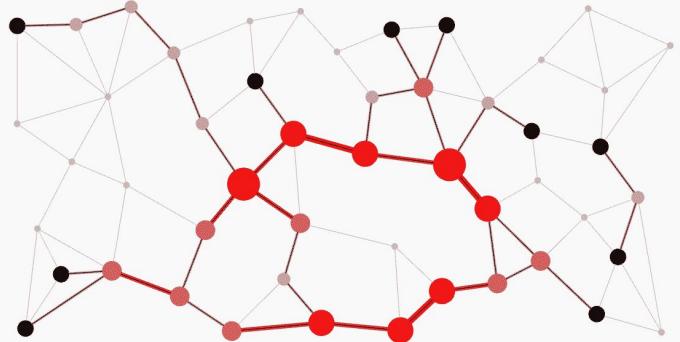


Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, Huaiyu Wan



[Image source]

The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting

Shengnan Guo,^{1,2} Youfang Lin,^{1,2,3} Ning Feng,^{1,4} Chao Song,^{1,2} Huaiyu Wan,^{1,2*}

¹School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

²Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing, China

³CAAC Key Laboratory of Intelligent Passenger Service of Civil Aviation, Beijing, China

{guoshn, yflin, fengning, chaosong, hywan}@bjtu.edu.cn

Abstract

Forecasting the traffic flows is a critical issue for researchers and practitioners in the field of transportation. However, it is very challenging since traffic flows usually show high nonlinearities and complex patterns. Most existing traffic flow prediction models can only capture the static spatial correlations of traffic data, thus cannot yield accurate predictions. In this paper, we propose a novel attention-based spatial-temporal graph convolutional network (ASTGCN) model to solve traffic flow forecasting problem. The proposed model decomposes the traffic flow into three independent components to respectively model three temporal properties: daily periodicity, weekly periodicity and weekly-periodic dependencies. More specifically, each component plays a major part in the spatial-temporal attention mechanism to better capture the spatial-temporal correlations in traffic data. 2) the spatial-temporal convolutional layer is used to capture the spatial-temporal correlations in traffic data. 3) the spatial-temporal convolutional layer is used to capture the spatial-temporal correlations in traffic data. The three components are weighted fused to generate the final prediction results. Experiments on two real-world datasets from the California Performance Measurement System (CPMS) demonstrate that the proposed ASTGCN model outperforms the state-of-the-art baselines.

Introduction

Recently, many countries are committed to vigorously develop the Intelligent Transportation System (ITS) (Zhang et al. 2011) to help for efficient traffic management. Traffic flow forecasting is one of the most important parts of ITS, the highway which has large traffic flows and fast driving speed. Since the highway is relatively closed, once a congestion occurs, it will spread rapidly. Therefore, accurate traffic flow forecasting is a fundamental measurement reflecting the state of the highway. If it can be predicted accurately in advance, according to the prediction results, traffic authorities will be able to guide vehicles more reasonably to enhance the running efficiency.

Highway traffic flow forecasting is a typical problem of spatial-temporal data forecasting. Traffic data are recorded at fixed points in time, and the data are distributed

*Corresponding author. Email: hywan@bjtu.edu.cn

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

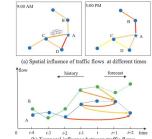


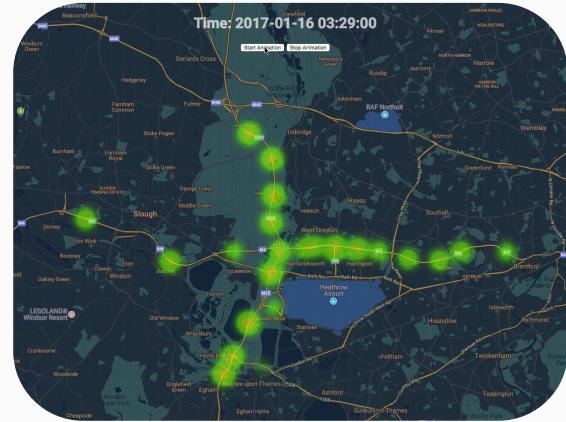
Figure 1: The spatial-temporal correlation diagram of traffic flow

