management and early warning for public transportation. Accurately predicting passenger flow has always been the key work of traffic management departments, especially under the background of rapid growth of traffic capacity and the constant changing of passenger flow demand. Therefore, passenger flow prediction, especially short-term passenger flow prediction has gradually become a research focus, which can effectively improve the level of vehicle scheduling, passenger flow evacuation, and ensure the safety of citizens. It is of great significance to urban management, security personnel allocation and real-time passenger flow monitoring [1], which is an important basis for intelligent transportation system.

In the recent ten years, there are many abundant studies focusing on the short-term passenger flow prediction in different public transportation system, such as rail transit, ordinary bus, bus rapid transit (BRT), taxi, etc. However, most researchers prefer to choose a single station as the predict target since the historical passenger flow data is more regular and simpler, which ignores the strong dynamic correlation of flow data in spatial and temporal dimensions. The passenger flow dynamics at a station not only

[1] Corresponding author, E-mail: ljliu@xmut.edu.cn

depends on the sequential pattern in temporal dimension, but also depends on inbound and outbound passenger flow from the other stations in spatial dimension. Spatial correlation between different locations is highly dynamic, depending on real-time traffic conditions and network topology. Thus, there is still a large research space to exploit the spatio-temporal features. Compared with the research on passenger flow prediction for a single station, the research on passenger flow prediction for multiple stations in public transport has been more valuable in practical application.

Recently, there is a trend that more and more studies have been taken into account to extract both spatial-temporal dependence in the short-term passenger flow prediction for multiple stations or even the whole network. Zhao and Song (2020) proposed a temporal graph convolutional network (T-GCN) model combining the GCN and the gated recurrent unit (GRU) for traffic flow prediction [2]. Zhang and Chen (2021) developed a novel origin-destination (OD) flow predicting method that considers the unique characteristics of the passenger flow, which mainly consists of a channel-wise attention mechanism and split convolutional neural network (CNN) [3]. However, the above models are difficult to deep capture complex dynamic spatio-temporal dependencies due to the static graphs constructed in these models.

Attention mechanism is widely used in various fields because of high efficiency and flexibility in modeling dependencies. The core of attention mechanism is to adaptively focus on the most relevant features based on the original input data. More recently, researchers have applied the attention mechanism for graph structured data to model the spatial relevance. Multi-head attention mechanism could more abundantly to capture the features in different respective, which have been shown to be more suitable for processing time series. Therefore, the temporal and spatial correlation of passenger flow characteristics have been proposed to design a more accurate passenger flow prediction model for multiple stations based on multi-head attention mechanism in this paper.

2. Methodology

2.1. Overall Architecture

The proposed architecture is shown in Figure 1. It consists of two spatio-temporal blocks called ST-Block, spatio-temporal embedding and two fully connected layers. Specifically, each ST-Block consists of a temporal attention block and a spatial attention block to jointly extract the spatio-temporal features of passenger flow among multiple stations. Spatio-temporal embedding and passenger flow are used as the inputs to the ST-Block. Fully connected layer 1 is the input of model and fully connected layer 2 is to aggregate these spatio-temporal features for the final passenger flow prediction.

2.2. Problem Formulation

The topology of the station network is naturally represented as a graph $G = (V, E, A)$. Here, V is a set of $N = |V|$ vertices, representing all the stations; E is a set of edges representing the connectivity between any two stations; $A \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where A_{v_i,v_j} represents the proximity between V_i and V_j. The $X_t \in$

$\mathbb{R}^{N\times D}$ represents a graph signal at time step t on graph G, where D is the real passenger flow.

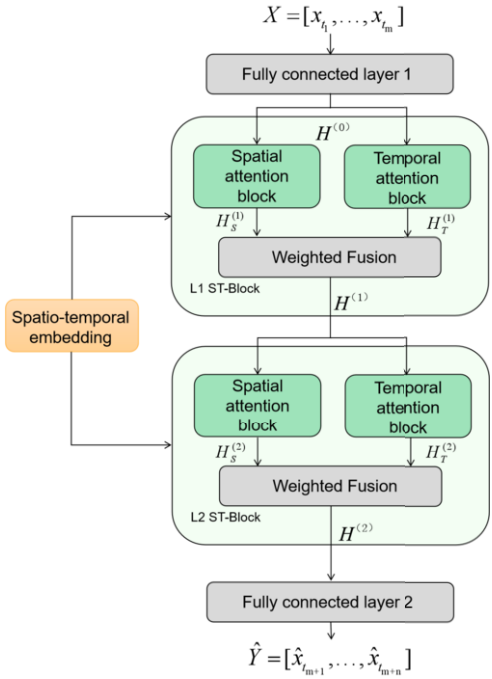

Figure 1. Overall Architecture.

