Stanford University

A Novel Approach for Predicting and Understanding Road Danger in the Developing World:

Deep Video-Classification of Roads in Nairobi, Kenya

Alexandr Lenk, Matias Cersosimo, Negin Raoof[1]

Abstract

With a road traffic death rate of 27.8 per 100,000 inhabitants in 2016, Kenya has nearly twice as much road fatalities as the world average. Hence, understanding the factors determining road danger is key and we are the first ones to attempt at achieving this goal by constructing a deep learning model entirely based on videos of several road segments from Nairobi. The best-performing model is a pre-trained Res-Net3D with shorter video clips which results in a test set accuracy of 50%. Comparing the results with the rest of the proposed models, we are able to infer that road danger is not only a function of quality of the road, but also of the density of road and pedestrian activity within a given timeframe. While similar models for video classification of daily activities reach an accuracy of 70%, we believe that given the increased complexity of our classification task (road danger), we fare rather well as a first pass. Click here to access the GitHub repo of this project.

1. Motivation and Related Work

Providing traffic safety and lowering the rate of road accidents in Nairobi, Kenya is a major concern. With a road traffic death rate of 27.8 per 100,000 inhabitants in 2016, Kenya has nearly twice as much road fatalities as the world average (WHO, ongoing). Collaborating with the World Bank, we are the first ones to construct a deep learning model entirely based on videos of several road segments from Nairobi. The model allows us to analyze different road conditions and predict danger level of roads.There have been previous studies which used deep learning to evaluate the risk of traffic accidents such as Hébert, Antoine, et al. (2019), Chen et al. (2016) and Yuan et al. (2018). These studies have mainly focused on training models based on features such as weather, human mobility, road conditions and satellite images. However, we aim to develop models which process raw video data capturing traffic patterns in order to classify road danger level. Karpathy et. al. (2014) write the seminal paper on video classification using 3D CNN-based models. Later papers such as Abu-el-Haija et. al. (2016) and Diba et. al. (2017) come up with deeper and more advanced architectures and incorporate transfer learning in order to improve performance. The current state of the art model is given by Carreira & Zisserman (2017) who reach 80.9% accuracy on HMDB-51 dataset and 98.0 % on UCF-101 dataset. Most of the papers use daily activity datasets, which is arguably a much easier classification task that the road danger evaluation that we aim to conduct.

2. Data

The data we obtained from the World Bank consists of the following datasets:

- A geojson file containing 912 unique entries, each entry corresponding to a 100-meter long road segment.
- A geojson file with 1428 crash hotspots linked to the number of annual crashes between 2012-2018 along with the number of fatalities occurring at each hotspot.
- A folder with 852 videos of length varying between 2 and 8 minutes, all videos taken in the morning and capturing traffic and pedestrian activity on different road segments (the videos do not capture crashes, but rather road conditions that are typical of that particular location and time of day).

Our first task consisted in finding the number of crashes associated to each road segment. We achieve this by matching road segments to crash

[1]The three team members have contributed equally to the development of this project.

hotspots using geographical coordinates in the following way. We calculate the distance between each road segment and hotspot and match segment i to hotspot j as long as the distance between i and j is less than 130m.[2] As a result, it happens that some road segments could be matched to multiple hotspots: 2 roads have no match, 288 roads have a unique match, 424 roads have two matches, 165 roads have three matches, 21 roads have four matches, 11 roads have five matches and 1 road has six matches.

Figure 1: SEGMENT (BLUE) MATCHED WITH 1 HOTSPOT (RED)

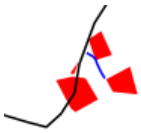

Figure 2: SEGMENT (BLUE) MATCHED WITH 4 HOTSPOTS (RED)

For road segments matched to multiple hotspots, the final number of crashes associated to a road was calculated as the average of the number of crashes occurring in all hotspots matched to that road where each matched hotspot is weighed by the inverse of its distance from the road (i.e., hotspots closer to a road get a larger weight).[3] We finally note that only fewer than 200 out of the 1428 hotspots were used as matches for the roads. We notice, in fact, that those hotspots report more crashes than the average number of crashes across the 1428 hotspots, which means that our task should be interpreted as *analyzing road danger intensity in locations that are relatively more dangerous at the outset*. This observation is in line with the World Bank's strategy to take videos of the most dangerous hotspots.

