Spatiotemporal Crime Rate Prediction using Regional LSTMs

Selim Firat Yilmaz
Bilkent University
Ankara/Turkey
firat.yilmaz(at)ug.bilkent.edu.tr

Hamdi Dibeklioglu
Bilkent University
Ankara/Turkey
last_name(at)cs.bilkent.edu.tr

Abstract

Crime rate prediction aims to predict the crime intensities in the future. Althoguh crimes seem highly complex events, criminology studies argue that crimes are not random events but very dependent on the components of crime such as environment, target, performer, and their trajectories. In this study, we propose new evaluation framework and validate it through our baseline methods. Also, we propose Regional LSTMs model which explains 16% *of the variation within the Chicago Police Department's crime data.*

1. Introduction

Criminology studies criminal actions and its patterns [22]. Although crime seems as a random event for many people due to its complexity, crime pattern theory argues that patterns exists in such complexities [22]. According to the crime pattern theory, every component of criminal events complexly related to its past trajectory, status, routines, and the environment [22].

Crime rate prediction aims to solve a problem of the real life. Understanding the patterns of criminal events would lead us to understand crimes and their motives. Predicting crime rates of regions helps polices to fight crime and improve the safety in the public areas. Crime rate prediction can be used to optimize the patrol routes [9]. Crime rate prediction also helps to optimize the distribution of funding and logistic supplies to the police departments.

Crime rate prediction is a challenging problem. Crimes are obviously dependent on the existence of target(s) and performer(s), however, this information, or even people's locations, is not available. There is many different styles of segmenting regions(gridding) and none is validated through an ablation study. Although the crime datasets has many occurrences, their spatiotemporal distribution is uneven, highly sparse and imbalanced for some crime types [21]. In our case, we consider different types of crimes as separate classes so it increases the sparsity of the data. Crime statistics as well as information obtained from GIS exist, however, these information change often over time [16]. Hence, crime rate prediction requires to incorporate geolocational(spatial) relations, temporal relations, and spatiotemporal relations.

Our contributions are as follows.

- We give a new representation of the crime rate prediction.
- We propose a new evaluation framework and validate it via our baseline methods.
- We propose a novel LSTM based Regional LSTMs model that incorporates spatial and temporal information.
- To provide reproducibility and a starter code of our experimental framework, we make all source code and data visualizations publicly available at https://github.com/selimfirat/st-crime-rate-prediction.

2. Related Work

2.1. Crime Prediction

To our knowledge, there is no common evaluation framework exists for crime rate prediction problem. Likewise, to our knowledge, there is no previous work either makes predictions for districts or does regression. Thus, our results are not comparable to the previous works.

Crimes are a non-random events happening in place and time between an offender and a target (a living or nonliving) [3]. Environmental criminology analyzes the relations between people's behaviors and external factors such as law, time, and geography and focuses on exploiting specific crime patterns within an environmental context [24]. Social criminology focuses on relations between criminal occurrences and socioeconomic, cultural, demographic indicators [24]. Both environmental and social criminology seem to have many useful features for machine learning tasks. Many integrible information such as census data of specific locations is not available for space-time tuples but

Work	Features	Gridding	Model	Task
Nyugen et al. [13]	Demographic	Point	SVM	Classification
Kang and Kang [10]	Demographic + Street View	Point	MLP	Classification
Porzi et al. [17]	Street View	Point	CNN Based Ranking	Classification
Chen et al. [4]	Tweets' Sentiment (Lexicon)	Rectangle	KDE	Classification
Gerber [6]	Tweets' Sentiment (LDA)	Rectangle	KDE	Classification
Wang et al. [21]	Crime Intensities + Weather	Rectangle	St-ResNet	Classification
Our's	Crime Intensities	**District**	**Regional LSTMs**	**Regression**

Table 1: Summary of previous works and our's.

usually is available in only regional or temporal or recent. Hence, this section aims to answer following questions.

- What are the used features?
- How prediction(gridding) regions are segmented?
- Which algorithms/architectures of machine learning are used for crime prediction?

Nyugen et al. [13] incorporates demographic, economic, educational and ethnic information into crime prediction. They concatenate the information vector and feeds it into support vector machines(SVM), Random Forest(RF), gradient boosting tree, and multilayer perceptron. Latter, they deal with imbalanced nature of the crime data by undersampling the majority crime classes.

