Towards Crime Rate Prediction through Street-level Images and Siamese Convolutional Neural Networks

Conference Paper · December 2017

CITATIONS
0

READS
195

1 author:

Marco Birck
Universidade Federal de Pelotas
6 PUBLICATIONS **1** CITATION

SEE PROFILE

Towards Crime Rate Prediction through Street-level Images and Siamese Convolutional Neural Networks

Virginia Ortiz Andersson, Marco Antônio Ferreira Birck, Ricardo Matsumura Araújo and Cristian Cechinel

[1]PPGC – Federal University of Pelotas (UFPel)
Rua Gomes Carneiro, 1 - Centro - CEP 96010-610 – Pelotas – RS – Brazil

{vandersson,mafbirck,ricardo}@inf.ufpel.edu.br,cris16marte@yahoo.com.br

Abstract. *The analysis of the environment for crime prediction is based on the premise that criminal behavior is influenced by the nature of the environment in which occurs. Street-level images are the closest digital depiction available of the urban environment, in which most street crimes take place. This work proposes a crime rate prediction model that uses street-level images to classify street crimes into four crime rates levels (from low to high). For that, we use a 4-Cardinal Siamese Convolution Neural Network (4CSCNN) and train and test our analytic model in a region of Chicago, US, that showed high street crime concentrations between the years of 2014 and 2015. With this experiment, we investigate the use of convolutional neural networks (CNN) for the task of crime rating trough visual scene analysis and found possibilities towards automatic crime rate predictions using CNN models.*

1. Introduction

Several works have been developed to understand the factors that trigger a criminal event carried out by an individual, as well as the risks involved and the measures to avoid it. Over the years, theories have been developed to map criminals behavior and crime itself. "Environmental theories" consider crime as a confluence of *offenders*, *targets* and specific *laws and settings* at *particular times and places* [Wortley and Mazerolle 2008] [Brantingham and Brantingham 1995]. In here, offenders are not the central object of interest, but one element of a *crime event*.

According to [Wortley and Mazerolle 2008], the environmental perspective is based on the premise that criminal behavior is influenced by the nature of the environment in which occurs, i.e., the environment plays a fundamental role in initiating crime and shaping its course. The distribution of crime in time and space is non-random and it depends on environmental factors and situations. Control and crime prevention are then a result of understanding the role of the environment on crimes patterns. Environmental theories can focus on (i) how offenders react in the environment, such as the *Routine Activity* [Cohen and Felson 1979], the *Rational Choice* [Brantingham and Brantingham 1993], and the *Crime Pattern Theory* [Eck and Weisburd 1995], and (ii) how to map the physical environment in which the criminals operate, such as the well known *Broken Windows Theory* [Wilson and L. 1982] and the *Routine Active of Places* [Sherman et al. 1989].

Street-level images are the closest depiction of the human environment available in digital form, and their use in daily life is gradually increasing, mostly to aid navigation. The Google Street View [Google 2017] service popularized the access to such

street images and computer vision models are being used together with street-level images to relate a city's physical appearance with crime statistics [Arietta and Efros 2014] [Khosla et al. 2014] [Gebru et al. 2017].

In this work, we present an initial study on crime rate prediction models that use street-level images as input. We propose a 4-Cardinal Siamese Convolution Neural Network (4-CSCNN) together with a Multi-layer Perceptron to classify visual scenes depicted in street-level images into four categories of crime rate, from low to high rates. We built a dataset with street-level images from a region of Chicago City, US, that showed high street crimes rates between the years of 2014 and 2015. This region was divided into equal sized cells containing the total crime events that happened inside the cell region. Images belonging to the cell received the label according to the total crimes in that region. We use this street-level images dataset to train and test our proposed model, obtaining 54.3% of overall accuracy and 77% of average accuracy per class, in the classification of the four crime rate categories. The achieved results are discussed in Section 4.

This paper is organized as follows. Section 2 presents background and related work to Crime Prediction Models, Visual Scene and Environment Analysis and Siamese Convolutional Neural Networks. Section 3 describes how data was collected together with the methodology followed by our approach. Section 4 discusses the most important findings of our work, and Section 5 presents some conclusions and ideas for future work.

