Modeling Spatiotemporal Geographic-Semantic Dynamics for Urban Hotspots Prediction

Guangyin Jin
College of System Engineering, National University of Defense Technology
Changsha, China
jinguangyin@nudt.edu.cn

Hengyu Sha
College of System Engineering, National University of Defense Technology
Changsha, China
s24271722@163.com

Yanghe Feng
College of System Engineering, National University of Defense Technology
Changsha, China
fengyanghe@yeah.net

Qing Cheng
College of System Engineering, National University of Defense Technology
Changsha, China
sgggps@163.com

Jincai Huang
College of System Engineering, National University of Defense Technology
Changsha, China
huangjincai@nudt.edu.cn

ABSTRACT

Urban hotspots spatiotemporal prediction is a long-term but challenging task for urban management and smart cities construction. Accurate urban hotspots spatiotemporal prediction can improve urban planning, scheduling and security capability, reduce resource consumption. Existing deep spatiotemporal prediction methods mainly utilize geographic grid based image, some given network structure or some additional data to capture spatiotemporal dynamics. However, we observed that mining some latent self-semantics from raw data and fusing them with geospatial based grid images can also improve the performance of spatiotemporal predictions. In this paper, we propose Geographic-Semantic Ensemble Neural Network (GSEN), a novel deep learning approach to stack geographical prediction neural network and semantical prediction neutral network. GSEN model integrates structures of Predictive Recurrent Neural Network (PredRNN), Graph Convolutional Predictive Recurrent Neural Network (GC-PredRNN) and Ensemble Layer to capture spatiotemporal dynamics from different views. And this model can also be correlated with some latent high-level dynamics in real-world without any additional data. We evaluate our proposed model on three different domains real-world datasets and the experimental studies demonstrate generalization and effectiveness of GSEN in different urban hotspots spatiotemporal prediction tasks.

Deepspatial '20, August 24th, 2020, San Diego, California, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → **Geographic information systems**.

KEYWORDS

Spatiotemporal prediction, Semantic modeling, Predictive recurrent neural network, Graph convolutional neural network

ACM Reference Format:
Guangyin Jin, Hengyu Sha, Yanghe Feng, Qing Cheng, and Jincai Huang. 2020. Modeling Spatiotemporal Geographic-Semantic Dynamics for Urban Hotspots Prediction. In *Deepspatial '20: 1st ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data, Applications, and Systems, August 24th, 2020, San Diego, Claifornia, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

According to the urban development report by World Bank [1], the process of urbanization is becoming more and more rapid. Globally, over 50% of the population lives in urban areas today. By 2045, the world's urban population will increase by 1.5 times to 6 billion. The rapid urbanization process has brought a huge amount of data to be processed. We can discover some interesting patterns to guide urban planning and decision-making in many domains based on urban big data, for instance, traffic management, environmental monitoring, urban safety and so on. This kind of technology is the foundation for future smart cities construction, and we define it as urban computing, whose concept was proposed by Yu Zheng et al in 2014 [1]. For understanding complex spatiotemporal dynamics in urban systems, spatiotemporal prediction is one of the most basic but significant loop in urban computing [2].

[1] https://www.worldbank.org/en/topic/urbandevelopment

In general, the most common spatiotemporal prediction technology is expected to predict the urban hotspots distribution in spatial and temporal scales. In spatial perspective, some inner interactions between different urban regions such as human flow, traffic flow, could contribute a lot to urban hotspots distribution. In temporal perspective, some event hotspots in some urban regions reveal a cyclical or even seasonal fluctuation pattern. Hence, the biggest challenge is to capture the spatial and temporal correlations from complex dynamics simultaneously. To address this challenge, many advanced methodologies were proposed, which can be divided into three main categories, stochastic process based, statistical learning based, deep learning based, respectively. Compared with deep learning based approach, the stochastic process based and statistical learning based approaches are more explainable and elegant in mathematical form but some high-level nonlinear spatiotemporal correlations cannot be easily captured. The successful application of two neutral network models, Convolutional Neutral Network (CNN) [3] and Recurrent Neutral Network (RNN) [4] in image recognition and time series processing respectively has promoted the development of deep learning in spatiotemporal prediction. In addition, with the advent of the era of big data, abundant urban sensing data provides a solid foundation for deep learning. Hence, deep learning based spatiotemporal prediction approach has become much more popular in recent years. In the existing works, the most common data preprocessing method is to equally divide the urban area into m*n small grid according to the latitudes and longitudes. The hotspot statistics in each grid can be seen as the pixel values of the Spatial Hotspot Map (SHM) and the temporal SHMs are initial inputs of the CNN-based models [5]. This method is intuitive and the high-level geospatial correlations can be captured easily by CNN models in this way. However, there are still some limitations in previous works. (a) The hotspots distribution of geo-grid space only represents spatial correlations in Euclidean space but ignores some deeper non-Euclidean inner correlations. As we know, there are many different functional regions in urban area such as residential, commercial, industrial regions, and there are some implicit long-term correlations and pattern similarities among these regions, even they are not geographically adjacent. (b) To improve the accuracy of the prediction models, the traditional method is to fuse more additional relevant information. However, some additional data is hard to access and the auxiliary information only works in certain domain. This means the transferability and extensibility of the model is limited.

