2017 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2017, Cancun, Mexico, December 4-7, 2017

Deep Learning for Real Time Crime Forecasting

Bao Wang†*, Duo Zhang†, Duanhao Zhang†, P. Jeffrey Brantingham‡, and Andrea L. Bertozzi†

†Department of Mathematics, UCLA
‡Department of Anthropology, UCLA
Email: wangbaonj@gmail.com, jasonzhang56886@gmail.com, adylucioforzm@live.com,
branting@g.ucla.edu, bertozzi@math.ucla.edu

Abstract—Accurate real time crime prediction is a fundamental issue for public safety, but remains a challenging problem for the scientific community. Crime occurrences depend on many complex factors. Compared to many predictable events, crime is sparse. At different spatio-temporal scales, crime distributions display dramatically different patterns. These distributions are of very low regularity in both space and time. In this work, we adapt the state-of-the-art deep learning spatio-temporal predictor, ST-ResNet [Zhang et al, AAAI, 2017], to collectively predict crime distribution over the Los Angeles area. Our models are two staged. First, we preprocess the raw crime data. This includes regularization in both space and time to enhance predictable signals. Second, we adapt hierarchical structures of residual convolutional units to train multi-factor crime prediction models. Experiments over a half year period in Los Angeles reveal highly accurate predictive power of our models.

Keywords: Crime modeling, Spatio-temporal deep learning, Real-time forecasting, Sparsity.

1. Introduction

Real time crime forecasting is an important scientific and sociological problem. It is directly related to our quality of life. Recent efforts have been devoted to the mathematical modeling of crime. Short et al developed novel differential equation models for dynamics of crime hotspots.[8] Considering crime as self-exciting, Mohler et al applied the classical epidemic type aftershock sequence (ETAS) model and its variants to crime modeling.[5,6] These types of models provide a microscopic representation of the crime events with predictive power. The aforementioned models are built from historical data. There is also interesting work on crime prediction using social network data, e.g., Twitter.[10]

Deep learning, a cutting-edge technology for automatic feature identification via a deep neural network (DNN), gives state-of-the-art performance on many predictive scenarios, such as image classification, computer vision, speech recognition.[4] Moreover, recurrent neural networks (RNN), which are superior in mining long-range dependencies, enables highly accurate prediction of sequential data.[2] Recently, deep learning paradigms have been applied to study spatio-temporal data. There are two frameworks of interest. One is the convolutional neural network (CNN) for learning spatial traits and RNN to learn temporal features.[11] The other is to embed spatio-temporal data into an RNN framework via graphical models.[3]

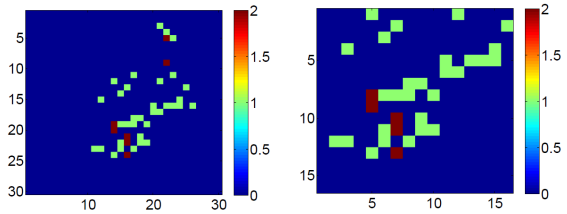
There are many challenges in accurate real time spatio-temporal crime prediction. For example, compared to the traffic flow data studied in,[11] crime data has much less spatio-temporal regularity, i.e., the number of events in adjacent time intervals and spatial cells differs hugely.[7] Furthermore, crime types are diverse. Our contribution in this work is three-fold. First, we select the appropriate spatio-temporal scale at which crimes are predictable. Second, we provide different approaches for data regularization in both the spatial and temporal dimensions. Third, we adapt deep learning architectures for predictions. Both convolutional and non-convolutional models are considered. In the first case, spatial connectivity is included. In the second case, grids are independent of each other.

We structure this paper as follows: In section 2, we discuss data sets. In section 3, the feature and deep learning architectures are presented. In section 4, we present an example of spatio-temporal crime prediction over the last two weeks of 2015 for the whole Los Angeles (LA) area. Future work is discussed in section 5.