2. Related Work

Our paper is related to (i) Crime Prediction Models, that are the main purpose of this work, (ii) Visual Scene and Environment Analysis, which inspired the use of visual attributes in our proposed model and (iii) Siamese Convolutional Neural Networks, the deep learning technique which is used in our model.

2.1. Crime Prediction Models

Based on *Routine Activity of Places Theory* [Sherman et al. 1989], crime *hot spots* are regions where high concentration of crime events is observed. The *hot spots* were initially used as a criminal data visualization technique, and further became a prediction model, with the advance of statistical and geographic information (GIS) tools, and the support of the observed characteristics of crime events *repeat* and *near-repeat* [Sherman et al. 1989]. *Hot spots* were used by [Block 1998], in the *Early Warning System*, that used data collected by the community and law enforcement to produce the *hot spots* and point out possible near regions affected by violent crimes.

In [Bowers et al. 2004], authors proposed the use of a technique to distribute crime in a geographic surface and calculate risk assessment of crime events - the prospective risk surface - to obtain *hot spots*. This technique consists of using a two-dimensional grid with n equally sized cells overlying the geographic region of interest. In the model, a weight is associated with geographically located crimes. Recent crimes that happened near the center of a cell receives a higher weight. The weights of all crimes near the center are added together to produce the risk index of the belonging cell. To evaluate the proposed model, the authors used "theft" data from Merseyside County, England, from the year of 1997. The model presented 62% to 64% of accuracy in predicting crimes for a 2-day time window.

Later, [Chainey et al. 2008] proposed the use of *Kernel Density Estimation* (KDE) to map *hot spots* overlaying a geographic area. It smooths the criminal data over a region according to a kernel density estimator function, mapping probabilities for a crime event happening under a specific area. In [Johansson et al. 2015], the authors evaluate the KDE approach reaching 76% to 84% of accuracy in predicting crimes in a 3-month time window. More recently, [Gerber 2014] proposed a modification in KDE estimator to hold independent variables. The author used *tweets* from a specific location in Chicago city, Illinois, correlated with crime events in the location.

Risk Terrain Modeling (RTM) was proposed by [Caplan et al. 2011] to assess crime risk over a region. RTM consist of acquiring crime-related factors and standardize each factor to a common geographic region, usually assigning a weight to the presence or absence of this factor at every place covered by the region of interest. As the risk value over a region gets higher, the probability of a crime event occurring in that region also gets higher. According to [Caplan et al. 2011], the RTM technique produces maps that indicate regions with greater risks of becoming *hot spots* in the future. In [Drawve et al. 2016], the authors propose an aggregate neighborhood risk of crime (ANROC) measure, applying RTM model to forecast neighborhood-level of violent crimes in Little Rock, Arkansas. They identify 14 risk factors, such as banks, bus stops, check-cashing, convenience stores, fast-food restaurants, grocery stores, hotel/motels, liquor stores, pawn shops, tattoo/piercing shops that were expected to influence crime. The ANROC measure was obtained by averaging the risk of crime per cell by neighborhood.

2.2. Visual Scene Environment Analysis

Computer vision and machine learning are used extensively to discover environment attributes in street-level images. According to [Dubey et al. 2016], visual scene and urban imagery analysis can be focused on different objectives, such as predicting perceptual responses from images, understanding cities by their visual urban scene, understanding the connection between urban appearance and socioeconomic factors and rank or compare different urban environments. The strains which the present work is interested in are understanding cities and the connection between urban appearance and socioeconomic factors.