To address all these limitations above, we proposed Geographic Semantic Ensemble Neural Network (GSEN), which integrates geographic information and self-semantic information without any external auxiliary information. In geographic perspective, we initially arrange the spatiotemporal data as SHMs while in semantic perspective we model the Spatial Self-Semantic Graphs (SSGs) by spatial correlation coefficients. Based on the spatiotemporal LSTM variation, Predictive Recurrent Neural Network (PredRNN) [6], Graph Convolutional Predictive Recurrent Neural Network (GC-PredRNN) model is proposed in this paper for processing SSGs. Then, Predictive PredRNN and GC-PredRNN are applied respectively to capture the spatiotemporal correlations from temporal SHMs and SSGs. Subsequently, the geographic-based and semantic-based predictions are fused by ensemble convolution layers. The latent correlations between geographic space and semantic space are established in this model, and the limitations of traditional deep learning model can be overcome by involvement of SSGs.

To summarize, we make the following main contributions in this paper:

- To the best of our knowledge, it is the first exploration to propose a universal geographic-semantic information fusion framework to address multi urban hotspots spatiotemporal prediction tasks.
- Our proposed model GSEN is a soft computational framework without any other additional data. This promote its transferability and extensibility compared with some additional-data-dependent frameworks.
- We evaluate our approaches on three real urban datasets. The results show that our approach reduces the almost all evaluation metrics error compared to traditional state-of-art methods, which demonstrates that GSEN model has superiority and universality in urban hotspots spatiotemporal prediction.

2 RELATED WORK

Early methods focus on spatiotemporal prediction were based on some mathematical modeling approaches, especially stochastic process modeling and time series modeling. Self-exciting point process is a kind of point stochastic process, which have been widely applied in different fields where spatio-temporal events can be observed [7–9]. Based on the classical time series model ARIMA, many variant models were also introduced in many different domains [10–12]. The advantage of mathematical modeling approach is that some real-word spatiotemporal dynamics can be approximated by some explainable mathematical forms. But the disadvantage is also obvious, the mathematical model must be rebuilt in the different environments, even for the same event type, and this greatly limits the scalability of the models.

Compared with the methods based on mathematical modeling, the spatio-temporal prediction methods based on statistical learning is more convenient, and do not require repetitive modeling for the same type of task. In traffic prediction, some variants of SVM and Random Forest methods were involved to predict short-term traffic flow [13–15]. In environmental forecasting, many improved methods based on Ensemble Decision Trees, Random Forest have been proposed [16, 17]. Although statistical learning models have achieved some breakthroughs in spatiotemporal prediction, it is still difficult for them to capture high-level spatial and temporal dynamics synchronously.

With advent of the era of big data and the breakthroughs on deep learning models in image recognition and sequence

learning respectively, the deep learning based spatiotemporal prediction methods have gradually become one of the most mainstream methods in recent years. Xingjian et al proposed Conv-LSTM model [18], which was the first attempt to combine deep learning model CNN with RNN in precipitation prediction. Since then, many spatiotemporal prediction methods based on such hybrid deep model have been proposed in many different domains, for instance, crime prediction, traffic prediction and environmental monitoring [19–24]. Based on Conv-LSTM framework, Youngjoo et al proposed Graph Convolutional recurrent Neural Network (GCRNN) [25], which is similar to Conv-LSTM in their structures but the convolution operation is replaced by graph convolution operation. Compared with convolution operation, graph convolution operation has natural advantage in processing real-world structured data, especially in traffic domain [26–28]. To improve the accuracy of spatiotemporal prediction, some information fusion approaches were applied. Some proposed works fuse some multi inner correlation information without additional information. In work [29], correlations in different crime types were considered and different crime hotspot maps were fused to predict future crime situation. The citywide flow prediction deep model ST-ResNet [30], proposed by Yu Zheng, information from different time scales was considered and fused together. In taxi demand prediction model DMVST-Net [31] and OD matrix prediction model GEML [32], some latent semantic information in raw data was extracted and fused with raw geographic information. Some other works reflect the prediction authenticity of real-world by involving external multi-source information. Some deep models fuse some POI information, for instance, ST-MGCN proposed by Xu Geng et al [28] and DeepCrime proposed by Chao Huang et al [33] are multiple information fusion models which integrates additional information POI vectors.

Although some external information fusion deep models improve prediction performance, they sacrificed the models transferability and scalability, and the acquisition of external data and greater computational burden have brought bottlenecks to the model. Motived by the work [31] and [32], we find that combination of geographic information and semantic information can promote spatiotemporal prediction from different views. Hence, we can discover some inner correlation semantic information from raw spatiotemporal data and fuse them with geographic information to improve final prediction performance. Compared with previous work, the aim of our work is to develop a soft-computing model that combines geographic information and latent semantic information, which can be transferred in many different tasks without any external data.

3 DATASET PROCESSING

In this paper, we use three datasets in different domains to demonstrate effectiveness of our model, urban theft crime records dataset, urban ride-haling demand dataset and urban fire record dataset, respectively. The urban crime and fire dataset in this research originates from a public safety data repository managed by San Francisco government[2], and the region of focus is defined in San Francisco city. The urban ride-haling demand dataset is from New Yorks Uber demand statistics[3]. To avoid some records from being too sparse and incomplete, we adjust the spatial and temporal range of the raw data. Considering adequate statistics, we identifies one week as a time slot in theft crime dataset, two hours as a time slot in ride-hailing dataset and one day as a time slot in fire dataset. And some details about these three adjusted datasets are shown in Table 1.

Table 1: Details about three different dataset

Dataset	Spatial range (Lat*Lon)	Temporal range	Number of time slots
Theft crime	$[37.71, 37.80]*$ $[-122.51, -122.38]$	2003/01/01− 2018/06/30	5600
Ride-hailing	$[40.628, 40.830]*$ $[-74.05, -73.88]$	2014/04/01− 2014/08/31	3628
Fire	$[37.71, 37.80]*$ $[-122.51, -122.38]$	2014/06/31− 2019/06/31	5600

Definition 1: Spatial Hotspot Map. Given the region lattice indexed (i,j), the time slot indexed t, we de?ne the element x_{ij}^t as the occurrence count of the speci?c event in such spatio-temporal background. And the SHM is composed of basic elements $X_t = [x_{ij}^t]_{I*J}$, where I*J is the set of geographical indexes.