Kang and Kang [10] incorporates demographic, housing, education, economic, weather and Google street view images. The categorizes these information into spatial, temporal, and environmental context groups. They propose a multi-modal deep neural network architecture by feeding them into 3 distinct sequence of hidden layers, concatenating their last layers and again 3 hidden layers. Prediction is done via softmax layer in output layer. They train whole network jointly.

Porzi et al. [17] proposes convolutional neural network based architecture for ranking safety levels of spatial regions. They feed scenes of urban crime location's scene to the network and predicts safety levels of the specific locations.

Chen et al. [4] and Gerber [6] integrates the information from Twitter. They collect the GPS-tagged data and calculates features per grid point. Chen et al. [4] applies lexicon based algorithm sentiment detection to collect polarity information per grid in addition to the weather information. Gerber [8] applies latent dirichlet analysis(LDA) to detect topical distribution of geo-tagged tweets. Latter, both [4] and [6] applies kernel density estimation.

Wang et al. [21] employs spatiotemporal ResNet (St-ResNet) architecture of [23] and emsembles that architecture with external features such as weather information feeded to fully connected. Summary of previous crime prediction works can be found in Table 1.

2.2. Sequence Modeling

Long Short Term Memory(LSTM) networks are widely used deep learning technique for sequence modeling such as modeling language [18]. LSTM models proven to be effective to propagate information over long sequences [8]. Today's crime prediction techniques such as [3, 24, 13, 10] are at least incorporates sequential modeling of the data. In sequence modeling, autoregressive procedure is also applied [2]. Many works integrate auto regressive procedure to the LSTMs [7].

3. Methodology

3.1. Problem Description

In the problem of crime rate prediction, we predict count of future in a region of a particular crime type. Given set of crime types $C = \{C_1, C_2, ..., C_p\}$, set of distinct regions $K = \{K_1, K_2, ..., K_r\}$, sequence of past time bins $T = 1, 2, ..., t$ and crime occurrence history tensor X with dimensions $(|K|, |C|, |T|)$. An entity $X^t_{k,c}$ in the tensor X represents the number of crimes of type $c \in C$ in region $k \in K$. Our goal is to predict $X^{t+1}_{k,c}$.

3.2. Evaluation Metric

R^2 (often referred as R-squared) shows the amount of the variation explained by a model. $R^2 = 0.1$ tells that 10% of the variation within the data is explained by the model. The greater the R^2 value, the better the model. We use the generalized version of the R^2 metric. The coefficient of determination R^2_c for class $c \in C$ is defined as

$$SS_r = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$SS_t = \sum_{i=1}^{m} (y_i - \bar{y}_i)^2$$

$$R_c^2 = 1 - \frac{SS_t}{SS_r}$$

$$R^2 = \frac{1}{|C|} \sum_{c \in C} R_c^2$$

Our task is multi-target regression so we use generalized R^2 metric which is uniform average of all R_c^2 scores calculated for each crime type separately.

3.3. Experimental Setup

We see crime prediction as a discrete problem and so using precision, recall, and F1 metrics due to the sparsity of the datasets we use. Currently, since project is ongoing, only accuracy metrics exists but their effectivenes can still be observed regardless of the sparsity of the datasets.

Currently, accuracy is calculated as each crime type of each location-time tuples are distinct problems so that if there occur a crime in a location-time tuple and the model says 1 or there does not occur a crime and model says 0 for it is counted as correct classification. Otherwise, it is counted as false classification.

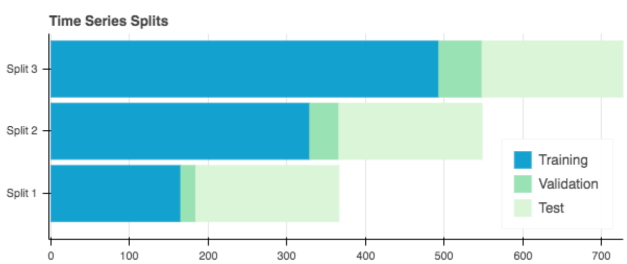

Figure 1: Visualization of time series splits we employ.

Experimentation protocol follows cross validation for time series. Visualization of splits can be seen in Figure 1. To split the data, the data are ordered in terms of time.

- Choose the part of (0, 25%] as training data, validate on the part of (25%, 35%], and test on the part of (35%, 50%].
- Choose the part of (0, 50%] as training data, validate on the part of (50%, 60%], and test on the part of (60%, 75%].
- Choose the part of (0, 75%] as training data, validate on the part of (75%, 85%], and test on the part of (85%, 100%].
- Then, average the R^2 scores for the experiments above.