Next, we create labels for those roads in terms of danger, which we assume to be an increasing function of the number of crashes.[4] In theory, it is is possible to use the number of crashes directly as our output variable and perform a regression task. However, the crash data is very likely to be subject to measurement error, which would make predictions noisy had we decided to go ahead with a continuous label. Therefore, we were more confident adopting an ordinal categorical approach in which we classify the roads into 4 categories using a k-means algorithm. The number of clusters has been chosen by playing around such that we try to keep k relatively small, but at the same time capture a decent level of danger heterogeneity. We get the following distribution of road danger:

Table 1: Distribution of Road Danger

Danger level	# of Roads	Mean # of Crashes within a Category	Percentile of the Corresponding Mean
1	344	6.40	18 %
2	367	14.94	59 %
3	130	26.30	86 %
4	71	44.27	96 %

3. Training and Validation Datasets

Our videos are processed as follows. First, we crop all videos to a sub-clip of a pre-determined length (see below for details), starting from a random point in time. Table 1 shows that we suffer from class imbalance which could negatively affect our training. We deal with this issue using two distinct approaches. The first approach consists in going ahead with the class imbalance issue and reporting a variety of evaluation metrics rather than just aggregate accuracy which we know can be misleading in this context. The second approach consists in directly dealing with class imbalance by including more observations for categories 3 and 4. In particular, rather than just sampling copies of those videos, we instead decide to crop videos for categories 1 and 2 and keep longer videos for categories

[2] This was the rule that the World Bank recommended since the geometry of each hotspot was constructed based on reporting approximate car accident locations which could have occurred in a larger radius.

[3] We consider only the sum of crashes between 2015 and 2018 since records have not been consistent in earlier years.

[4] We initially considered the severity of accidents as well; however, when normalizing by number of crashes, the severity in different hotspots was similar, so we disregard that feature. As a robustness check, we performed a second classification where we did include severity, but this did not change the label classification substantially.

3 and 4. In particular, we set the length to 30 seconds for categories 1 and 2, 75 seconds for category 3 and 150 seconds for category 4. Notice that categories 3 and 4 are respectively 2.5 times and 5 times longer than category 1 and 2 videos, which corrects for the fact that the number of roads in category 3 and 4 is respectively 2.5 and 5 times lower than the number of roads in categories 1 and 2. As is standard in the literature of video classification, we then cut each cropped video into a number of shorter video clips and use those as our final inputs to the algorithm. Notice that with the same video clip length across categories, the number of video clips across categories becomes comparable (since category 3 and 4 videos are longer) and hence we obtain a balanced training set even though the number of actual full-length videos is still imbalanced, but this does no longer affect training itself. For the first approach that does not deal with class imbalance directly, we crop all videos to a length of 150 seconds, which we further cut into video clips of same length for each category. Clearly, this will keep the class imbalance in terms of number of video clips per category and hence our training set remains imbalanced.

For both approaches, we divide our video clip data set into train, validation and test using a proportion of 80:10:10. We resize video frames to 112 pixels height x 112 pixels width, and we normalized the pixels to floating point values between 0 and 1, using calculated mean and standard deviation over the training samples. Hence, each video clip has a shape of $(C, T, H, W) = (3, 15, 112, 112)$.

4. Method

Following a Caviar-like approach, we choose to train 3 main types of deep neural networks:

- A ConvNet-3D which we train from scratch. This model has two 3D convolution layers (Conv1: 32 filters 5 x 5 x 5, Conv2: 64 filters 3 x 3 x 3) each followed by a 3D BatchNorm, ReLU and Dropout ($p = 0.2$) layer. The convolution block is followed by a 3D MaxPool (2 x 2 x 2) and two fully connected (FC1: 256, FC2: 128) + ReLU layers, and a 4-class Softmax. We use Cross-Entropy loss function to train this model optimized using Adam Optimizer.
- A pre-trained 18-layer ResNet-3D (Tran et al. (2017)) to which we add two hidden fully connected (FC1: 256, FC2: 128) + ReLU + Dropout (p=0.2) layers, and a 4-class Softmax. The ResNet-3D 18 model is pre-trained on Kinetics-400 dataset. During training, we freeze the weights of the ResNet 18 block, and train the two additional fully connected layers. We use Cross-Entropy loss function to train this model optimized using Adam Optimizer as well.
- A ConvNet-2D model based on ResNet-18 where we just sample random single frames from each video-clip. Again, we have two fully connected (FC1: 256, FC2: 128) + ReLU + Dropout ($p = 0.2$) layers, and a 4-class Softmax added to the pre-trained model and we train these layers only, keeping the ResNet weights fixed. The same Cross-Entropy loss function optimized using Adam Optimizer is used for training.