The instance of SHM processing is shown in Figure 1. Target area are uniformly partitioned into 10*10 grids in this research.

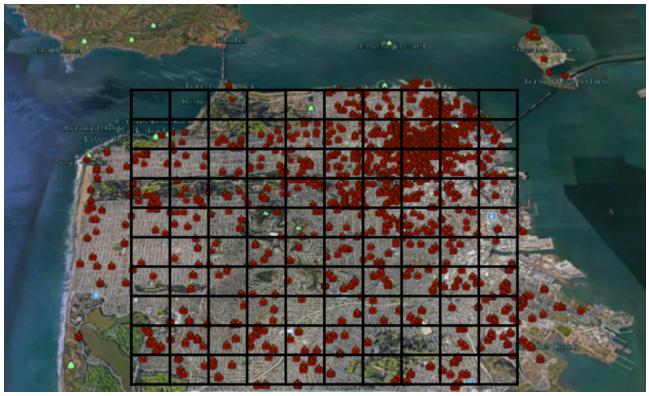

Figure 1: Spatial Hotspot Map grid division processing diagram.

Definition 2: Spatial Self-Semantic Graph. Given the region lattice indexed (i,j) and the temporal statistics in each grid x_{ij}^t, we take a grid region as a vertex and we can construct the adjacency matrix by absolute Pearson correlation coefficient. And the feature matrix of SSG is the temporal statistics in each node. The instance of SSG

[2]https://datasf.org/opendata/
[3]https://www1.nyc.gov/nyc-resources/agencies.page

construction is shown in Figure 2. The adjacency matrix of SSG is defined as:

$$r_{ij} = \begin{cases} p_{ij} = & |\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}| \\ 0 & if \quad p_{ij} < 0.05 \end{cases} \tag{1}$$

$$a_t^r = \begin{bmatrix} 1 & r_{0,1} & \dots & r_{0,98} & r_{0,99} \\ r_{1,0} & 1 & \dots & r_{1,98} & r_{1,99} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ t_{98,0} & r_{98,1} & \dots & 1 & r_{98,99} \\ r_{99,0} & r_{99,1} & \dots & r_{99,98} & 1 \end{bmatrix} \tag{2}$$

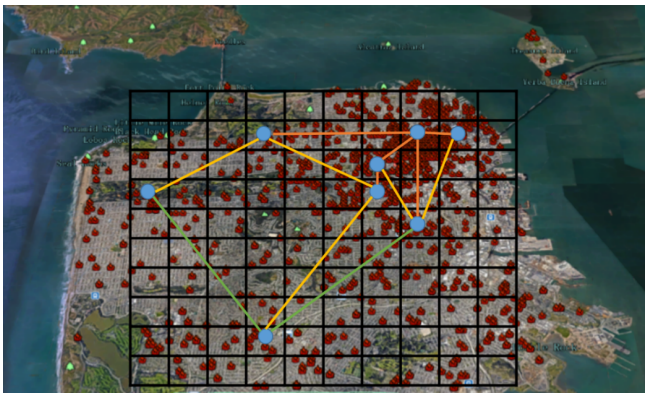

Figure 2: Spatial Self-Semantic Graph processing diagram. We can find that geographical neighborhood is no longer a limiting factor in semantic representation.

The adjacency matrix is stationary with time but feature matrix is constantly changing with time. However, the adjacency matrix is constructed based on long-term statistics in each small grid. Hence, the correlation graph can reflect the long-term semantic correlation between different regions. As we discuss above, we can beyond the limitations in Euclidean space geographic information. In addition, such graph construction does not need any other auxiliary information, so we call it as Self-Semantic Graph.

4 PROPOSED METHODOLOGY

In this section, we provide details of our proposed model Geographic-Semantic Ensemble Neural Network (GSEN).

Definition 3. Urban Hotspots Spatiotemporal Prediction. Given the M-1 step sequential SHMs $[X_{t+1}, X_{t+2}, , X_{t+M-1}]$ and the adjacency matrix A, we wish to predict the situation of the next step SHM X_{t+M}. These SHMs are continuous but do not overlap.

As illustrated in Fig 3: GSEN is mainly comprised of three components, PredRNN model, GC- PredRNN model and the ensemble layer. The function of PredRNN is to predict SHMs in geographical view while the function of GC- PredRNN is to predict SHMs in semantical view. The ensemble layer is to stack two different views and output the final SHMs.

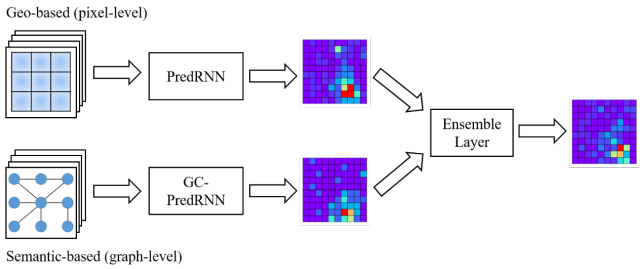

Figure 3: The overview of Geographic-Semantic Ensemble Neural Network.