2. Data Set and Preprocessing

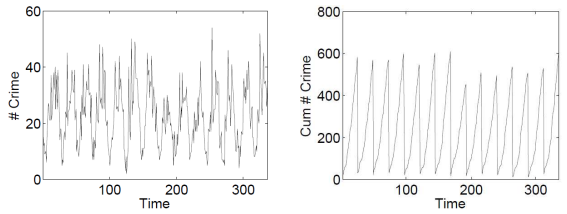
Data set description In this work, we consider all types of crime in LA over the last six months of 2015. In total there were 104,957 crimes. Each crime record includes crime start, end times and location. To avoid ambiguity, we regard start time of each event as the time slot. Geographically, the latitude and longitude intervals spanned by these crimes are $[33.3427°, 34.6837°]$ and $[-118.8551°, -117.7157°]$, respectively. The spatial distribution is highly heterogeneous. A large portion of the area contains only a little crime. Therefore, we only consider crimes that happened within the region $[33.6927°, 34.3837°] \times [-118.7051°, -118.1157°]$. In our study, we partition this selected region into a 16×16 lattice. Figure 1 shows the crime distribution at 1:00 p.m on Dec 20th, 2015. The left panel is the crime distribution over the whole LA area. The right panel depicts the crimes in the restricted region. The restricted region contains more than

*corresponding author



(a) (b)

Figure 1: Crime distribution at 1:00 p.m, Dec 20th, 2015. Chart (a) depicts crime distribution over the whole LA area; chart (b) depicts crime distribution over the selected region. The units are described in Sec. 2.



(a) (b)

Figure 2: Chart (a) depicts the hourly crime intensity of the last two weeks of 2015 over the whole LA area; chart (b) draws the cumulated crime intensity corresponding to (a). Units: x-axis: time; y-axis: number of crimes.

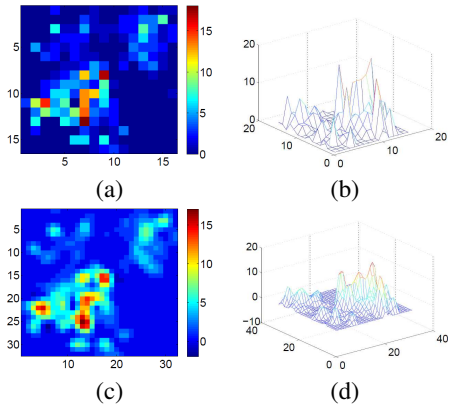
95 percent of all crimes.

Data preprocessing Figure 2 reflects low regularity in the temporal dimension. Nevertheless, the hourly crime intensity indicates a strongly predictable signal; this signal is indeed periodic with a period of 24 hours. This periodic temporal pattern exists in each grid cell. It is more spiky for the grid-wise crime intensity (see Fig.6). Despite the predictability, the data is still inappropriate for deep learning which requires some regularity of the inputs, especially for grid-wise intensity. To address this, we map the original crime intensity $\{X(t)\}$ to $\{Y(t)\}$ via a periodic integral mapping:

$$Y(t) = \int_{nT}^{t} X(s)ds, \tag{1}$$

for t within the time interval $(nT, (n+1)T]$. Figure 2 (a) shows the crime intensity over the whole LA region. Figure 2 (b) plots the cumulative intensity of (a) after the mapping defined by Eq.(1).

To address lack of spatial regularity, we can use a super resolution technique involving cubic spline interpolation. For computational efficiency, we resolve by a factor of 2 in each dimension. After spatial super resolution, the dimension of the input data becomes 32×32. From Fig. 3, we see that this cubic spline super resolution significantly improves spatial regularity. A merit of this preprocessing is that it improves the signal without losing information as-



(a) (b) (c) (d)

Figure 3: Cumulated crime intensity at 11:00 p.m, Dec 31st, 2015. Chart (a) depicts crime distribution over the selected area; chart (b) provides the mesh plot of the chart (a); chart (c) depicts super resolution version of chart (a); and char (d) is mesh plot of chart (c).

sociated with the crime data.

3. Models

Problem formulation Given the historical filtration $\{(X_t, E_t)\}_{t=1,2,\cdots n}$ and $\{E_{n+1}\}$, to predict X_{n+1}, where $X_1, X_2, \cdots, X_{n+1}$ are the tensors representing the crime spatial distributions at times $t_1, t_2, \cdots, t_{n+1}$, $E_1, E_2, \cdots, E_{n+1}$ are external features that affect the crimes, (e.g., holiday, time, weather), our protocol is divided into the following three steps:

- Apply the spatio-temporal regularization to the historical data to get $\{(\hat{X}_t, E_t)\}_{t=1,2,\cdots,n}$.
- Predict regularized cumulated crime intensity $\hat{X}^p_{n+1}$.
- Perform the spatio-temporal inverse map to get the crime intensity prediction X^p_{n+1}.

Deep neural network structure The deep neural network architecture is depicted in Fig. 4. We test two different structures: one is plotted in Fig. 4; the other one is similar but without convolutional layers. For the second model, we apply an ensemble of ResNet to learn the time series on each grid, without considering the transition of crimes between different grids. The first model is more realistic. Through convolutional layers, crime dynamics can be captured. In both networks, all features are fused with the crime data itself via a parametric-matrix based fusion technique used in.[11] The detailed description of the network structure can be found in.[11] We implement our models using Keras[1] on top of Theano[9] software.