3.4. Baseline – Meta Predictors

In this section, s represents the number of future time bins, p represents the number of crime types, and r represents the number of distinct regions. Also, $1_{a \times b \times c}$ is an (a, b, c) dimensional tensor consisting of only 1s and $0_{a \times b \times c}$ is an (a, b, c) dimensional tensor consisting of only 0s.

3.4.1 All Ones Meta Predictor

$$\hat{y} = 1_{p \times r \times s}$$

All ones meta predictor predicts 1 for each of time, region pairs regardless of the training data. It corresponds to the prediction that there will be a crime for each crime type in each region.

3.4.2 All Zeros Meta Predictor

$$\hat{y} = 0_{p \times r \times s}$$

All Zeros meta predictor predicts 0 for each of time, region pairs regardless of the training data. It corresponds to the prediction that there will be no crime at all.

3.4.3 All Total Mean Meta Predictor

$$\mu = \frac{\sum_{i=1}^{p} \sum_{j=1}^{r} \sum_{k=1}^{s} y_{ijk}}{p \times r \times s}$$

$$\hat{y} = \mu \times 1_{p \times r \times s}$$

Total mean meta predictor predicts the total mean μ for each of time, region pairs regardless of the training data. It corresponds to the prediction that the crime rate will be as much as mu regardless of the region and crime type.

3.4.4 Mean Per Class Meta Predictor

$$\mu_c = \frac{\sum_{j=1}^{r} \sum_{k=1}^{s} y_{cjk}}{r \times s}, \forall c \in C$$

$$\hat{y}_c = \mu_c \times 1_{r \times s}$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \dots & \hat{y}_{p-1} & \hat{y}_p \end{bmatrix}^T$$

Mean per class meta predictor predicts the class mean μ_c of each class of c for each time, region pairs regardless of the training data. It corresponds to the prediction that the crime rate will be as much as μ_c of crime type c regardless of the region.

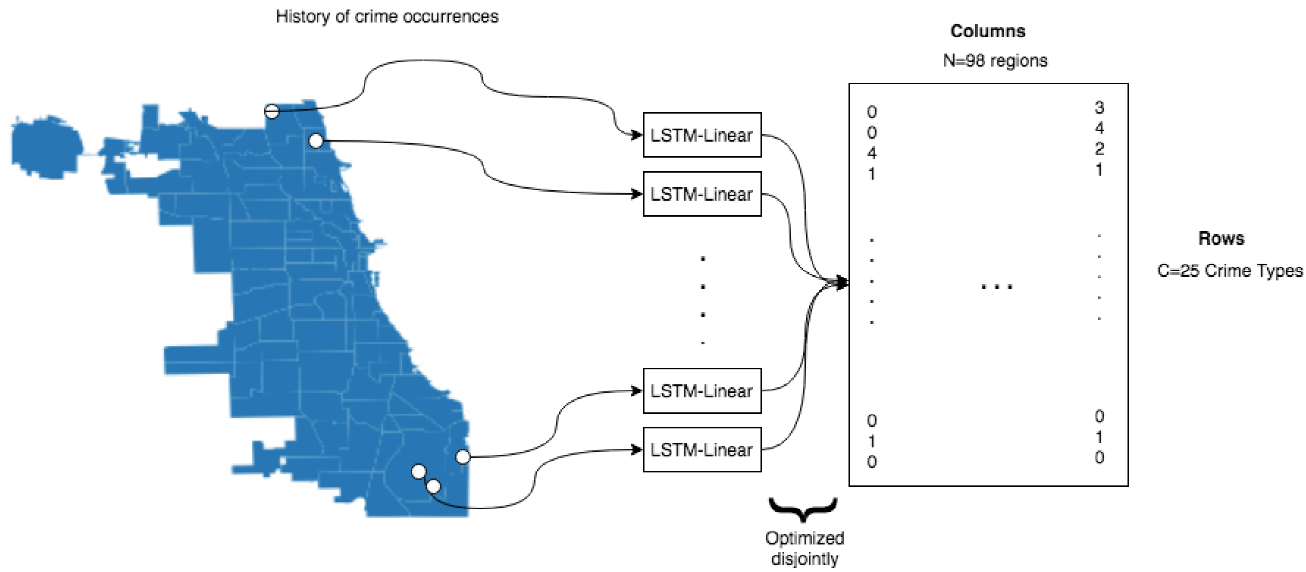

Figure 2: A diagram that describes Regional LSTMs Model.