Predictive Recurrent Neural Network

For a series of available SHMs x_t, we seek some recurrent model to capture their temporal correlations. But the SHMs are 3d tensors so combination of CNN and LSTM is suitable structure to process such data. Conv-LSTM is a common deep model for spatiotemporal prediction but the disadvantage of the model is the independence of memory information between the ConvLSTM units of each layer, which could cause the loss of low-level layer information. The motivation to apply Predictive Recurrent Neural Network (PredRNN) is that the spatial dynamics and temporal dynamics can be captured by convolutional layer and LSTM unit synchronously. Another motivation is that the involvement of spatiotemporal memory in PredRNN can overcome the biggest disadvantage of ConvLSTM.

The computation framework of PredRNN is similar to LSTM but involve the spatiotemporal memory. In mathematics form, the process is formulated as:

$$g_t = tanh(W_{xg} * x_t + W_{hg} * h_{t-1}^l + b_g) \tag{3}$$
$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1}^l + b_i) \tag{4}$$
$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1}^l + b_f) \tag{5}$$
$$C_t^l = f_t \odot C_{t-1}^l + i_t \odot g_t \tag{6}$$
$$g_t^{'} = tanh(W_{xg}^{'} * x_t + W_{mg}^{'} * m_{t-1}^l + b_g^{'}) \tag{7}$$
$$i_t^{'} = \sigma(W_{xi}^{'} * x_t + W_{mi}^{'} * m_{t-1}^l + b_i^{'}) \tag{8}$$
$$f_t^{'} = \sigma(W_{xf}^{'} * x_t + W_{mf}^{'} * m_{t-1}^l + b_f^{'}) \tag{9}$$
$$m_t^l = f_t^{'} \odot m_{t-1}^l + i_t^{'} \odot g_t^{'} \tag{10}$$
$$o_t = (W_{xo} * x_t + W_{ho} * h_{t-1}^l + W_{co} * C_t^l + W_{mo} * m_t^l + b_o) \tag{11}$$
$$h_t^l = o_t \odot tanh(W_{1*1} * [C_t^l, m_t^l]) \tag{12}$$

Where the notation * represents convolution operator, $\odot$ represents dot product, W represents convolution filter weights, b represents bias in each gate, m_t^l is spatiotemporal memory of l_{th} layer, C_t^l is cell output of l_{th} layer, h_t^l is hidden output of l_{th} layer. The overview of PredRNN and the internal structure of PredRNN cell are shown in Fig 4 and Fig 5 respectively.

Graph Predictive Recurrent Neural Network

For a series of available SSGs G_t, we also need temporal graph convolution model to process. Although some spatiotemporal graph neural network model were proposed, for

instance, GCRNN [25], GC-LSTM [34], they have common limitations with Conv-LSTM model. Hence, we proposed Graph Predictive Recurrent Neural Network (GC-PredRNN) in this case. This method is simple but effective. The computation framework of GC-PredRNN is similar to PredRNN, which only replace the convolution operator by graph convolution operator. In this paper, our model is based on the spatial GCN proposed by Kipf et al [35], which doesn't depend on the graph Laplace matrix. The graph convolution operation is defined as:

$$H^{l+1} = \sigma(\hat{D}^{-1/2}A\hat{D}^{-1/2}H^{l}W^{l}) \tag{13}$$

Where H^{l+1} is the hidden state in the $(l+1)_{th}$ layer, H^{l} is the hidden state in the l_{th} layer, W^{l} is the parameter of the l_{th} GCN layer. When $l = 0, H^{0}$ is the feature matrix of the graph. A denotes the adjacency matrix of the graph and $\hat{A} = A + I$. This transformation is to achieve self-accessible in graph convolution operation. $\hat{D}$ denotes the degree matrix of each vertex in $\hat{A}$. The operation of $\hat{D}^{-1/2}A\hat{D}^{-1/2}$ is to normalize $\hat{A}$ for training stability.

The total mathematical formula of GC-PredRNN is almost the same as PredRNN, the equations from (3)-(12), but the input x_t should be replaced by $G_t = (A, y_t)$, and the notation $*$ represents graph convolution operator in this case. Note that, in GC-PredRNN, the input element x_t should be reshaped as feature matrix y_t, whose size is (100,1), and the first dimension of feature matrices correspond to the number of nodes in the adjacency matrix. We also can see the overview of GC-PredRNN and the internal structure of GC-PredRNN in Fig 4 and Fig 5.

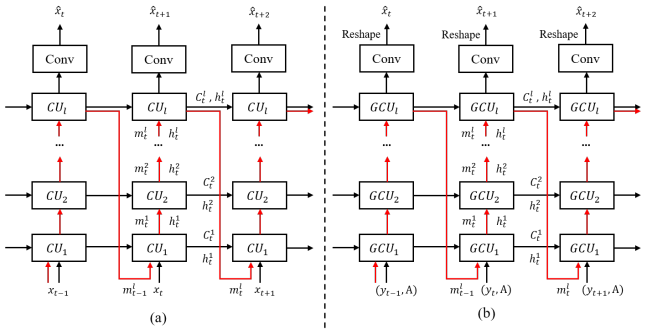

Figure 4: The overview of PredRNN and GC-PredRNN. The sub-figure (a) is PredRNN and the sub-figure (b) is GC-PredRNN. Where CU_l is the basic PredRNN unit with convolution operation, GU_l is the basic GC-PredRNN unit with graph convolution operation. We also can find that a convolutional layer is connected after the last PredRNN and GC-PredRNN unit to output the final predictions. Note that, in GC-PredRNN, we also need reshape the outputs from the last GC-PredRNN unit as the input format of the convolutional layer. The spatiotemporal memory flow is marked by red line in this overview.