In [Doersch et al. 2012], the authors proposed a methodology to automatically find visual elements (e.g. windows, balconies, street signs) from street-level images that are distinctive for a specific geographic area, i.e., they occur much more frequently in that area than other areas. The authors used Google Street View, a street-level image database, and clustered images using Histograms of Oriented Gradient (HOG) and color components to compose descriptors for specific patches (e.g. windows, signs, doors). They clustered the patches using Nearest Neighbor algorithm, dividing positive and negative data clusters into l equally-sized subsets. Next, they iteratively trained an SVM detector for each visual element using the detectors trained in the previous iteration in a new unseen subset, selecting top k detection for retraining. As results, for Paris, the authors achieved 83% of accuracy (where chance yields 50%) and for Prague City, 92% of accuracy. Later, [Arietta and Efros 2014] followed the proposed model of [Doersch et al. 2012] and applied Clustering and Support Vector Regression (SVR) to USA cities, discovering predictive relationships between visual elements from the environment and non-visual variables like crime and theft rates, housing prices, population density, graffiti density and percep-

tion of danger. They compared the use of HOG+color descriptors with the fifth convolutional layer of Caffe's ImageNet CNN model and concluded that HOG+color descriptors were more visually consistent but captured less city semantics.

In [Khosla et al. 2014], the authors explored the ability to use visual scenes to predict distances of surroundings establishments such as hospitals and fast-food restaurants. They also explored the possibility of predicting crime rates in an area using visual scenes. The authors experimented different descriptors, e.g. GIST, Texture using Local Binary Pattern (LBP), Color using Locally-Constrained Linear Coding (LLC), HOG+Color and FC7 layers from Caffe's ImageNet CNN model. They used an SVR machine on the image features obtained by each descriptor. The results achieved in finding hospitals and fast-food restaurants showed similar results between Color, HOG+color and Deep Learning descriptors, varying from 0.58% to 0.61% of accuracy. To predict crime rates and danger perception, the authors used HOG+color descriptors and the SVR machine and compared to a human test prediction. The results achieved for crime rate prediction was an accuracy of 72.0%, compared to 59.6% in human tests.

2.3. Siamese Convolutional Neural Networks

In [Bromley et al. 1993], Siamese Neural Networks are defined as "twin" Neural Networks (NNs) which share their parameters. The main reason to use twin NNs is that the inputs will be mapped to a very near space since they are being processed by the same function e.g. Convolutional Neural Networks (CNNs). The Siamese NNs are widely used to discriminate and match tasks - e.g. [Bromley et al. 1993],[Taigman et al. 2014], [Lin et al. 2015], [Zagoruyko and Komodakis 2015]. Recently, focusing in city places safeness classification through street-level images, [Dubey et al. 2016] proposed the Place Pulse 2.0, a novel dataset containing a pairwise comparison of "safe" and "unsafe" images from 56 cities across the world. The authors also proposed two different Siamese CNN architectures: Streetscore-CNN (SS-CNN), to predict a winner between two images comparison as safe and unsafe, and a Ranking Streetscore-CNN (RSS-CNN), where a ranking function is attached to the SS-CNN to rank the street-level images into high to low safeness scores.

A 4-Cardinal Directions pseudo-Siamese Convolutional Neural Network was proposed by [Lieman-Sifry 2016], where the author used CNNs to extract a low-level representation of images from geographical points in Colorado, USA state, aiming at determining the location of the images. The CNNs architecture used was based on ResNet [He et al. 2015] and AlexNet [Krizhevsky et al. 2012], reserving two networks per cardinal direction, using the method of ensembles in conjunction. The network learns "from scratch" to classify regions from Colorado state. The representations mapped by the author's CNNs for each cardinal direction are concatenated and passed through two fully connected Multi-layer Perceptron layers (MLP) that classify the representation according to the given labels.

3. Methodology

We present in this work a model to classify crime rates in a city region, using only visual attributes extracted from street-level images from specific locations. The main objective of the model is, given a georeferenced location (point), predict the historical crime density

in that location based on the images that surround it, in the four cardinal orientations. We intend to understand how the visual scene influences the crime rate in each place.

3.1. Crime Data

The crime data acquired for this work was retrieved from Chicago City data portal [CityOfChicago 2017]. The dataset contains records of crime events georeferenced since 2001. We chose to focus on crime data between in the year of 2014 and 2015. To visualize and identify crime hot spots in Chicago, we create a grid-like data structure of shape 50 x 50, dividing the limits of the city into 2.500 equally spaced cells and distributed the crime records that happened in the street or sidewalk into the corresponding cells, using latitude and longitude. Figure 1(a) shows crime events that happened in "street" or "sidewalk" in the year 2014 grouped by a cell. Darker shades of blue are used to denote that at least one crime occurred in that area, and more crimes per cell are indicated by lighter and hotter shades of blue, green, orange and red. We choose the area with higher crime rate concentration, highlighted with a red square in Figure 1(a), for our study.