in a continuous space. Apparently, the observations made at neighboring locations at same timestamp are independent but dynamically correlated with each other. Therefore, the key to solving the traffic flow forecasting problem is to mine the spatial-temporal correlations of data. Fig. 1 demonstrates the spatial-temporal correlations of traffic flows (also can be vehicles). The lines connecting the two locations between two points represents their mutual influence strength. The darker the color of line is, the greater the influence is. In the spatial influence (Fig. 1(a)), the traffic flows at different locations have different impacts on A and even a same location has different impacts on A at different times. In the temporal dimension (Fig. 1(b)), the historical observations of different locations have varying impacts on A's traffic states at different times. Therefore, the spatial-temporal correlations in traffic data on the highway network show strong dynamics in both the spatial dimension and temporal dimension. How to effectively mine the spatial-temporal correlations in traffic data to discover its inherent spatial-temporal patterns and to make accurate traffic flow forecasting is a very challenging issue.

Presently, with the development of the transportation industry, many cameras, sensors and other information collection devices have been deployed on the highway. Each

Motivation

- Intelligent Transportation System (ITS)
- Forecasting traffic flows are essential part of ITS
- Existing methods lack ability to capture dynamic correlations in spatial-temporal (ST) traffic data
- Key to solve the forecasting challenge lies in extracting ST correlations effectively from the data

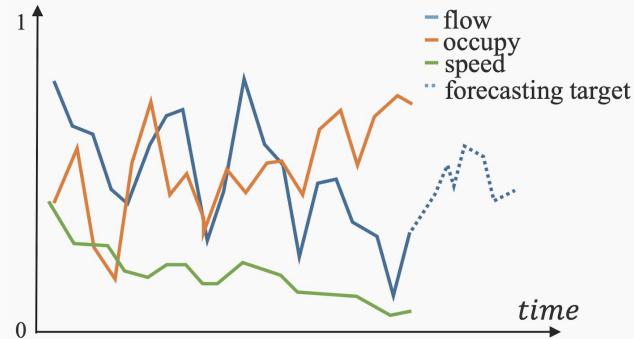
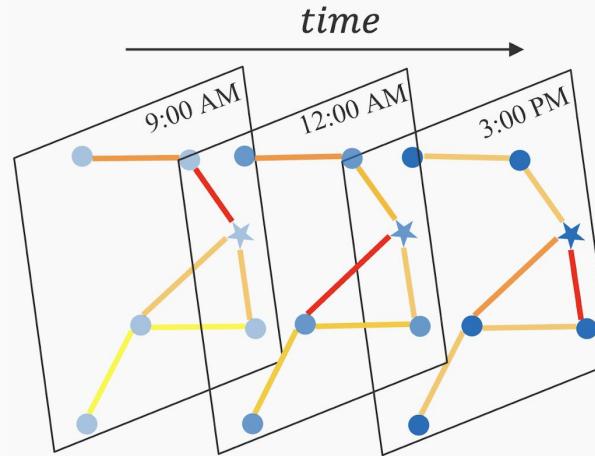


[[Image source](#)]

Preliminaries

Traffic Networks

- Traffic data is recorded
 - at fixed points in time
 - at fixed locations
- $G = V, E, A$
- Every sensor detects F features
 - at equal sampling frequencies