Ensemble Layer

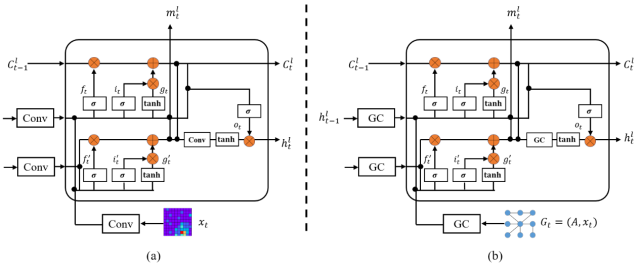

Figure 5: The basic cell unit of PredRNN and GC-PredRNN. The sub-figure (a) is PredRNN cell unit and the sub-figure (b) is GC-PredRNN cell unit. The black bold nodes in the figure represent the intersection of information.

Ensemble learning approach has been developed rapidly in recent years. The categories of ensemble learning are mainly divided into three, Bagging, Boosting and Stacking respectively [36]. In this paper, we apply stacking-based approach, which can obtain the final output by calculating the stacking output of multiple models. The motivation is that normal convolution and graph convolution respectively focus on different categories of features and fuse them through nonlinear operation to achieve more comprehensive perception of spatial features. Considering the outputs x^{g}_{t+M+1} from PredRNN and outputs x^{s}_{t+M+1} from GC- PredRNN should be combined nonlinearly, we utilize convolutional layer as ensemble layer to obtain the final outputs $\hat{x}_{t+M+1}$. The structure of convolutional layer is shown in Fig 6 and the formula is defined as:

$$\hat{x}_{t+M+1} = W_e * [x^{g}_{t+M+1}, x^{s}_{t+M+1}] \tag{14}$$

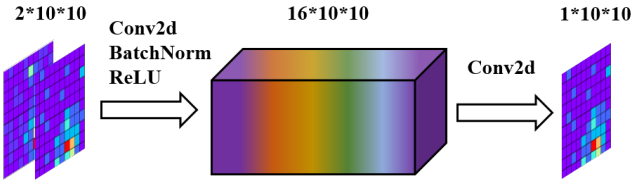

Figure 6: The structure of convolutional ensemble layer.

5 EXPERIMENTS AND ANALYSIS

In this section, the training details are described, and some numerical analysis are presented to show the performance of our model.

5.1 Training Details and Comparison Algorithm

The Adam optimizer is used in training process of GSEN. There are three Adam optimizers need to be set in three components of GSEN, PredRNN, GC-PredRNN, Ensemble Layer respectively. The batch is set as 16. The learning rates

are all set as 0.005 in these three optimizers. And we use early stop strategy in training process to prevent over-fitting if the loss of test set begins to decline no longer. In comparative experiment step, seven algorithms are presented to compare with GSEN, including ARIMA [10], Random Forest [37], XGBoost [38], Conv-LSTM [18], Deep Multi-View Spatial-Temporal Network (DMVST-Net) [31], PredRNN [6] and our proposed GC-PredRNN.

ARIMA: The auto-regressive term is set as 1, the difference order term is set as 1 and the moving average term is set as 1.

Random forest: The number of trees is set as 100, the maximum depth of each tree is set as 4, the features selection rate is 0.5 and the subsample rate is 0.5. In this case, we reshape the SHMs as 1D vector form as the input of Random forest.

XGBoost: The number of trees is set as 300, the maximum depth of each tree is set as 5, the subsample rate is 0.6. In this case, we reshape the SHMs as 1D vector form as the input of as the input of XGBoost.

Conv-LSTM: The number of cell unit layer is set as 2, the number of filters is set as [16,16], the size of filters are set as 3*3, the learning rate is set as 0.005. The batch size is set as 16. The input length remains the same as the length of GSEN determined by subsequent experiments.

DMVST-Net: For spatial view, the number of convolutional layers is set as 2, size of filters are set as 3*3, and dimension of the output is set as 64. For the temporal view, the input length remains the same as the length of GSEN determined by subsequent experiments. The learning rate is set as 0.001. The batch size is set as 16.

PredRNN: The number of cell unit layer is set as 2, the number of filters is set as [16,16], the size of filters are set as 3*3, the learning rate is set as 0.005. The batch size is set as 16. The input length remains the same as the length of GSEN determined by subsequent experiments.

GC-PredRNN: The number of cell unit layer is set as 2, the number of filters is set as [20,20], the learning rate is set as 0.005. The batch size is set as 16. The input length remains the same as the length of GSEN determined by subsequent experiments.

5.2 Result Analysis

We evaluate the effectiveness of proposed framework GSEN in this section, and concerning the performance of spatial dynamics and temporal dynamics capturing. In all experimental analysis steps, the quality of predicted SHMs is the most concerned. Three metrics are used in evaluation the prediction performance of each algorithm, respectively as the Root Mean Square Error (RMSE), Mean Absolute Percent Error (MAPE), Structural Similarity Index (SSIM).

Spatial Performance Analysis

For spatiotemporal prediction tasks, the spatial information extraction capability is one of our most concerned issues. Since PredRNN and GC-PredRNN are stacked by multiple convolutional units or graph convolutional units to extract spatial information, determining the number of basic unit layers is the most critical. As we know, multiple convolutional units or graph convolutional units could improve the capability of high-level information extraction but this also could lead to over-fitting or over-smooth in convolution operation or graph convolution operation if the spatial distribution complexity is lower than model complexity. Hence, we have to determine the suitable number of layers by experiments in this section. We fix the filters of convolution layers in PredRNN cell unit as 16, the filters of graph convolution layers in GC-PredRNN cell unit as 20. And we conduct experiments for PredRNN and GC- PredRNN respectively on three datasets and observe RMSE, MAPE and SSIM these three metrics. We select the optimal parameters of PredRNN and GC-PredRNN respectively. After all, the ensemble model will achieve the best performance if each component can achieve their best performance. And the range of cell layers is [1, 2, 3, 4]. In addition, we ignore the impact of the time slots length in this step so the input time slots length is fixed as 6 in this experiment.