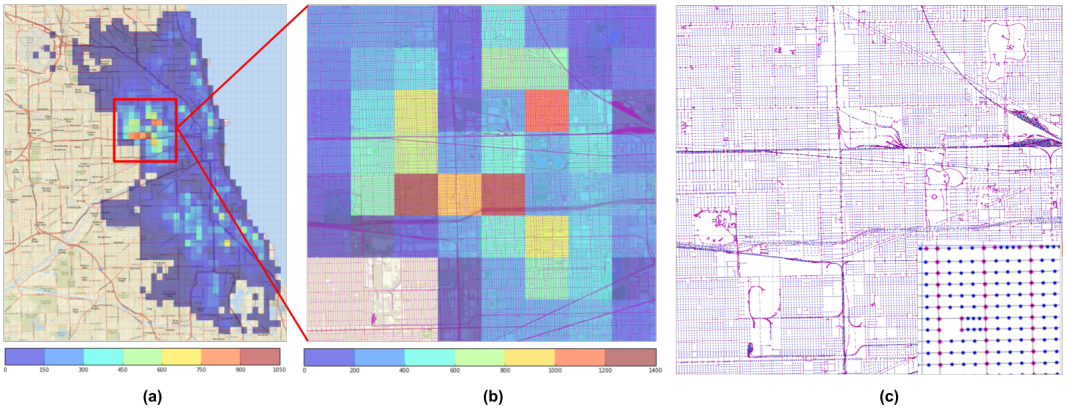

Figure 1. (a) Hot spots of street and sidewalk crimes in Chicago, in the year 2014. Crimes are distributed in a 50 x 50 grid, with low to high concentrations depicted in the color scale, showing total crime range by color. The region with more concentrated crimes between neighbors cells is highlighted with a square. In (b), streets of the region of interest with street and sidewalk crime hot spots in the year of 2014. The points that shape the streets of the selected area in Chicago are presented in (c), with the street corners depicted (lower right detail) as diamonds and the interpolated points as circles.

3.2. Obtaining and Labeling Street-level imagery

From the selected region in Figure 1(a), we create a new grid data structure of shape 8 x 8 with 64 equally spaced cells, corresponding to the region of interest in the original grid. Next, we obtain a *shapefile* from streets distribution of the selected region using Mapzen Metro Extracts service [Mapzen 2017], superimposing the *shapefile* vectors on the map, as depicted in Figure 1(b). The *shapefile* of Chicago streets contains 2D points, or vertices, that shape the lines of streets. We transformed the vertices into latitude and longitude coordinates and interpolated points between the vertices, as depicted in Figure 1(c). With Google Street View API [Google 2017] we obtained 4 images, from the vertices and interpolated points that shape the streets, corresponding to the 4 cardinal directions. Each point with their 4 images was distributed by grid cell. Next, each cell

received a label according to the crime distribution in the region of interest. We categorized the distribution into "blue", "green", "orange" and "red" to indicate low to high street crime rates.

Using the methodology described, we built a street-level imagery dataset with a total of $20,056$ points, with $80,226$ images for each point. Each point belongs to a specific cell and a specific label according to the crime rate in the cell region. To build the train and test dataset, a number of georeferenced distinct points were randomly chosen for each label. Table 1 shows the total number of points for each label - blue, green, orange and red - and total number of images in the train set. The test set was adjusted to have the same number of points based on the smallest number of examples a label could have.

Table 1. Image dataset composition arranged by labels, in train and test set.

