Using Hetero-ConvLSTM to predict criminal activity in Chicago, Illinois

Group 1
Ruben Ahrens, Lucas de Wolff

30th November 2022

1 Introduction and summary of the selected paper

The general topic of this paper [1] is predicting traffic accidents to improve public safety. The main idea is that being able to predict traffic accidents will enable the prevention of traffic accidents. With this, transportation, public safety and safe routing can all be improved. Predicting accidents is challenging as accidents are rare in space and time resulting in sparse data. This paper uses heterogeneous spatial data as opposed to previous research which focused only on a homogeneous study e.g. a city. This paper uses a Convolutional Long Short-Term Memory Neural Network model to predict traffic accidents. The NN model is trained on spatial-temporal data. This data describes environmental features related to driving conditions in the State of Iowa. Features such as weather, environment, road condition, and traffic volume.

2 Problem statement

In the Chicago area, (violent) crime is a huge issue, its violent crime rate is higher than the US average. The police force works hard to fight against criminal activity. There have been numerous studies [2][3] that developed machine learning models capable of predicting crimes. Crime prediction can help police forces to anticipate criminal activity and for example, achieve efficient police patrolling routes. However, to our best knowledge, there are no models that tackle the heterogeneous data in a similar (and effective) way as [1] did in the predicting traffic accidents context. We will apply the Hetero-ConvLSTM used in [1] to criminal data. As mentioned, deep learning has already been applied to crime prediction, in particular, [2] used a deep NN to predict criminal activity in Chicago. LSTM models have also already been tried for crime prediction [4]. However, in this paper they predict the crime rate in a neighbourhood, instead of pinpointing locations, which is what we will be doing. Because of it's crime rate, previous research, like the two mentioned papers, on criminal activity is often times done on the city of Chicago. This will make our results comparable to other research. Our solution should be able to generate accurate new data on criminal activity in the future. So specific crime locations should be predicted.

The fundamental difference in data sources between our project and the original paper are space, scale and features. The space of the paper's data are from instances based in Iowa. In our research, we will focus on Chicago. Chicago will have other data sources and those sources might have different formats. We will need to preprocess this data to be able to use this for the the Hetero-ConvLSTM model.

Another challenge, is the change in scale. In the previous paper, the authors assumed that rural, urban and mixed areas needed to be modelled separately. In our case, we only consider the urban environment, thus the assumption in our context will be that urban and suburban regions have to be modelled separately. Our intuition behind this is that we expect more crime to take place in suburban neighbourhoods than in urban neighbourhoods.

As the Hetero-ConvLSTM is specifically applied for predicting traffic accidents, the features are expected to be different. For example, a feature like the weather is expected to have more impact on traffic accidents than on criminal activity. In general, a big challenge is finding the right features for predicting criminal activity. We introduce different and new data like light intensity, that have not been used in the original paper.

3 Research questions

Research question:

Is Hetero-ConvLSTM capable of providing accurate predictions of criminal activity in Chicago?

Subquestions:

- How will the Hetero-ConvLSTM perform in predicting criminal activity compared to the historical average method?
- How do we tackle the heterogeneity in the crime data?
- How do we handle the temporal and spatial sparsity of the crime data?

4 Methodology

We will start by collecting the data described in section 6. We will then apply the methods in the original paper to criminal activity prediction. These methods require us to further investigate some techniques covered in the lectures. We will investigate:

1. Spatial auto-correlation, which the convolution layer specifically tries to capture.
2. Temporal auto-correlation, which the LSTM specifically tries to capture.
3. LSTM, to create a neural network architecture that takes temporal patterns into consideration.
4. Convolution, to extract information out of spatial/image data and compress it into features.

Combining the knowledge of these techniques, we will be able to understand and build our own Hetero-convLSTM. Hopefully, we will be able to outperform the historical average baseline.

5 Evaluation approach

Metrics: To evaluate our work we will use the same metrics used by the original paper: Mean Squared Error, Root Mean Squared Error and cross Entropy. Evaluating whether a prediction is in the same grid as the real data.
Baselines: We will compare the results of the Hetero-ConvLSTM with the results of the historical average method. We will also try to analytically compare our results with other studies that tried to predict the crime location in Chicago e.g. [2] and [4]. The methods may not be exactly comparable to other research, because the data used or evaluation approach might not be exactly the same. Therefore, we will not be able to conclude one method is better or worse, but we can compare them analytically and discuss their differences.

6 Data sources and other resources

Criminal records data of Chicago: Chicago crime data - https://data.cityofchicago.org/Public-Safety/Crimes-One-year-prior-x2n5-8w5q/data

Light intensity map mask created from ultra high-resolution satellite image: Light intensity map - https://www.nasa.gov/sites/default/files/thumbnails/image/26247384716_9281df96cc_o.jpg

Weather data (temp, rainfall, wind): Weather data - https://www.visualcrossing.com/weather/weather-data-services

Daytime satellite image, as well demographic data, education data. Public Schools - https://www.kaggle.com/datasets/chicago/chicago-public-schools-data

Shapefile of Chicago with neighborhoods: Shapefile - https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9

Ethical statement

A concern for predicting crime is that the historical data can be skewed towards locations where minorities reside. When there's more police enforcement in a certain area, there will also be more data distributed in that area. To combat this we can make the algorithm a bit more exploitative so it doesn't continue in a vicious cycle of minority oppression. Another ethical consideration is that it might be good to disregard harmless crimes like drug possession, as it might not be desirable to predict such light offences. A rational police force would want to spend its time fighting violent and or organized crime.

Division of workload

We share responsibility for each part of the work. The project consists of the following tasks:

- Data
  - Data Collection: Ruben
  - Data Preprocessing/Cleaning: Lucas
  - Data Exploration: Ruben
  - Explore spatial/temporal autocorrelation: shared
- Algorithms
  - Implement ConvLSTM: shared
  - Historical average baseline: Lucas
- Results
  - Heatmap on Chicago shapefile: Ruben
  - Scatterplot on chicago shapefile of predictions: Lucas
  - Compare analytically with [2],[3],[4]: shared

Code

GitHub repo: https://github.com/boomerr1/CrimeChicagoUC

References

[1] Zhuoning Yuan, Xun Zhou and Tianbao Yang. “Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD ’18. London, United Kingdom: Association for Computing Machinery, 2018, 984–992. ISBN: 9781450355520. DOI: Yuan:Traffic. URL: https://doi.org/10.1145/3219819.3219922.

[2] Hyeon-Woo Kang and Hang-Bong Kang. “Prediction of crime occurrence from multi-modal data using deep learning”. In: *PloS one* 12.4 (2017), e0176244. URL: https://doi.org/10.1371/journal.pone.0176244.

[3] Xinyu Chen, Youngwoon Cho and Suk Young Jang. “Crime prediction using Twitter sentiment and weather”. In: *2015 Systems and Information Engineering Design Symposium.* 2015, pp. 63–68. DOI: 10.1109/SIEDS.2015.7117012.

[4] Xinge Han, Xiaofeng Hu, Huanggang Wu, Bing Shen and Jiansong Wu. “Risk Prediction of Theft Crimes in Urban Communities: An Integrated Model of LSTM and ST-GCN”. In: *IEEE Access* 8 (2020), pp. 217222–217230. DOI: 10.1109/ACCESS.2020.3041924. URL: https://ieeexplore.ieee.org/abstract/document/9276416.