Our models incorporate external features, i.e., weather, holidays. Due to the periodic pattern and self exciting property of crimes,[5] we adopt trend, periodic, and nearby fea-

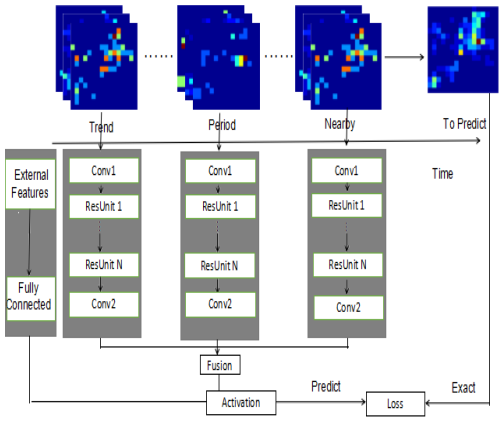


Figure 4: Structure of the deep neural network models.

tures. The time spacing of these features are at weekly, daily, and hourly levels, respectively.

4. Numerical Results

In our tests, for both DNNs, the parameters are set as follows: We use 6 layers of residual units to assemble the ST-ResNet. The number of filters in the convolutional layers are all set to be 64×64. The size of the convolution filters are fixed to 3×3. The length of the trend, period, and nearby dependencies are all chosen to be 3. Six layers of residual units are used in each dependency. In the training stage, the cross validation ratio is 0.1. The number of epochs at cross validation and fine tune periods are 200 and 50, respectively. The learning rate is chosen to be 0.0005. Batch normalization is used. We adopt the ADAM optimizer to optimize the loss function. We select the last two weeks of 2015 as the test set; all the other data is used to train the DNN models. All the experiments are carried out with a single Nvidia Quadro-K4000 graphics card. To speed up the training process, we make use of the deep neural network specific library cuDNN. We show an example of snapshots in time in Fig.5. The predictions match quite well with the ground truth. All crime hotspots are correctly predicted.

For a given grid, the crime intensity over a given time interval is also accurately predicted. As shown in Fig. 6, the maximum difference between the ground truth and the prediction is 2 crimes.

To quantitatively measure the performance of the models, we use the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N * T} \sum_{i,t} (I_{it} - I_{it}^p)^2} \qquad (2)$$

where N is the total number of grids that we partition the restricted area into, T is the number of time slots considered, I_{it}, I_{it}^p are the exact and predicted crime intensity in grid i at time t, respectively. By using the same parameters as above, the training and testing RMSEs are listed in Table 1. Using the convolution technique to incorporate spatial connectivity gives better predictions.

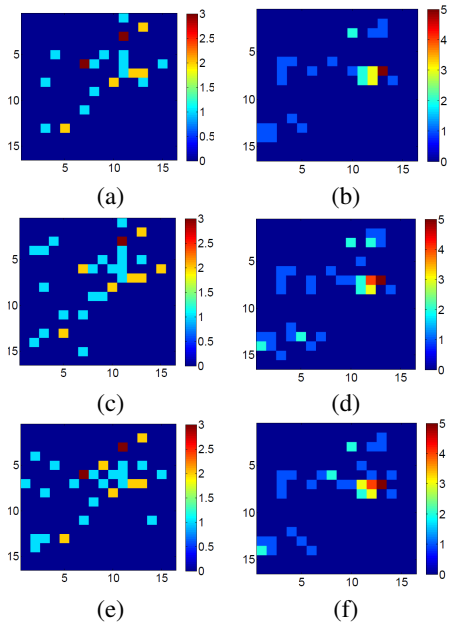


Figure 5: Predicted vs. exact crime spatial distribution. Panels (a), (b) plot the crime spatial distribution at 1 p.m. of Dec 19, 27, 2015, respectively. Panels (c), (d) are the predicted results without convolution layers. (e), (f) are the predicted results with convolution layers.

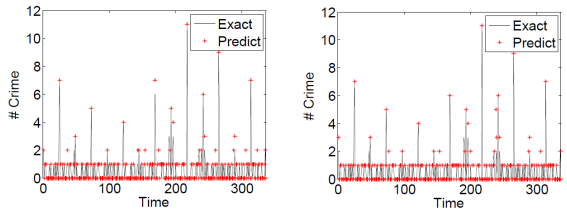
Table 1: Performance comparison between two models. Units for Train and Test columns: Number of crimes.