Train set			**Test set**	
Label	Points	Images	Points	Images
Blue	3,128	12,512	1,073	4,292
Green	4,128	16,512	1,073	4,292
Orange	4,327	17,308	1,073	4,292
Red	4,181	16,724	1,073	4,292
Total	15,764	63,058	4,292	17,168

3.3. Proposed Siamese CNN

Our work uses as inspiration the conjunction of ideas from [Zagoruyko and Komodakis 2015] and [Lieman-Sifry 2016]. While [Lieman-Sifry 2016] uses a variant of a Siamese network, as [Zagoruyko and Komodakis 2015] stated, this method can be named as a Pseudo-Siamese, because the CNNs used are free to learn specific features to each image subset. Our proposed model, however, uses the concept of Siamese Neural Network in which all the CNN have their weights "frozen", and the previous weights domain reused. Figure 2 shows the proposed Siamese CNN architecture, showing the 4-cardinal CNNs with shared and frozen weights trained with *ImageNet* and fully connected Multi-layer Perceptron (MLP) layers (FCs). The CNNs follows the AlexNet architecture [Krizhevsky et al. 2012], making our model more robust compared to the simpler CNN architecture used in [Lieman-Sifry 2016]. Each CNN is frozen, following a common approach of *transfer learning*, allowing the use of pre-trained weights from the *ImageNet* domain. This way, the model is less likely to overfit and has the ability to leverage knowledge from the *ImageNet* domain.

In the proposed model architecture, the resulting CNNs outputs, a smaller spatial representation of the images, are inputted in the *Fully Connected FC 1024* MLP units that are free to learn. Next, the 4 outputs are concatenated and passed trough another classifier, an MLP that has been proved to be very efficient in conjunction with CNNs. The MLP has two fully connected layers, the *FC1 4096* and *FC2 4*, where the backpropagation algorithm [Rumelhart et al. 1988, LeCun et al. 1998] is applied, and the output is passed to a *softmax* classifier to extract the distributed probability over the labels.

3.4. Experiments

The proposed model was trained using Keras [Chollet 2015] and Tensorflow [Martín 2015] on a NVIDIA graphic processing unit (GPU) with 8GB of memory. As

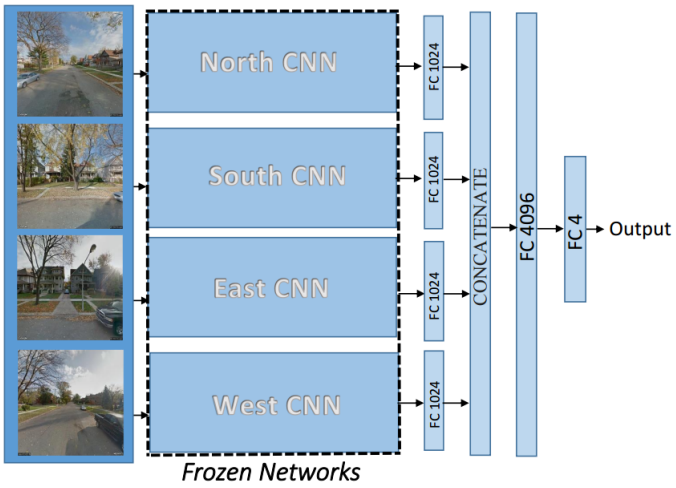

Figure 2. Proposed architecture with 4-cardinal Siamese CNNs, with Street View images following 4 cardinal orientations, using frozen *ImageNet* weights and fully connected layers followed by a *softmax* classifier.

the CNNs layers were frozen, the training process was faster compared to training the CNNs "from scratch", allowing only the fully connected (FC) layers to be trained. The loss function chosen was categorical *cross-entropy*, due to the multi-label nature of our problem. The *adaDelta* optimizer [Zeiler 2012] was used to train the model, with the ability of automatically adapt the learning rate. To improve generalization, we select 32 random images of our image dataset and applied a data augmentation method, consisting in random (i) image cuts with $range = 0.2$ i.e. the shear angle in counter-clockwise direction as radians, (ii) image zooms, with $zoom - range = 0.2$ in a random mode defined as $[lower, upper] = [1 - zoom - range, 1 + zoom - range]$ and also (iii) horizontal flips, helping our model to generalize better and deal with the data bias.