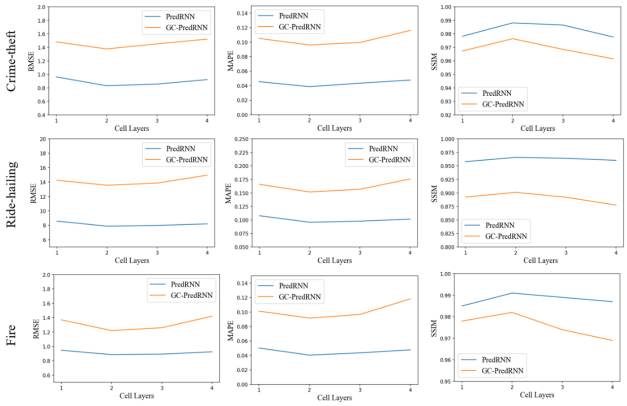

Figure 7: The visualization of RMSE, MAPE and SSIM changing with the number of cell layers on three different datasets.

From Fig 7, we can find that the optimal results can be obtained for PredRNN and GC-PredRNN if the number of cell layer equals 2. We can also find that the change in the number of cell layers causes greater fluctuations in outputs of GC-PredRNN than PredRNN. It may be that on the small scale graphs (100*100), too many graph convolution operations are more easily to cause over-smoothing. We determine the number of cell layers in both PredRNN and GC-PredRNN as 2 in subsequent experimental steps.

Temporal Performance Analysis

For spatiotemporal prediction tasks, the temporal information extraction capability is also one of our most concerned issues. Theoretically, the more historical information will make the models have stronger fitting ability. However, too long historical sequence information may lead to the loss of long-term memory, resulting in a reduction in prediction performance and even over-fitting. Hence, we also have to

Table 2: Comparison results of eight state-of-art algorithms for different spatiotemporal prediction tasks.

Algorithms	Metrics	Crime-theft	Ride-hailing	Fire
ARIMA	RMSE	2.5679±0.0147	23.5782±0.0204	1.9763±0.0081
	MAPE	0.2275±0.0078	0.2437±0.0048	0.2079±0.0063
	SSIM	0.8614±0.0026	0.8301±0.0059	0.8774±0.0028
Random Forest	RMSE	2.1289±0.0122	18.3297±0.0158	1.8124±0.0076
	MAPE	0.1882±0.0056	0.2112±0.0068	0.1725±0.0059
	SSIM	0.8855±0.0034	0.8676±0.0045	0.8965±0.0019
XGBoost	RMSE	2.0367±0.0095	18.6672±0.0165	1.7306±0.0085
	MAPE	0.1725±0.0062	0.2176±0.0075	0.1671±0.0054
	SSIM	0.8911±0.0029	0.8655±0.0036	0.9023±0.0023
Conv-LSTM	RMSE	1.3289±0.0046	12.4486±0.5097	1.1934±0.0045
	MAPE	0.0948±0.0031	0.1495±0.0113	0.0892±0.0026
	SSIM	0.9766±0.0012	0.9122±0.0041	0.9855±0.0021
DMVST-Net	RMSE	1.0278±0.0052	8.0142±0.4796	1.1645±0.0044
	MAPE	0.0477±0.0018	0.9642±0.0023	0.0548±0.0016
	SSIM	0.9788±0.0014	0.9641±0.0022	0.9802±0.0010
PredRNN	RMSE	0.7856±0.0052	7.8871±0.5618	0.8757±0.0044
	MAPE	0.0362±0.0013	0.0957±0.0028	0.0387±0.0015
	SSIM	0.9893±0.0011	0.9658±0.0019	0.9965±0.0012
GC-PredRNN	RMSE	1.3396±0.0052	13.5631±0.5618	1.1827±0.0048
	MAPE	0.0956±0.0023	0.1518±0.0121	0.0895±0.0016
	SSIM	0.9774±0.0017	0.9011±0.0034	0.9867±0.0025
GSEN	RMSE	0.6425±0.0057▲	3.4143±0.0137▲	0.7443±0.0061▲
	MAPE	0.0228±0.0017▲	0.0416±0.0021▲	0.0302±0.0014▲
	SSIM	0.9972±0.0014▲	0.9901±0.0012▲	0.9989±0.0018

determine the suitable input time slots length by experiments. In this step, we also conduct experiments for PredRNN and GC- PredRNN respectively on three datasets and observe RMSE, MAPE and SSIM these three metrics. We apply the optimal parameters of cell layers and the range of input length is [2, 4, 6, 8, 10, 12] in this experiment.

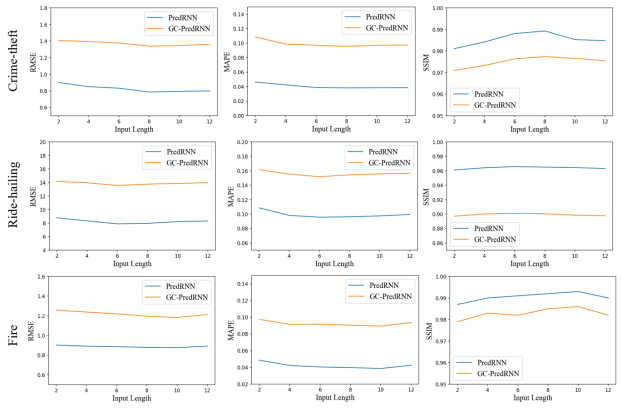

Figure 8: The visualization of RMSE, MAPE and SSIM changing with the length of input sequence on three different datasets.