Given the observations at N vertices of historical m time steps $X = [x_{t_1}, x_{t_2}, \ldots, x_{t_m}] \in \mathbb{R}^{m\times N\times D}$, a prediction model is learned to predict passenger flow in the future n time steps $\hat{Y} = [\hat{x}_{t_{m+1}}, \hat{x}_{t_{m+2}}, \ldots, \hat{x}_{t_{m+n}}]$.

2.3. Spatio-temporal Attention Block

In this section, we have developed a spatio-temporal attention block to integrate spatial attention and temporal attention, and jointly capture temporal and spatial features. As shown in Figure 1, the ST-Block includes a spatial attention part, a temporal attention part and a weighted fusion part. There are two such modules in our model. The input to the L1 ST-Block is a 3-dimension tensor. We denote the input of the L1 ST-Block as $H^{(0)} = \{h^{(0)}_{v_i,t_m}\}$. It is the hidden state of station v_i at time step t_m, which is represented as $h^{(0)}_{v_i,t_m}$. The outputs of spatial and temporal attention block in the L1 ST-Block are represented as $H_S^{(1)}$ and $H_T^{(1)}$. They are the hidden states of vertex v_i at time step, which are denoted as $hs^{(1)}_{v_i,t_m}$ and $ht^{(1)}_{v_i,t_m}$, respectively. We obtain the output of the L1 ST-Block represented as $H^{(1)}$. $H^{(2)}$ has the same processing with $H^{(1)}$.

2.3.1. Spation-temporal Embedding

The spatio-temporal dependencies between two stations would be determined by their connectivity and observed time steps. We use node2vec [4] to represent graph information as vectors for capturing spatial dependencies. The spatial embedding only provides static representations, which could not represent the dynamic correlations among stations. Therefore, we further propose a time embedding to encode each time step into a vector. Specifically, let a day have T time steps. We use a one-hot code to encode the week and time of each time step as $\mathbb{R}^7$ and $\mathbb{R}^T$, and connect them as vector $\mathbb{R}^{T+7}$.

$e_{v_i}^S \in \mathbb{R}^D$ and $e_{t_j}^T \in \mathbb{R}^D$ are learned as spatial and temporal embedding matrices, respectively, where $v_i \in V$ and $t_j = t_1, \ldots, t_{M,} \ldots, t_{M+N}$. These vectors are input into a two-layer fully connected neural network to transform the shape and adjust it to the same dimension as the passenger flow input data. We integrate the above spatial embedding and temporal embedding into spatio-temporal embedding. The spatio-temporal embedding is defined as $e_{v_i,t_j=}e_{v_i}^S + e_{t_j}^T$, which contains graph structure and time information, which has been combined with passenger flow data together as the input of ST-Block shown in Figure 2.

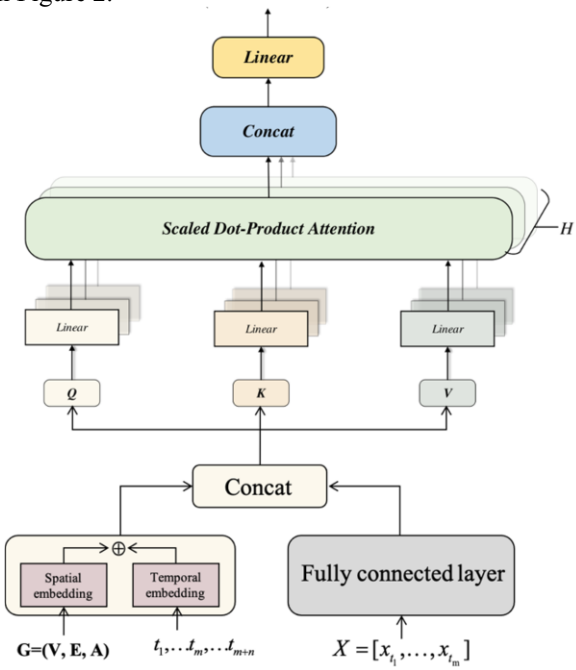

Figure 2. Spatial Attention Block.