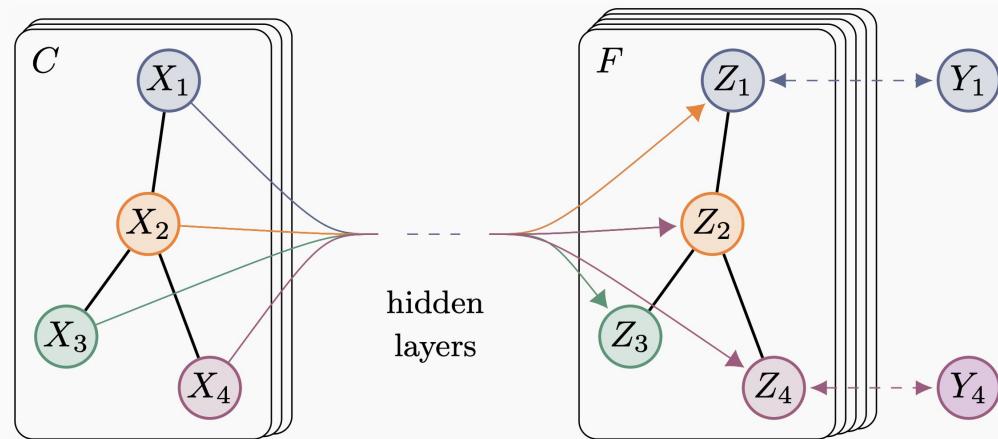
Related work

- **Statistical methods**
 - HA
 - ARIMA
 - VAR
- **Classical Machine Learning**
 - KNN
 - SVM
- **Deep Learning**
 - ST-ResNet
 - GeoMAN
 - DMVST-Net



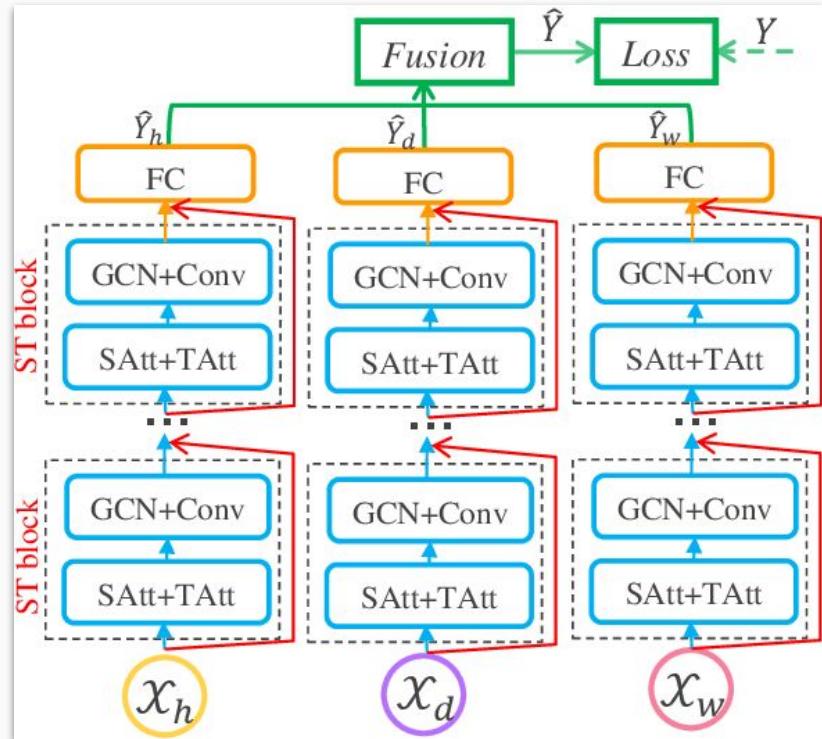
Graph Convolutional Networks

- Draws on the same idea as CNNs
 - Capture neighbourhood information for graph nodes
- Spatial
 - Aggregate neighbours feature information
- Spectral
 - Spectral graph analysis
 - Laplacian
 - $L = D - A$



Architecture

- Three components which are fused
 - Hourly
 - Daily
 - Weekly
- *Spatial-Temporal Blocks*
 - Spatial Attention + Temporal Attention
 - GCN + Conv
 - Fully Connected Layer
 - MSE as loss functions