Model	Train	Test	#parameters	Training Time
Convolution	0.143	0.207	1,350,911	104157.02s
No Convolution	0.363	0.379	165,119	23137.48s

For real applications, we focus more on the accuracy of the top N predictions. Let the top N crime hotspots be indexed as $I_N = \{i_1, i_2, \cdots, i_N\}$, where $i_1, i_2, \cdots, i_N$ are indices of the grids in the previous partition. Furthermore, denote the predicted indices of the top N crime hotspots as $I_N^p = \{j_1, j_2, \cdots, j_N\}$. We define the following accuracy of top N ranking:

$$\text{Acc}_\text{N} = \frac{|I_N \bigcap I_N^p|}{|I_N|}, \qquad (3)$$

where $|*|$ is the cardinality of the set $*$. This top N recommendation system provides a practical guidance to control crimes efficiently. Table 2 lists the top N accuracy averaged among all testing time slots. Using convolution provides slightly better results for all N, though the system without convolution also gives good recommendations, indicating



(a) (b)

Figure 6: Predicted vs. exact crime intensity in the area $[33.9519, 33.9951] \times [-118.3003, -118.2635]$ over the last two weeks of 2015. Charts (a) and (b) are results for models with and without convolutional layers, respectively. Units: x-axis: time; y-axis: number of crimes.

crime hotspots are quite stationary during this test time period.

Table 2: Top N prediction accuracy of the above two models. Unit: %.

Model	5	10	15	20	25
Convolution	80.77	83.90	84.35	84.60	84.78
No Convolution	80.48	82.26	82.90	83.29	83.95

5. Concluding Remarks

In this work, we apply the ST-ResNet to real time crime prediction on an hourly timescale. Due to the low regularity of the crime data in both space and time, we perform both spatial and temporal regularization of the data. More specifically, in the temporal dimension, we compute the diurnal cumulative crime per spatial region. In the spatial dimension, we use cubic spline interpolation super resolution. Compared to applying the residual network on each individual grid, our data preprocessing and CNN approach improves the prediction accuracy dramatically. The improvement is due to the fact that, in the convolutional model, the spatial information is no longer isolated allowing the crime hotspot transitions to be captured. Our predictions are extremely accurate in both space and time, which can provide reliable guidance for crime control.

There are still many things to do. In our current work, we have not classified the crime types. In reality, different types of crime may lead to different consequences. Intervention strategies should not be based on the crime intensity alone; crime type should also be addressed. Due to the super resolution regularization in space, the computational cost increases dramatically. One possible way to improve computational efficiency is to use an adaptive super resolution to train each layer of the DNN. Alternatively, we can improve the loss function of the deep learning to make the ResNet applicable to low regularity data. Another deficiency of the presented models comes from the ad hoc grid partitioning of the LA area. A better approach would be to use a graph to model the intrinsic geographic connectivity.

Acknowledgments

We thank Dr. Penghang Yin for fruitful discussions. This work was supported by ONR grant N00014-16-1-2119, ARO MURI grant W911NF-11-1-0332, NSF grants DMS-1118971, DMS-1417674, and DMS-1737770. The authors thank the Los Angeles Police Department for providing the crime data for this paper.

References

[1] F. Choillet. Keras. *https://github.com/fchollet/keras*, 2015.

[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput*, 9:1735–1780, 1997.

[3] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. *CVPR*, 2016.

[4] Y. LeCun, Y. Bengion, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[5] G. Mohler, M. Short, P. Brantingham, F. Schoenberg, and G. Tita. Self-exciting point process modeling of crime. *Journal of Americal Statistical Association*, 106(493):100, 2011.

[6] M. Short, G. Mohler, P. J. Brantingham, and G. Tita. Gang rivalry dynamics via coupled point process network. *Discrete and Continuous Dynamical System B*, 34(1459), 2014.

[7] M. Short, G. Mohler and P. J. Brantingham. The Concentration Dynamics Tradeoff in Crime Hot Spotting, Connecting Crime to Place: New Directions in Theory and Policy. *in press*.

[8] M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. Chayes. A statistical model of criminal behavior. *M3AS: Mathematical Models and Methods in Applied Sciences*, 18:1249–1267, 2008.

[9] Theano Development Team. Theano: A python framework for fas computation of mathematical expressions. *arxiv 1605.02688*, 2015.

[10] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from twitter posts. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 231–238, 2012.

[11] J. Zhang, Y. Zheng, and D. Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. *AAAI*, 2017.