In the training step, our model receives as inputs $n^i \in N$, $e^i \in E$, $s^i \in S$ and $w^i \in W$, where N, E, S, W are subsets of images from our dataset composed of Google Street View images, taken each from the cardinal direction North, East, South and West respectively. Additionally, i is a geographical point of a street that belongs to a cell with a labeled crime rate. Each one of these four images was passed trough the model in batches of 32 geographic points for the mini-batch training. Each image linked to one cardinal direction was passed trough the respective CNN to obtain a representation in a low dimensional space, a descriptor with deep features of the image. All image descriptors are then concatenated to represent a new low-dimensional descriptor of the whole geographical point. Furthermore, the resulting descriptor is passed through the fully connected MLP layer whose objective is to map the geographic point to one of the four crime rate labels.

Training was executed with a maximum of 600 epochs, and checkpoints in 150, 360 and 600 epochs were set to observe possible overfitting. In the MLP we applied an aggressive *dropout* [Hinton et al. 2012] to avoid overfitting due to the high entropic capacity of the model. We turned off 75% random units in the *FC 1024* and *FC 4096* layers, to prevent overfitting, forcing the model to learn with less architectural capacity, about 25% of the total units.

With the purpose of analyzing the results from the proposed 4-CSCNN, we built a simplified model, replacing the CNNs by Histogram of Oriented Gradients (HOG) descriptors [Dalal and Triggs 2005] and the fully-connected layers by a simple MLP classifier. We resize each image to 128x128 pixels and applied the HOG methodology using patches of 8x8 pixels, with 8 directions and 1 cell per block. Each cardinal HOG resulted in a feature dimension of 2048. Similar to the proposed 4-CSCNN model, the features were concatenated, with the resulting dimension of 8192. Next, the MLP algorithm was trained to classify each location point into one of the 4 crime rate labels. The baseline MLP was built with 2 hidden layers, with 4 and 2 neurons respectively, using the *stochastic gradient descent* optimization algorithm and 200 epochs. We train and test the baseline model 10 times, shuffling the datasets each time, and averaged the overall accuracy to obtain a score metric for further comparison.

4. Results and Discussion

As depicted in Figures 3(a) and 3(b), high values during training opposing lower values in validation or test step indicates overfitting, which was observed after 150 epochs in our model. The results obtained during training step are presented in Figure 3(a), including the loss value for each epoch. Figure 3(b) shows validation loss values and overall accuracy calculated during validation step. Loss value is the sum of the errors obtained during the evaluation of each example, and it's expected to diminish during training and to be low during validation or test step. In our model, the loss value obtained during validation was approximately 1.

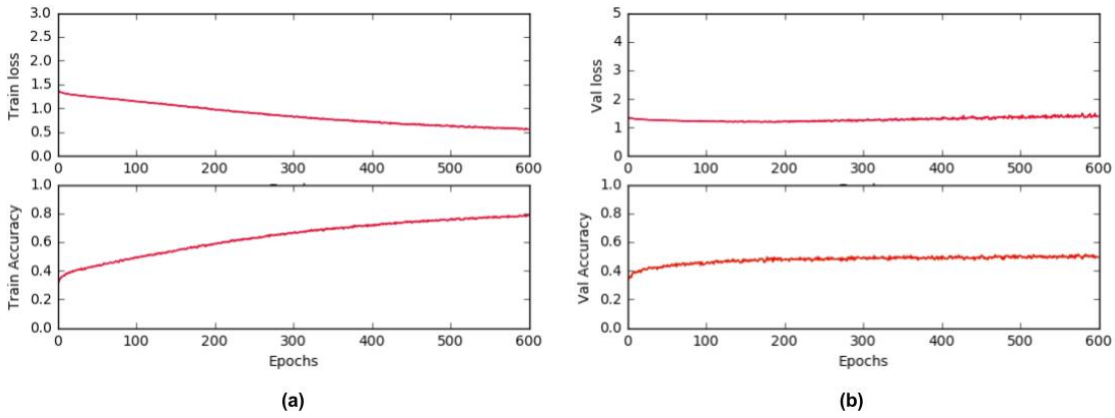

Figure 3. Loss and accuracy values during training (a) and validation (or test) (b) steps.