2.3.2. Spatial Attention Block

We designed a spatial attention mechanism to capture the spatial correlation of station-level dynamic passenger flow, which is to extract the spatial features among multiple stations. The key idea is to dynamically assign different weights to different stations at different time steps, as shown in Figure 2. The attention mechanism, also called scaled dot-product attention [5], can be regarded as a function. In our model, a query vector

$Q \in \mathbb{R}^{N \times D_q}$, a key vector $K \in \mathbb{R}^{N \times D_k}$, and a value vector $V \in \mathbb{R}^{N \times D_v}$ are set to be consistent. Their inputs are the same time series data, and their expressions are shown in Eqs. (1)-(3), respectively.

$$Q = W_q \cdot X^t \tag{1}$$

$$K = W_k \cdot X^t \tag{2}$$

$$V = W_v \cdot X^t \tag{3}$$

Where W_q, W_k and W_v denote the weight matrices for Q, K, V respectively, X^t is the model input and its expression shown in Eq. (4).

$$X^t = \langle h_{v_i,t_j}^{(l-1)} || (e_{v_i,t_j}, h_{v_i,t_j}^{(l-1)}) || e_{v_i,t_j} \rangle \tag{4}$$

Where || represents the concatenation operation, $\langle \quad \rangle$ denotes the inner product operator. After getting Q, K, V, we can calculate the spatial dependencies Z by dot-product shown in Eq. (5).

$$Attention(Q, K, V) = Z = softmax\left(\frac{QK^T}{\sqrt{d^K}}\right)V \tag{5}$$

Where the softmax layer makes non-linear changes to the input so that the output is between [0, 1], and d_K denotes the columns of Q, K matrix, which is used to avoid the softmax to make the gradient too small to backpropagate.

2.3.3. Temporal Attention Block

Time correlation is affected by historical passenger flow in different time steps. For example, the passenger flow in the next time step will be influenced by the passenger flow in current time step, even several ahead time steps. We consider the flow characteristics and time to measure the correlation between different time steps. Specifically, we connect the hidden state with spatio-temporal embedding, and use the multi-head method to calculate the attention score. The input of the temporal attention block is consistent with that of the spatial attention block. We adjust the dimension of the time module and put the time step in the first dimension to extract the context information.

2.3.4. Weighted Fusion

We design a weighted fusion mechanism to combine the output of spatial attention block ($H_S^{(l)}$) and temporal attention block ($H_T^{(l)}$). $H_S^{(l)}$ and $H_T^{(l)}$ are fused as Eqs. (6)-(7), respectively.

$$H^{(l)} = \alpha \odot H_S^{(l)} + (1 - \alpha) H_T^{(l)} \tag{6}$$

$$\alpha = \sigma(H_S^{(l)} W_{\alpha,1} + H_T^{(l)} W_{\alpha,2} + b_\alpha) \tag{7}$$

Where $W_{\alpha,1} \in \mathbb{R}^{D\times D}$, $W_{\alpha,2} \in \mathbb{R}^{D\times D}$ and $b_\alpha \in \mathbb{R}^D$ are learnable parameters, $\odot$ represents the dot product, $\sigma()$ denotes the activation function of sigmoid.

3. Experiment

3.1. Data Description

The used dataset in this paper is from automatic fare collection (AFC) in Xiamen BRT from March 1, 2019 to May 31, 2019. A total of 44 stations have been collected, and the statistical period is 5 minutes from 6:30 am to 10:10 pm. We only used the historical inbound passenger flow data to make prediction for the future 5 min inbound passenger flow for the 44 stations. After outlier processing, there are totally 17,296 pieces of data. Table 1 shows the specific data recorded in the original dataset. Each sample includes station, time, and passenger flow.

Table 1. Original Dataset.