3.5. Regional LSTMs

3.5.1 Long Short Term Memory Networks

In this section, W is the recurrent connection at the previous hidden layer and current hidden layer, U is the weight matrix connecting the inputs to the current hidden layer.

$$i_t = \sigma\big(x_t U^i + h_{t-1} W^i\big)$$

$$f_t = \sigma\big(x_t U^f + h_{t-1} W^f\big)$$

$$o_t = \sigma\big(x_t U^o + h_{t-1} W^o\big)$$

$$\tilde{C}_t = \tanh\big(x_t U^g + h_{t-1} W^g\big)$$

$$C_t = \sigma\big(f_t * C_{t-1} + i_t * \tilde{C}_t\big)$$

$$h_t = \tanh(C_t) * o_t$$

where we i, f, o are defined the input, forget and output gates, respectively.

3.5.2 Training Loss

We employ mean square error loss for our multi target regression task.

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of future time bins.

3.5.3 The Model

Regional LSTMs model is combination of multiple LSTM networks followed by a linear layer for regression for each region in K which are trained separately. We fit each LSTM in the network via autoregressive procedure [2], namely given previous day(s)' crime intensities, we to predict next crime intensity. We will refer Regional LSTMs as R-LSTMs. The Figure 2 describes the procedure of Regional LSTMs. The following Algorithm 1 summarizes the training procedure of Regional LSTMs model.

Algorithm 1 Regional LSTMs Training Procedure

procedure R-LSTMS
  $R \leftarrow$ A list containing indices of regions K
  $D \leftarrow$ historical data in the current split, tensor X
  for each c **in** R **do**
    model $\leftarrow$ Create an LSTM model
    $W \leftarrow$ Data of the region c, tensor D_r
    Train *model* using training split of W
      early stop based on validation split of W
    $\hat{y}_c \leftarrow$ Predict by *model* on test split of W
    $R_c^2 \leftarrow$ Calculate R_c^2 score using $\hat{y}_c$
  $R^2 \leftarrow \frac{1}{|R|}\sum_{c\in R} R_c^2$
  return R^2

4. Implementation Details

4.1. Technical Specifications

We used Pytorch to implement our LSTM based models. Pytorch is a python based deep learning library which handles automatic differentiation of the computational graph of neural network models efficiently [14]. We also used Scikit-learn [15], a machine learning library in Python, and numpy [20], a numerical computation library in Python, for utility methods such as R^2 calculation. We ran the experiments on a machine using an NVIDIA GeForce 1080 TI GPU with 12GB memory.

4.2. Preprocessing Dataset

The raw version of Chicago Crime dataset was unclean. We removed exact duplicate rows except first ones. We removed rows containing empty latitude, longitude, or type columns. We removed all rows except the ones in date range from January 1, 2015 to December 31, 2016 inclusive. We removed all crime entries outside of the Chicago boundaries. We removed crimes that belong to rare crime types, namely whose crime type's occurrence is less than 100. Chicago districts are obtained via the district boundaries downloaded from [1].

4.3. Hyperparameters

Time bins are chosen as a day for usability purposes. For comparability, we chose same hyperparameters for all neural network based models.

We trained R-LSTMs for each region separately with same hyperparameters for each region. We trained all LSTM based models with 2000 epochs, and early stopping patience 50. We chose learning rate as 0.1. Optimized with Adam optimization algorithm [11]. Previous inputs are tried for either a day or a week for interpretability purposes instead of choosing them arbitrarily. All LSTMs consist of 3 layers containing 16, 32, and 16 cells respectively. LSTM networks are followed by 25 units of linear layer for our regression task.

5. Dataset

We employ Chicago Police Department's crime dataset to model crime prediction task. The dataset contains many information such as street name, but we only use crime type, location in terms of latitude and longitude, the time of occurrence of crime. There are 34 different types of crimes in the dataset: Arson, assault, battery, burglary, concealed carry license violation, sexual assault, criminal damage, criminal treepass, deceptive practice, gambling, homicide, human trafficking, interference with public officer, intimidation, kidnapping, liquor law violation, motor vehicle theft, narcotics, non-criminal, obscenity, offense involving children, other narcotic violation, other offense, prostitution, public indecency, public peace violation, ritualism, robbery, sex offense, stalking, theft, weapons violation. The location boundaries of crimes are limited with the city of Chicago. Also, the time of crimes' occurences are from 2002 to 2017.