From Fig 8, we can find that PredRNN and GC-PredRNN are not so sensitive to the input sequence length parameters in general. The change trend of the three evaluation metrics with the input length parameter is marginal. But we can still select relatively good parameters from this experiment. For Crime-theft dataset, the optimal parameter M equals 8. For Ride-hailing dataset, the optimal parameter M equals 6. For Fire dataset, the optimal parameter M equals 10.

Comparative Experiments

In Table 2, the comparison results on ARIMA, Random Forest, XGBoost, Conv-LSTM, DMVST-Net, PredRNN, GC-PredRNN and GSEN are displayed. We conducted 10 independent experiments with different random seeds and the results by the best performer in each column are in boldface. We assume that ten results are consistent with a normal distribution and statistical significance of pairwise differences of the structure GSEN vs. the other state-of-art Models is determined by a t-test (▲/▼ for = 0.1).

From Table 2, we can find that mathematical modeling model ARIMA and statistical learning model Random Forest, XGBoost still have some gaps with deep learning methods in prediction performance. In contrast, deep learning methods have certain advantages in capturing complex spatiotemporal latent dynamics. We can also find that GSEN is a significant optimal algorithm, and has achieved excellent performance on almost every metric. Compared with PredRNN, GSEN reduces MAPE by 0.0134, 0.0541, 0.0085 respectively and increase SSIM by 0.0079, 0.0243, 0.0024 respectively on Crime dataset, Ride-hailing dataset and Fire dataset. Compared with GC-PredRNN, GSEN reduces MAPE by 0.0728,

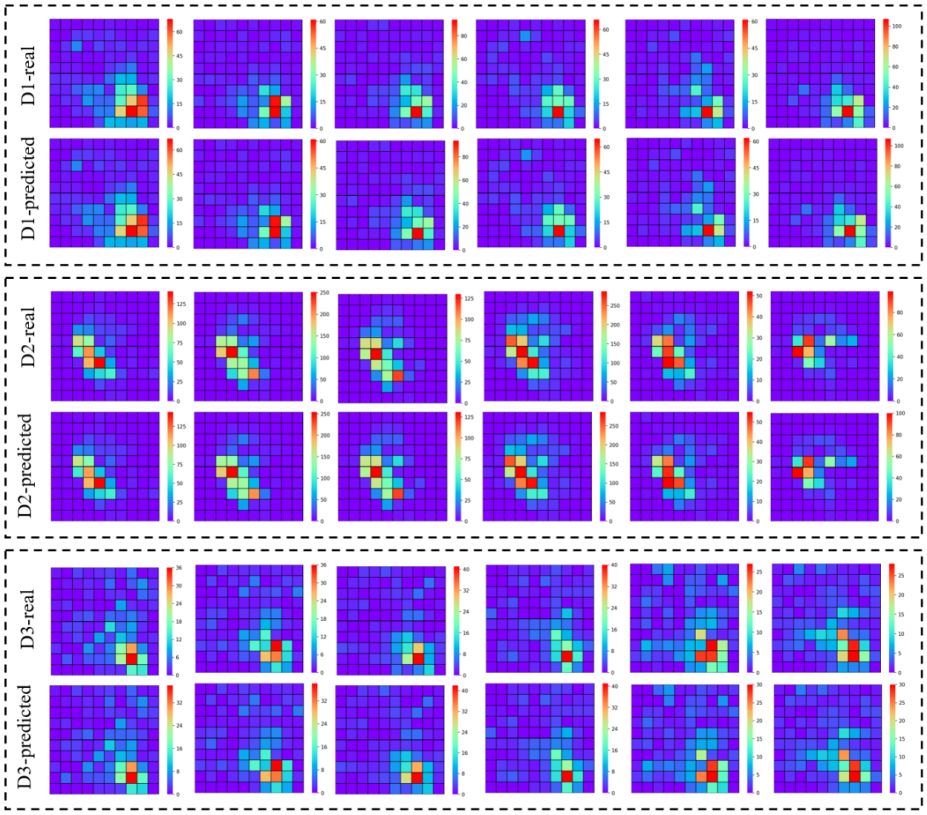

Figure 9: Visualization of the prediction performance of the GSEN model on three datasets. (D1: Crime-theft, D2: Ride-hailing, D3: Fire).

0.1102 and 0.0593 respectively and increase SSIM by 0.0198, 0.089, 0.0122 respectively on Crime dataset, Ride-hailing dataset and Fire dataset. This suggests that the fusion of geographic information and semantic information is effective and does improve the accuracy of different urban spatiotemporal hotspots prediction tasks, whether from numerical or distribution perspective.

As the instantiation, the predicted SHMs are initialized from one selected time slot in each dataset when M=8 on Crime-theft dataset, M=6 on Ride-hailing dataset and M=10 on Fire dataset. We choose six consecutive time slots predicted SHMs to show performance of GSEN. Figure 1 from top to bottom are Crime-theft dataset, Ride-hailing dataset and Fire dataset. In each dataset unit, the first row is the real SHMs and the second row is the predicted SHMs. Crime-theft is initialized from September 1st, 2016, Ride-hailing is initialized from 10 am, June 1st, 2014 and Fire is initialized from October 1st, 2018. We can find that our proposed model GSEN has achieved state-of-art prediction performance on each dataset. All hotspots in predicted SHMs and even image details are more accurately predicted.