Table 2 shows the confusion matrix of the results obtained during validation step, at the 150 epoch, right before the model shows signs of overfitting. For each crime rate labels, 1073 location points containing 4 images were validated. From all points tested, about 34.2% of all predictions were classified as "red", or as a local with high crime rates, 28.3% as "orange", 21.6% as "green" and 15.7% of all predictions were classified as "blue", with low crime rates. The presence of fewer examples for low crime rates and the chosen distribution of the labels may explain the tendency of the proposed model to classify most images as medium or high crime rates. The accuracy per label, calculated as one-against-all using the equation $(TP + TN)/(TP + TN + FP + FN)$ is presented

in Table 3. Following the values for each one of the labels, the accuracy per label varied between 73% to 81%, and the average accuracy obtained was approximately 77%. The overall accuracy, calculated as $\sum(mainDiagonal)/population$, at the 150th epoch was 54.3%.

Table 2. Confusion matrix showing the classification between four labels indicating low to high crime rates resulted from validation step.

Predicted / Actual	Blue	Green	Orange	Red	Total
Blue	**470**	144	210	249	1073
Green	97	**512**	195	269	1073
Orange	57	141	**636**	239	1073
Red	50	134	176	**713**	1073
Total Predicted	674	931	1217	1470	4292

Following Table 3, from the total predicted examples as "blue", 69% of the time the model was correct, i.e., the precision of the model in predicting a local as "blue", and, from all the times that the model should have predicted a local as having low crime rates, it correctly predicted about 43% of the examples i.e., the recall. The values for high crime rates, or "red", were about 48% of precision and 66% of recall, due to more examples classified as "red".

Table 3. Results obtained during testing: accuracy, precision and recall for all classified labels.

	Blue	Green	Orange	Red
Population	4292	4292	4292	4292
TP: True Positive	470	512	636	713
TN: True Negative	3015	2800	2638	2462
FP: False Positive	204	419	581	757
FN: False Negative	603	561	437	360
Accuracy	0.811	0.771	0.762	0.739
Precision	0.697	0.549	0.522	0.485
Recall	0.438	0.477	0.592	0.664
Avg. Label Accuracy	0.77	(77%)		
Overall Accuracy	**0.543**	(54.3%)		

Table 4 shows the achieved results with the baseline model, compared with the results from the proposed 4-CSCNN network. CNNs need more examples than classic computer vision descriptors, that only needs few examples to achieve reasonable results. The results obtained with the simplified model demonstrate the non-trivial nature of the problem addressed, necessitating the investigation of a model that better serves the classification of the criminal rate through street-level images. Our proposed 4-CSCNN performed better considering the overall accuracy of the simple baseline-MLP model.

Table 4. Comparison between the proposed 4-CSCNN and the baseline model with HOG+MLP classifier.

Model	Overall Accuracy
Baseline-MLP	0.455
4-CSCNN	0.543

5. Conclusion

This paper presents a preliminary study on predicting crime rates from street-level images, which represent the urban environment where street crime occurs. For this, we proposed a new 4-Cardinal Siamese Convolutional Neural Network (4-CSCNN) architecture to predict urban crime rates, given a georeferenced location point. The model uses 4 images surrounding the given point, facing north, south, east and west. Each image is the input of one CNN, with pre-trained frozen weights from *AlexNet* architecture [Krizhevsky et al. 2012]. At the output of each CNN, a *Fully Connected* (FC) layer was attached, and the resulting descriptors were merged into one only descriptor, that is finally classified by a Multi-layer Perceptron (MLP) into one of the four crime rate labels.

The CNNs are responsible for learning features of the environment images which may affect crime rates. The use of 4 images surrounding a location gives more information about the environment than using a single image as input. The obtained results, 54.3% of overall accuracy, indicates that the architecture can infer a possible relation between environment features and crime rates, using only street-level images. The CNNs achieved better results when compared with the baseline model with HOG descriptors, considering the overall accuracy score. Although the result obtained for overall accuracy is less than the related works, our 4-CSCNN model can be distinguished from related works by differences in how we tackle the input image. Our model does not focus on specific attributes in the image, instead, it uses the whole scene as input, and it is still a preliminary study to investigate the use of CNNs to discover the most important features in images for the crime prediction task.