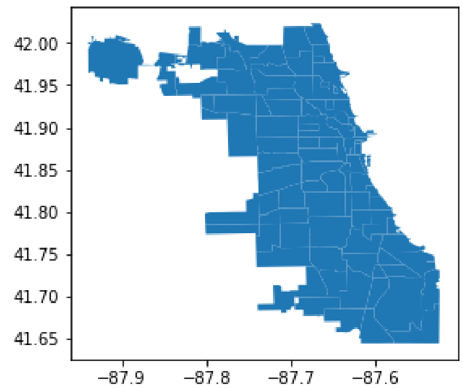

Figure 3: Visualization of boundaries of districts in the City of Chicago.

Crimes in the dataset are sparse when crime types are separated. The size of the neighborhood regions are not equal as seen in Figure 3. We also summarized the statistics of the Chicago P.D.'s Crime dataset after preprocessing steps in Table 2.

Table 2: Chicago Police Department's Crime dataset statistics after our preprocessing step.

# of Crime Types	25
# of Regions	98
# of Unique Locations	205.593
# of Days	730
Date Range	[Jan 1, 2015 - Dec. 1, 2016]
Latitude Range	(41.64, 42.02)
Longitude Range	(-87.92, -87.52)

5.1. Cross Validation Splits

The Chicago crime dataset is splited to time series splits as described in section 3.3. In each split, days are consecutive. Also, splits are ordered such that training set always places before the validation and the validation set always places before the test set. The bumber of consecutive days in the splits of the Chicago P.D.'s crime dataset after our preprocessing can be found in Table 3.

Table 3: Number of consecutive days in the splits of the Chicago P.D.'s crime dataset after our preprocessing.

Split	**Training**	**Validation**	**Test**
Split 1	165	19	182
Split 2	329	37	182
Split 3	493	55	182

Table 4: Comparison of Single vs. Regional LSTMs on Chicago P.D.'s crime dataset.

Model	AR Parameter	Validation R^2	Test R^2
Regional LSTMs	**AR(1)**	**0.137**	**0.161**
Regional LSTMs	AR(7)	0.134	0.160
Single LSTM	AR(1)	0.136	0.148
Single LSTM	AR(7)	0.113	0.114

6. Experiments

To our knowledge, there is no previous work either makes predictions for districts or does regression. Thus, our results are not comparable to the previous works.

6.1. Validating Experimental Setup

Classification Experiments are done in order to compare our framework which is based on regression with classification. Also, many metrics and experimental setups suffer from the class imbalance, and sparsity. Thus, a validating the experimental setup is needed. To validate our experimental setup, we tested whole data with 4 different meta predictors. The scores for meta predictors are calculated via testing on whole data without splitting into train and test since they do not require training phase.

We convert the problem to multi label classification task [19] by rounding the predictions greater than or equal to 0.5 to 1, and the values less than 0.5 to 1. After calculating the accuracies per class, we take average of accuracies uniformly. Thus, classification accuracy metric is between 0 and 1.

Classification accuracy and R^2 scores can be found in Table 5. The differences in R^2 scores of regression task seems more interpretable compared to accuracy scores of classification task. We concluded to make experiments with using regression settings for further experiments.

Table 5: Results of experiments with meta predictors on whole Chicago P.D.'s crime dataset.

Model	R^2 Score	Classification Accuracy
Mean Per Class	**0.000**	**0.852**
All Ones	-57.288	0.155
All Zeros	-0.262	0.845
All Total Mean	-1.295	0.845

6.2. Ablation Study of R-LSTMs

To show the improvement of training an LSTM for each region instead of one for all, we developed a Single LSTM model as an ablation study.

Single LSTM model is exact same component of R-LSTM's one region's LSTM with its linear component applied to all regions one by one. We trained an Single LSTM model with 2000 epochs, and early stopping patience 50. We chose learning rate as 0.1. Optimized with Adam optimization algorithm [11]. Previous inputs are tried for either a day or a week for interpretability purposes and to faciliate ablation study with regional LSTMs. We will refer Single LSTM model as S-LSTM. LSTM consist of 3 layers containing 16, 32, and 16 cells respectively. LSTM networks are followed by 25 units of linear layer for our regression task.