6 CONCLUSION

In this study, we gain deep comprehension into the urban hotspots spatiotemporal prediction and a novel ensemble deep learning benchmark model GSEN is proposed. The experimental results of GSEN are remarkable in comparison to some traditional state-of-art models and the visualization also demonstrates that integration of geographic information and semantic information makes the spatiotemporal dynamics about different urban hotspots more foreseeable.

However, there are still some limitations in this work. (a) To avoid spatial hospots sparsity of SHMs, the granularity of the division of urban areas is not fine enough, resulting in some difficulties in spatiotemporal in fine small regions. (b) The average division method of the spatial grid does not consider the actual regional semantics, which may lead to the lack of practical significance for the hotspots prediction in the divided areas.

In the future, it is potential to make further improvements on the structure of GSEN through modelling more convincing semantic graphs and seeking some adaptive multi-scale prediction methods to make the model more explainable and comprehensive.

REFERENCES

[1] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

[2] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41, 2018.

[3] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[4] Tao Cheng and Jiaqiu Wang. Application of a dynamic recurrent neural network in spatio-temporal forecasting. In *Information Fusion and Geographic Information Systems*, pages 173–186. Springer, 2007.

[5] Senzhang Wang, Jiannong Cao, and Philip S Yu. Deep learning for spatio-temporal data mining: A survey. *arXiv preprint arXiv:1906.04928*, 2019.

[6] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, pages 879–888, 2017.

[7] Sebastian Meyer, Johannes Elias, and Michael Höhle. A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616, 2012.

[8] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[9] Yosihiko Ogata, Ritsuko S Matsu'ura, and Koichi Katsura. Fast likelihood computation of epidemic type aftershock-sequence model. *Geophysical research letters*, 20(19):2143–2146, 1993.

[10] Billy M Williams and Lester A Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

[11] Virginijus Radziukynas and Arturas Klementavicius. Short-term wind speed forecasting with arima model. In *2014 55th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*, pages 145–149. IEEE, 2014.

[12] Yanchun Pan, Mingxia Zhang, Zhimin Chen, Ming Zhou, and Zuoyao Zhang. An arima based model for forecasting the patient number of epidemic disease. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–4. IEEE, 2016.

[13] Tigran T Tchrakian, Biswajit Basu, and Margaret O'Mahony. Real-time traffic flow forecasting using spectral analysis. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):519–526, 2011.

[14] Narjes Zarei, Mohammad Ali Ghayour, and Sattar Hashemi. Road traffic prediction using context-aware random forest based on volatility nature of traffic flows. In *Asian Conference on Intelligent Information and Database Systems*, pages 196–205. Springer, 2013.

[15] Wei-Chiang Hong. Traffic flow forecasting by seasonal svr with chaotic simulated annealing algorithm. *Neurocomputing*, 74(12-13):2096–2107, 2011.

[16] Ruiyun Yu, Yu Yang, Leyou Yang, Guangjie Han, and Oguti Ann Move. Raq–a random forest approach for predicting air quality in urban sensing systems. *Sensors*, 16(1):86, 2016.

[17] Amy Mcgovern, Timothy Supinie, II Gagne, Troutman N DJ, MATTHEW Collier, Rodger A Brown, JEFFREY Basara, and J Williams. Understanding severe weather processes through spatiotemporal relational random forests. In *2010 NASA conference on intelligent data understanding (to appear)*. Citeseer, 2010.

[18] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[19] Bao Wang, Penghang Yin, Andrea Louise Bertozzi, P Jeffrey Brantingham, Stanley Joel Osher, and Jack Xin. Deep learning for real-time crime forecasting and its ternarization. *Chinese Annals of Mathematics, Series B*, 40(6):949–966, 2019.

[20] Guangyin Jin, Qi Wang, Xia Zhao, Yanghe Feng, Qing Cheng, and Jincai Huang. Crime-gan: A context-based sequence generative network for crime forecasting with adversarial loss. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1460–1469. IEEE, 2019.

[21] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, Yanwei Yu, and Zhenhui Li. Modeling spatial-temporal dynamics for traffic prediction. *arXiv preprint arXiv:1803.01254*, 2018.

[22] Congcong Wen, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu, and Tianhe Chi. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of the Total Environment*, 654:1091–1099, 2019.

[23] Van-Duc Le, Tien-Cuong Bui, and Sang-Kyun Cha. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 55–62. IEEE, 2020.

[24] Nicholas G Polson and Vadim O Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017.

[25] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *International Conference on Neural Information Processing*, pages 362–373. Springer, 2018.

[26] Ling Zhao, Yujiao Song, Min Deng, and Haifeng Li. Temporal graph convolutional network for urban traffic flow prediction method. *arXiv preprint arXiv:1811.05320*, 2018.

[27] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[28] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3656–3663, 2019.

[29] Qi Wang, Guangyin Jin, Xia Zhao, Yanghe Feng, and Jincai Huang. Csan: A neural network benchmark model for crime forecasting in spatio-temporal scale. *Knowledge-Based Systems*, 189:105120, 2020.

[30] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[31] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] Yuandong Wang, Hongzhi Yin, Hongxu Chen, Tianyu Wo, Jie Xu, and Kai Zheng. Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1227–1235, 2019.

[33] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. Deepcrime: attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1423–1432, 2018.

[34] Jinyin Chen, Xuanheng Xu, Yangyang Wu, and Haibin Zheng. Gc-lstm: Graph convolution embedded lstm for dynamic link prediction. *arXiv preprint arXiv:1812.04206*, 2018.

[35] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[36] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[37] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[38] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.