Spatial- and Temporal Attention

Spatial attention mechanism

- Model complex spatial correlations between the different locations

$$\mathbf{S} = \mathbf{V}_s \cdot \sigma((\mathcal{X}_h^{(r-1)} \mathbf{W}_1) \mathbf{W}_2 (\mathbf{W}_3 \mathcal{X}_h^{(r-1)})^T + \mathbf{b}_s)$$

$$\mathbf{S}'_{i,j} = \frac{\exp(\mathbf{S}_{i,j})}{\sum_{j=1}^N \exp(\mathbf{S}_{i,j})}$$

$\mathbf{V}_s, \mathbf{b}_s \in \mathbb{R}^{N \times N}, \mathbf{W}_1 \in \mathbb{R}^{T_{r-1}}, \mathbf{W}_2 \in \mathbb{R}^{C_{r-1} \times T_{r-1}}, \mathbf{W}_3 \in \mathbb{R}^{C_{r-1}}$

Temporal attention mechanism

- Capture the dynamic temporal correlations between different time slices

$$\mathbf{E} = \mathbf{V}_e \cdot \sigma(((\mathcal{X}_h^{(r-1)})^T \mathbf{U}_1) \mathbf{U}_2 (\mathbf{U}_3 \mathcal{X}_h^{(r-1)}) + \mathbf{b}_e)$$

$$\mathbf{E}'_{i,j} = \frac{\exp(\mathbf{E}_{i,j})}{\sum_{j=1}^{T_{r-1}} \exp(\mathbf{E}_{i,j})}$$

where $\mathbf{V}_e, \mathbf{b}_e \in \mathbb{R}^{T_{r-1} \times T_{r-1}}, \mathbf{U}_1 \in \mathbb{R}^N, \mathbf{U}_2 \in \mathbb{R}^{C_{r-1} \times N}, \mathbf{U}_3 \in \mathbb{R}^{C_{r-1}}$

Spatial-Temporal Convolution

- **Graph Convolutions**
 - Capture neighbouring information for each node on the graph
 - **Temporal Convolution**
 - Extract temporal dependencies from nearby time slices

Graph convolution in spatial dimension The spectral graph theory generalizes the convolution operation from the grid-based data to graph structure data. In this study, the traffic network is a graph structure in nature, so the features of each node can be regarded as the signals on the graph. Thus, we adopt graph convolutions based on the spectral theory to directly process the signals in the spatial dimension. The spectral method transforms a graph into an eigenbasis form to analyze the graph structures. Specifically, the connectivity of the graph structure is analyzed by the corresponding Laplacian matrix, a graph signal processing technique.

where $T_2(x) = 1$, $T_1(x) = x$. Using approximate expansion of Chebyshev polynomial to solve this formulation corresponds to extracting information of the surrounding nodes to the $(k-1)^{th}$ order neighbor centered on each node g_{ij} . The graph convolution graph by the convolutional Linear Unit (ReLU^U) as the final module uses $\text{ReLU}^U(g_{ij}(x))$.

graph Laplacian matrix as $\hat{s} = U^T \hat{L} U$. According to the properties of the graph Fourier transform of the corresponding inverse Fourier transform is $\hat{x} = U^{-1} \hat{s} U$. The convolution operation is a convolution operator that diagonalizes the graph Laplacian matrix (Hornik, Bruna, and LeCun 2015). Based on the graph signal x on the graph G is filtered by a kernel g :

$$g \circ *_G x = g(L)x = g(UAU^T)x = U(g(A)U^T)x \quad (5)$$

where $g \circ _G$ denotes a graph convolution operation. Since the convolution operation of the graph signal is equal to the product of these signals which have been transformed to the spectral domain by graph Fourier transform (Sinošek and Komadomský 2017), the inverse transform can be applied as $(g \circ _G x)^{-1} = g^{-1}(x)$. First, the graph signal x is transformed to the spectral domain by graph Fourier transform, then multiplying their transformed version and finally applying their inverse transform to get the result of the convolution operation. To handle it, if the spectral domain of the graph signal is large, then performing the eigenvalue decomposition on the graph signal is difficult. Therefore, we propose to directly perform the eigenvalue decomposition on the graph signal. This approach is called Chebyshev polynomials approximation, which is often used in graph signal processing (Sinošek and Komadomský 2017).