Station	Time	Passenger Flow
First Pier Station	06:10-06:15, March 1, 2019	2
Kaihelukou Station	15:15-15:20, March 10, 2019	55
Xiamen North Railway Station	16:20-16:25, March 10, 2019	57

3.2. Model Specification

We use 70% of the data for training, 10% for validation, and 20% for testing, the Adam optimizer is adopted, the number of iterations is 100, the batch size is 50, the learning rate is 0.001, and the number of the head of attention mechanism is 8. The time granularity is 5 minutes, and the time step of a day is 188.The computer used in the experiment is configured as GPU (Nvidia GTX 1080), and CPU (Intel Xeon E5).

The evaluation indexes include root mean square error (RMSE) and mean absolute error (MAE). MAE and RMSE are used to describe the prediction error. The smaller the results, the closer to the true values, and the better the model fitting. The MAE and RMSE are calculated in Eqs. (8)-(9), respectively.

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|h_i - \hat{h}_i| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(h_i - \hat{h}_i)^2} \tag{9}$$

Where h_i refers to real passenger flow, $\hat{h}_i$refers to the predicted passenger flow, and M refers to the number of testing dataset.

3.3. Experimental Results

To evaluate the performance of our proposed model, we compared it with several popular models, including graph convolutional network (GCN) [6], graph attention network (GAT) [7] and spatio-temporal graph convolutional network (STGCN) [8]. GCN and its variants are widely considered to capture the spatial dependencies of traffic flows to explore the inherent traffic topology. A weighted sum of adjacent node features has been

proposed using attention mechanism in GAT. The weights of adjacent node features depend entirely on node features and are independent of graph structure. Graph convolutional layers and convolutional sequence learning layers have been combined in STGCN.

Table 2. Experiment Results.

Model	MAE	RMSE
GAT	13.05	19.93
GCN	6.63	10.65
STGCN	6.49	10.12
Ours	5.35	8.08

Table 2 shows the performance of our multi-head attention model compared to the above benchmark models. All models use the historical one-hour ridership to calculate the future 5-min passenger flow. Here, one-hour refers to use the inbound passenger flow of 12 timesteps inbound passenger flow with 5-min for one timestep. In the three compared benchmark models, GAT and GCN only capture the spatial dependencies. STGCN captures both temporal and spatial dependencies. As we expected, our proposed model achieved the best results in Xiamen BRT with MAE and RMSE of 5.35 and 8.08, respectively.

4. Conclusions

A multi-station passenger flow prediction model based on multi-head attention mechanism has been proposed to improve the prediction accuracy in Xiamen BRT. It can dynamically capture the deep spatio-temporal features for passenger flow among multiple stations with 8-head attention mechanism. Experimental results on real datasets show that the proposed model has the superior performance, especially in spatio-temporal modeling. Because Xiamen BRT adopts the closed viaduct mode, the proposed model could be easily applied in metro passenger flow prediction task. In the future, we will focus on the long-term passenger flow prediction in spatio-temporal modeling, such as the next one hour, and verify our model on more real datasets in different public transport, such as metro system.

Acknowledgement

This work was partly supported in part by the National Natural Science Foundation of China (No. 62103345), Fujian Provincial Natural Science Foundation of China (No. 2020H0023, No. 2020J02160), Xiamen Youth Innovation Fund Project (No.3502Z20206076), and the High-Level Talents Research Launched Project of Xiamen University of Technology (Grant No. YKJ19012R).

References

[1] J. J. Li, L. Z. Xu, L. Tang, S. Y. Wang, and L. Li, Big data in tourism research: A literature review, Tourism Management 68(10) (2018), 301-323.

[2] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, T-GCN: a temporal graph convolutional network for traffic prediction, IEEE Transactions on Intelligent Transportation Systems

21(9) (2020), 3848-3858,
[3] J. L. Zhang, and H. S. Che, Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method, Transportation Research Part C-Emerging Technologies 124 (2021), 1-20.
[4] A. Grover, and Leskovec, Node2vec: Scalable Feature Learning for Networks, Knowledge Discovery and Data Mining (2016), 855-864.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762, 2017.
[6] TN. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907, 2016.
[7] P. V eli˘ckovi´c et al., Graph attention networks, International Conference on Learning Representations, Vancouver, BC, Canada, 2017.
[8] B. Yu, H. Yin, and Z. Zhu, Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, International Joint Conference on Artificial Intelligence, 2018.