For future works, we intend to implement techniques of deep visualization in the neurons, to show what image inputs cause higher activation in units. By doing this, we search for the features that the CNNs learned to be related to low or high crime rates. This can be useful for social and law enforcement analysis of the urban environment. Also, our model has potential to be applied to different problems i.e. 4-Cardinal images of one georeferenced point can be related to different statistics and environment characteristics, e.g. city region classification as downtown and suburbs, that requires surrounding visualization.

Acknowledgements

We acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

References

Arietta, S. M. and Efros, A. A. (2014). City Forensics : Using Visual Elements to Predict Non-Visual City Attributes. *Transactions on Visualization and Computer Graphics*, 20(12):2624–2633.

Block, C. (1998). The GeoArchive: An information foundation for community policing. *Crime mapping and crime prevention*, pages 27–81.

Bowers, K. J., Johnson, S. D., and Pease, K. (2004). Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology*, 44(5):641–658.

Brantingham, P. J. and Brantingham, P. L. (1993). Environment, routine and situation: Toward a pattern theory of crime. *Advances in criminological theory*, 5:259–294.

Brantingham, P. P. and Brantingham, P. P. (1995). Criminality of place. *European Journal on Criminal Policy and Research*, 3(3):5–26.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

Caplan, J. M., Kennedy, L. W., and Miller, J. (2011). Risk terrain modeling: Brokering criminological theory and gis methods for crime forecasting. *Justice Quarterly*, 28(2):360–381.

Chainey, S., Tompson, L., and Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21:4–28.

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

CityOfChicago (2017). Chicago data portal. https://data.cityofchicago.org/.

Cohen, L. E. and Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, 44(4):588–608.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 886–893.

Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. a. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics*, 31(4):1–9.

Drawve, G., Thomas, S. A., and Walker, J. T. (2016). Bringing the physical environment back into neighborhood research: The utility of RTM for developing an aggregate neighborhood risk of crime measure. *Journal of Criminal Justice*, 44:21–29.

Dubey, A., Naik, N., Parikh, D., Raskar, R., and Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. pages 196–212.

Eck, J. E. and Weisburd, D. L. (1995). Crime Places in Crime Theory. *Crime Prevention Studies*, 4:1–33.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US. pages 1–41.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61(1):115–125.

Google (2017). Google Street View API. `https://developers.google.com/maps/documentation/streetview`. [Online; accessed 09-May-2017].

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Johansson, E., Gahlin, C., and Borg, A. (2015). Crime Hotspots: An Evaluation of the KDE Spatial Mapping Technique. In *EISIC European Intelligence and Security Informatics Conference*, pages 69–74, Manchester, UK. IEEE.

Khosla, A., An, B., and Lim, J. (2014). Looking beyond the visible scene. *2014 IEEE Conference on*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lieman-Sifry, J. (2016). Convolutional neural networks to predict location from colorado google street view images: Galvanize capstone project. https://github.com/jliemansifry/streetview/.

Lin, T.-Y., Cui, Y., Belongie, S., and Hays, J. (2015). Learning deep representations for ground-to-aerial geolocalization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5015. IEEE.

Mapzen (2017). Mapzen metro extracts. https://mapzen.com/data/metro-extracts/. [Online; accessed 09-May-2017].

Martín, Abadi, e. a. M. A. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Sherman, L. W., Gartin, P. R., and Burger, M. E. (1989). Hot spots of predatory crime : routine activitices and the criminology of place. *Criminology*, 27(June):27—-55.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

Wilson, J. Q. and L., K. G. (1982). Broken Windows. *Atlantic Monthly*, 249(3):29.

Wortley, R. and Mazerolle, L. (2008). *Environmental criminology and crime analysis: situating the theory, analytic approach and application*.

Zagoruyko, S. and Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.