$$g_{\theta} *_G x = g_{\theta}(\tilde{\mathbf{L}})x = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\mathbf{L}})x \quad (6)$$

where the parameter $\theta \in \mathbb{R}^K$ is a vector of polynomial coefficients, $\mathbf{L} = \lambda_{max}^{-2} \mathbf{L} - I_N$, λ_{max} is the maximum eigenvalue of the Laplacian matrix. The recursive definition of the Chebyshev polynomial is $T_k(x) = 2xT_{k-1}(x) - T_{k-2}$.

where the parameter $\theta \in \mathbb{R}^K$ is a vector of polynomial coefficients, $\mathbf{L} = \lambda_{max}^{-2} \mathbf{L} - I_N$, λ_{max} is the maximum eigenvalue of the Laplacian matrix. The recursive definition of the Chebyshev polynomial is $T_k(x) = 2xT_{k-1}(x) - T_{k-2}$.

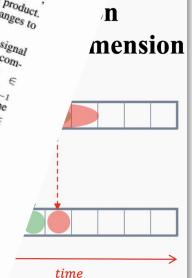
926

where $*$ denotes a standard convolution operation and the activation function is ReLU . In conclusion, a C_t -dimensional trace of r is able to work as an example:

Multi-Component Fusion
In this section, we will discuss how to fuse the three components, based on the whole trajectory. For example, we can use a spatial-temporal convolution module to capture the spatial-temporal features and a spatial-temporal attention module to stack the spatial-temporal feature maps. A spatial-temporal convolution module performs a spatial-temporal convolution on convolution kernel Φ to generate the fused features. The fused features are then passed through a spatial-temporal attention module to obtain the final output. This process is repeated for all three components. Finally, the outputs of all three components are concatenated to form the final output. The final output is then passed through a fully connected layer to predict the target action.

Multi-Component Fusion

...er uses RSL-2 as the activation function for the forecasting target. T1



In- and output

tures of all the nodes at time t . $\mathbf{\mathcal{X}} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_\tau)^T \in \mathbb{R}^{N \times F \times \tau}$ denotes the value of all the features of all the nodes over τ time slices. In addition, we set $y_t^i = x_t^{j,i} \in \mathbb{R}$ to rep-

$$\hat{\mathbf{Y}} = \mathbf{W}_h \odot \hat{\mathbf{Y}}_h + \mathbf{W}_d \odot \hat{\mathbf{Y}}_d + \mathbf{W}_w \odot \hat{\mathbf{Y}}_w$$

$$MSE = \frac{1}{n} \sum \underbrace{(y - \hat{y})^2}_{\text{The square of the difference between actual and predicted}}$$

Dataset

- Caltrans Performance Measurement System (PeMS)
- PeMSD4
 - 3848 detectors on 29 roads
 - Used **307** detectors
- PeMSD8
 - 1979 detectors on 8 roads
 - Used **170** detectors