AR parameter is the number of days for in every input of the sequence.

Experiment results for both R-LSTMs and S-LSTM can be found in Table 4. Our proposed R-LSTMs model with AR(1) parameter explains 16% of the variation within the test set. R-LSTMs show 1.3% improvement in explaining the variation in test set.

R-LSTMs seems more consistent for AR parameter compared to S-LSTM since there is much more difference when AR parameter of S-LSTM changes.

6.3. Exploration of Neighborhood Ensembles

We discovered temporal relations with LSTM based models and we explored spatial relations through region based R-LSTMs and showed their improvement over S-LSTM which is non-regional. This improvement led us to dig on spatial relations more so we built an ensemble model to explore the spatial neighborhood of regions. From R-LSTMs we have an LSTM model for each region. Using these LSTMs, we employ an ensemble of neighbor regions' models.

First, we train an LSTM for each region using the training data of each region separately. Later, we found the neighbors for each region based on the corners of the polygons representing the regions using the called "Queen's case" [5] and converted them into a graph as seen in Figure 4. Lastly, for a region, we take average of the predictions using that region's model and the region's neighbors' models. We repeat these procedures for each region. We will refer Ensemble of Neighbor R-LSTMs as ER-LSTMs. The comparison of our best model with exact same parameters exact ensembling procedure shown in Table 6. ER-LSTMs show 10% diminishment in explaining the variation within the data compared to our proposed R-LSTMs model with AR(1) parameter.

Figure 4: Districts as nodes where neighboring districts are connected where neighboring is based on "Queen's case" [5].

Table 6: Comparison with Neighborhood Ensembles on Chicago P.D.'s crime dataset.

Model	Validation R^2	Test R^2
R-LSTMs AR(1)	**0.14**	**0.16**
ER-LSTMs AR(1)	0.04	0.05

The results most probably caused because of the averaging procedure makes all regions' predictions closer to to global mean of each crime type since its R^2 is closer to the Mean Per Class model in Table 5.

7. Conclusion

We have presented a novel experimental setup with its baselines for the problem of crime rate prediction. We also proposed a novel model for crime rate prediction R-LSTMs and we have done its ablation study by comparing it with S-LSTM. Although we could not explain much of the variation in the data, we obtained 1.3% improvement by R-LSTMs compared to S-LSTM model and Class Per Mean model.

We recommend to employ Graph Convolutional Network architecture [12] by embedding each time bin using the graph we proposed in section 6.3.

References

[1] Boundaries - neighborhoods — city of chicago — data portal. 5

[2] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969. 2, 4

[3] P. J. Brantingham, P. L. Brantingham, et al. *Environmental criminology*. Sage Publications Beverly Hills, CA, 1981. 1, 2

[4] X. Chen, Y. Cho, and S. Y. Jang. Crime prediction using twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium*, pages 63–68. IEEE, 2015. 2

[5] G. Chi and J. Zhu. Spatial regression models for demographic analysis. *Population Research and Policy Review*, 27(1):17–42, Feb 2008. 6, 7

[6] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014. 2

[7] T. Guo, T. Lin, and Y. Lu. An interpretable lstm neural network for autoregressive exogenous model. *arXiv preprint arXiv:1804.05251*, 2018. 2

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[9] C. Johnson. Crime mapping and analysis using gis. In *Geomatics Conference on Geomatics in Electronic Governance, January*, 2000. 1

[10] H.-W. Kang and H.-B. Kang. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, 12(4):e0176244, 2017. 2

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 6

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 7

[13] T. T. Nguyen, A. Hatua, and A. H. Sung. Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology Vol*, 8(2), 2017. 2

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. 5

[16] D. J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in

geographic information systems. *Annals of the Association of american Geographers*, 84(3):441–461, 1994. 1

[17] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 139–148. ACM, 2015. 2

[18] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. In *INTERSPEECH*, 2012. 2

[19] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. 6

[20] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011. 5

[21] B. Wang, D. Zhang, D. Zhang, P. J. Brantingham, and A. L. Bertozzi. Deep learning for real time crime forecasting. *arXiv preprint arXiv:1707.03340*, 2017. 1, 2

[22] R. Wortley and M. Townsley. *Environmental criminology and crime analysis*. Taylor & Francis, 2016. 1

[23] J. Zhang, Y. Zheng, and D. Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[24] X. Zhao and J. Tang. Crime in urban areas:: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 20(1):1–12, 2018. 1, 2