Predicting Road Traffic Accidents

https://github.com/jfbrou/Accidents

Jean-Felix Brouillette

November 17, 2020

Abstract

Injuries from road traffic accidents is now the eighth leading cause of death globally, claiming about 1.35 million lives every year.[1] What if one could predict the likelihood of road traffic accidents to induce drivers to avoid hazardous road segments? In this project, we show how deep learning can be used to provide accurate high-resolution road traffic accident predictions in the form of an hourly road segment risk map.

1 Introduction

Road traffic accidents claim about 1.35 million lives every year, with 50 million more left injured or disabled. It is the leading cause of death globally for individuals between the age of 5 and 29. Despite it being an international priority, the United Nations sustainable development goal of halving road traffic accidents by 2020 has not been met. In fact, the slight decline in the death rate in recent years has not kept pace with the increasing absolute number of road accidents coming from rapid population growth and urbanization. This growing human and economic burden urges us to think about new approaches and solutions.

What if one could predict the likelihood of road traffic accidents to induce drivers to avoid hazardous road segments? One could think of introducing such predictions in modern GPS navigation applications where drivers could receive routing plan suggestions that avoid accident prone road segments. In this project, we build a deep learning model to produce such accurate high-resolution road traffic accident predictions for the city of Montreal. Our Multilayer Perceptron (MLP) model takes as inputs meteorological, spatial, temporal and geometric features of each road segment to hourly predict the binary outcome of whether or not an accident will occur on any given road segment.

[1] https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

2 Related Work

The problem of predicting road traffic accidents can be formulated in several ways. Depending on the spatial and temporal resolution of an observation, one can define the exercise as a classification or regression problem. In fact, if an observation is defined on a road segment at the hourly frequency, it is more convenient to perform a binary classification exercise in which an accident is treated as a Bernoulli random variable. In contrast, if an observation is defined on a coarser spatial unit or at a lower frequency, one can instead perform a regression exercise in which the model will predict the number of accidents that will occur within the specified area and time interval.

Depending on intended use of the model, any of the two formulations could be preferred, but for our purpose, it is more convenient to formulate the problem as a binary classification one since a navigation application would need relatively high-resolution predictions to produce optimal routing plans at any given moment. In that sense, Hébert, Guédon, Glatard and Jaumard (2019) is closest to what we are doing. In fact, they use a Balanced Random Forest (BRF) model with similar data sources to predict road traffic accidents in the city of Montreal. With this model, they achieve a recall of 85% and precision of 28%. In this project, we seek to improve upon this by instead using a deep architecture.

In contrast, most previous work on predicting road traffic accidents has been formulated as a regression problem. For instance, Chen, Song, Yamada and Shibasaki (2016) use human mobility data from mobile phone GPS records to predict traffic accident risk on a 500 by 500 meter grid in Japan, where risk is defined as the sum of the severity of accidents that occur in the area at a given hour. Using a deep architecture, they achieve a root mean-square error of unity. The main contribution of this article is to introduce human mobility data as a model feature to substitute for detailed road traffic data, which is typically harder to access. Najjar, Kaneko and Miyanaga (2017) also introduce a novel source of data to predict traffic accident risk by exclusively using satellite imagery together with a Convolutional Neural Network (CNN) to predict road traffic accidents in New York City and Denver. With different image dimensions and training strategies, they achieved a precision within the range of 74% and 78%. This is convincing evidence that visual features encoded in satellite imagery can accurately predict road traffic accidents. Finally, formulating the problem as a more standard regression exercise, Yuan, Zhou and Yang (2018) use a Convolutional Long Short-Term Memory (ConvLSTM) Neural Network to predict the daily number of accidents on 5 by 5 kilometer grid in Iowa. Their main contribution is to introduce a rather novel architecture for this task that accounts for long dependencies in both the spatial and temporal dimensions. With this, they achieve a root mean-square error lying between 0.08 and 0.14 for different configurations of their model and data.

3 Data

In this project, we will need data on (i) road traffic accidents, (ii) road segments, (iii) traffic lights and (iv) hourly local weather estimates. For (i) to (iii), the city of Montreal provides three public datasets on road traffic accidents, road segments and traffic lights. For (iv), the government of Canada provides publicly available hourly weather information measured at several weather stations in or near Montreal.

(a) Road traffic accidents

(b) Road segments

3.1 Road Traffic Accidents

This dataset provided by the city of Montreal records every single road traffic accident that occurred between 2012 and 2019 in Montreal.[2] For each accident, the dataset contains information on the date, time and location of the event as well as additional information on injuries, casualties, vehicles involved and road conditions, which will not be used. The dataset contains a total of 173,661 accidents, plotted in Figure 1a, for which the date, time and location of the event is observed.

3.2 Road Segments

The city of Montreal also publicly provides a dataset containing the line strings of every road segment defined by intersections in Montreal.[3] For each road segment, the dataset contains information on the road type and direction. This dataset contains a total of 47,567 segments, plotted in Figure 1b. From this data, we calculate additional road segment features such as segment length and sinuosity, the area of the segment's convex hull and the number of intersections with other segments.