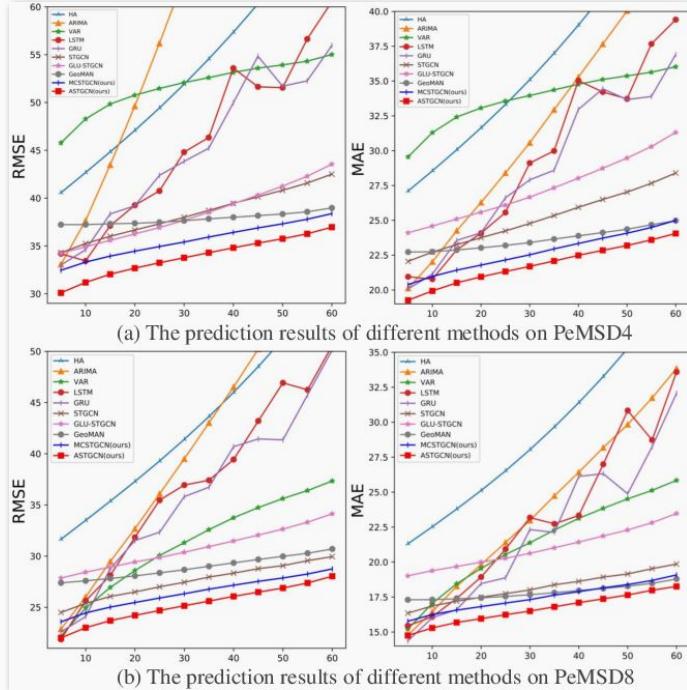


[[Image source](#)]

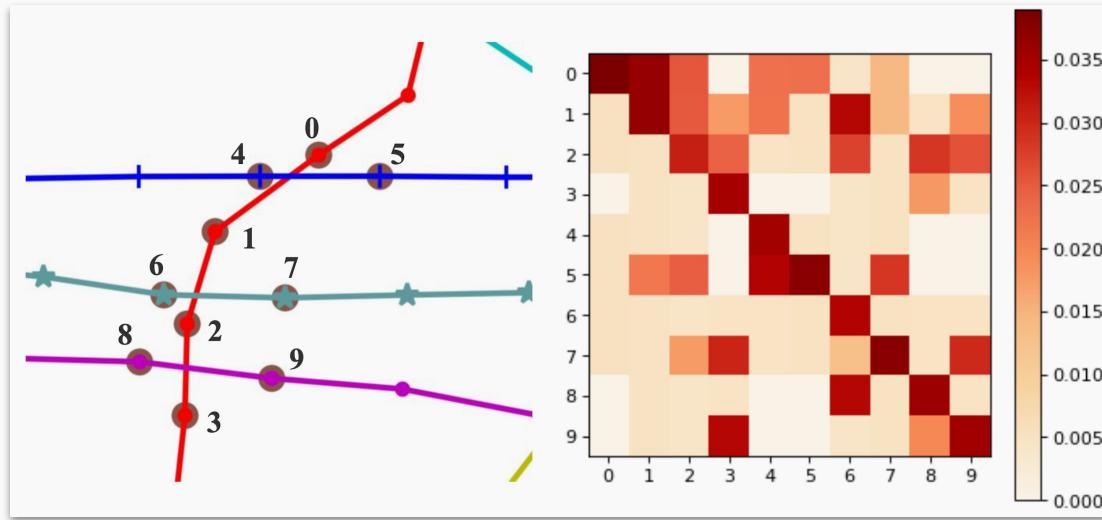
Experimental setting and results

Model	PeMSD4		PeMSD8	
	RMSE	MAE	RMSE	MAE
HA	54.14	36.76	44.03	29.52
ARIMA	68.13	32.11	43.30	24.04
VAR	51.73	33.76	31.21	21.41
LSTM	45.82	29.45	36.96	23.18
GRU	45.11	28.65	35.95	22.20
STGCN	38.29	25.15	27.87	18.88
GLU-STGCN	38.41	27.28	30.78	20.99
GeoMAN	37.84	23.64	28.91	17.84
MSTGCN (ours)	35.64	22.73	26.47	17.47
ASTGCN (ours)	32.82	21.80	25.27	16.63

Table 1: Average performance comparison of different approaches on PeMSD4 and PeMSD8.



Interpretability



Conclusions and thoughts

- A novel *Attention Based Spatial-Temporal GCN* has been proposed
- Superior to existing models at the time
- I really enjoyed researching the topic
- Not relevant for my thesis directly
- Code is available on GitHub
 - [MxNet implementation](#)
 - [PyTorch implementation](#)
- Found Graph Neural Networks very interesting



Thanks!