[2] http://donnees.ville.montreal.qc.ca/dataset/collisions-routieres/resource/5a81f6c5-e3e7-4c0e-9ccf-a4ba2cab77ba
[3] http://donnees.ville.montreal.qc.ca/dataset/geobase/resource/70c1f8c7-91a0-4553-b602-89c3edb959b5

3.3 Traffic lights

The last dataset we obtain from the city of Montreal contains the geographic location of all traffic lights in the city.[4] This data is used to count the number of traffic lights within a 100 meter radius of each road segment.

3.4 Weather

This dataset provided by the government of Canada contains hourly weather information measured at different weather stations in Canada.[5] More precisely, each station records the temperature, dew point, relative humidity, wind direction and speed, visibility, atmospheric pressure and other atmospheric phenomena such as snow, fog and rain at every hour. In total, there are 12 weather stations located within a radius of 50 kilometers of Montreal reporting at an hourly frequency. Weather conditions will therefore be interpolated between weather stations to obtain estimates at each road segment.

3.5 Combination

Combining the first three datasets, we obtain total of 173,452 positive examples, in which an accident occurred. As previously mentioned, since our data spans a period of 8 years and there are 47,567 road segments in Montreal, this amounts to a total of over 3.3 billion possible negative examples. Including all of those negative examples in the training, development and test sets would lead to a severe class imbalance problem. Therefore, we instead use a sampling approach by which we randomly sample accident records from the set of positive examples, randomly alter the associated date and time, and if those resulting examples are not already within the set of positive examples, we include them in our dataset. Repeating this procedure until there are enough negative examples (i.e. as much as the number of positive examples) circumvents the class imbalance problem. Combining the resulting dataset with weather information, we obtain a total of 157 normalized features with 841,861 examples, which we divide in the training, development and test sets with fractions 90%, 5% and 5%, respectively. Besides what is mentioned above, the features include binary variables for the month, week, day, weekday and hour of the examples as well as the elevation, zenith and azimuth of the sun given their location, date and time.

4 Methods

In this project we build a Multilayer Perceptron (MLP) with six hidden layers and 53,217 trainable parameters. Since we formulate the problem as a binary classification task, we use the binary cross-entropy loss function with a sigmoid activation function for the output layer. Each hidden

[4] http://donnees.ville.montreal.qc.ca/dataset/feux-tous/resource/b5153f25-37d0-4e5f-a367-2d02b3f1d826
[5] https://climate.weather.gc.ca/index_e.html

layer is activated with the ReLU function after which we apply batch-normalization. More precisely, here is a description of each layer of the model:

1. Input layer with 157 features.
2. Dense layer of 256 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization $\rightarrow$ dropout with probability 0.5.
3. Dense layer of 256 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization $\rightarrow$ dropout with probability 0.4.
4. Dense layer of 128 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization $\rightarrow$ dropout with probability 0.3.
5. Dense layer of 128 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization $\rightarrow$ dropout with probability 0.2.
6. Dense layer of 64 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization $\rightarrow$ dropout with probability 0.1.
7. Dense layer of 32 units $\rightarrow$ ReLU $\rightarrow$ batch-normalization.
8. Dense layer of 1 unit $\rightarrow$ sigmoid.

5 Discussion

To train the above model, we use mini-batch gradient descent with mini-batches of size 128 and the ADAM optimizer with a learning rate of 1×10^{-3}. After 100 training epochs, we obtain the confusion matrix in Table 1. This yields a recall of 68.9% and an area under the ROC and PR curve of 63.7% and 61.9%, which is not quite as high as what is achieved in Hébert et al. (2019); 85%, 92% and 69%, respectively.

Table 1: Confusion matrix

True/Predicted	Negative	Positive
Negative	24.8%	24.3%
Positive	16.6%	34.2%

6 Conclusion

The performance of the above model is unfortunately not satisfactory. I believe there are at least three reasons why. First of all, the random sampling solution to the severe class imbalance problem might no suffice. In fact, since correctly predicting accidents is so critical to this task, it could be that one should use relatively more negative examples, but instead modify the binary cross-entropy loss function in order to penalize deviations from positive examples proportionally more. This solution will be explored in future iterations of this project. Second, weather is measured with a substantial amount of noise in our data. In fact, we are linearly interpolating weather conditions at each point of the city of Montreal from only 12 weather

stations within a radius of 50 kilometers. It should be no surprise that weather conditions are not linear in space, and instead using data generated from a meteorological model could increase the performance of our model if weather is a critical predictor of road traffic accidents. Finally, using alternative features such as satellite images or road traffic data could be useful in predicting traffic accidents. In fact, satellite images and traffic data could encode a vast amount of information on how hazardous roads can be.

References

Chen, Quanjun, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," in "Thirtieth AAAI Conference on Artificial Intelligence" 2016.

Hébert, Antoine, Timothée Guédon, Tristan Glatard, and Brigitte Jaumard, "High-Resolution Road Vehicle Collision Prediction for the City of Montreal," in "2019 IEEE International Conference on Big Data" IEEE 2019, pp. 1804–1813.

Najjar, Alameen, Shun'ichi Kaneko, and Yoshikazu Miyanaga, "Combining satellite imagery and open data to map road safety," in "Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence" 2017, pp. 4524–4530.

Yuan, Zhuoning, 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" 2018, pp. 984–992.