Given time and financial constraints, we decided to tune hyperparameters that were common (mostly) to the three models. Thus, our main tuning parameter becomes the length of the video clips. We fix the number of frames per video clip to 15 and we set frames per second to either 1 or 10, which results in 15 and 1.5 seconds video-clips respectively.[5] A secondary tuning parameter is epoch length. The best epoch length for the pre-trained ResNet-3D models was 6 (validation accuracy started decreasing after the 6th epoch), while we did not observe substantial improvement after the 9th epoch for the rest of the model such that we trained those for 10 epochs. We set the learning rate to 0.0001, and the batch size to 40 for 1FPS clips and to 200 for 10FPS clips.

[5] For the Conv-2D model, since we are just sampling single frames, we know that this does not affect the length of video clip, hence the two models are essentially the same. Rather the difference between the two models is that we will be sampling more frames overall for the 10FPS models. Indeed, our results show that there are basically no differences in performance between the ConvNet-2D 1FPS and 10FPS, which confirms our prior.

The results with the imbalanced data set are summarized in the tables below where, in addition to aggregate accuracy, we also report category-specific accuracy to check whether the model performs well also in the classes that are scarcer:[6]

Table 2: Accuracy Results (in %): Training Set

	Training Accuracy				
Model	**1**	**2**	**3**	**4**	**All**
Conv3D, 1FPS	72.54	68.18	0	0	57.26
Conv3D, 10FPS	89.21	90.91	36.66	14.28	78.63
ResNet3D, 1FPS	100	2.27	0	0	44.44
ResNet3D, 10FPS	90.19	88.63	60	7.14	80.76
Conv2D, 1FPS	89.21	11.36	0	0	43.16
Conv2D, 10FPS	98.03	4.54	0	0	44.44

Table 3: Accuracy Results (in %): Validation Set

	Validation Accuracy				
Model	**1**	**2**	**3**	**4**	**All**
Conv3D, 1FPS	20	100	0	0	53.84
Conv3D, 10FPS	64.70	60.60	0	0	49.41
ResNet3D, 1FPS	100	0	0	0	38.46
ResNet3D, 10FPS	55.88	51.51	18.18	0	44.70
Conv2D, 1FPS	100	16.66	0	0	46.15
Conv2D, 10FPS	100	15.15	0	0	45.88

Analyzing the results, we see that ConvNet-3D, 10FPS and ResNet-3D, 10FPS have the lowest bias (lowest training set error in terms of aggregate accuracy) whereas ConvNet-3D, 10FPS has the highest aggregate accuracy on the validation set (lowest variance). However, looking more carefully at category-specific accuracies, ResNet-3D, 10FPS is actually the only model that is able to correctly classify a positive fraction of category 3 videos despite the fact that it does not perform best in terms of aggregate accuracy. To obtain additional evidence, we compare the true distribution of labels to the predicted distribution of labels on the validation set for each of the 6 models.

Table 4: Category Distribution (Validation Set): True vs Predicted

	Category Distribution			
Model	**1**	**2**	**3**	**4**
True Distribution	38.46	46.15	7.69	7.69
Conv3D, 1FPS	7.69	92.31	0	0
Conv3D, 10FPS	49.41	48.23	1.17	0
ResNet3D, 1FPS	100	0	0	0
ResNet3D, 10FPS	47.05	47.05	5.88	0
Conv2D, 1FPS	88.46	11.53	0	0
Conv2D, 10FPS	89.41	10.58	0	0

We observe that the predicted distributions[7] given by ResNet-3D, 10FPS and ConvNet-3D, 10FPS are closest to the true distributions with ResNet3D, 10FPS being somewhat closer. In conjunction with the aggregate accuracy and class-specific accuracy results, it seems that ResNet-3D, 10FPS is the best model among the six with our first approach. The other models are incapable of recognizing categories 3 and 4 at all, which while certainly related to the class imbalance issue, do not dispute the fact that ResNet-3D, 10FPS performs best as it can recognize those higher categories despite the class imbalance.

Additional inspection of the results shows that the temporal dimension of the video does matter: the ConvNet-3D and ResNet-3D models do perform better overall compared to the ConvNet-2D model. This suggests that road danger is not just given by background characteristics of the video that are fixed (e.g., road and pavement quality, presence of zebra crossings an lights, etc..), but also depend on features that change over time (e.g., circulating traffic and flow of pedestrians). Furthermore, the 10FPS videos, which are the shorter video

[6] Accuracy can be calculated using two approaches. The first one averages predicted probabilities across video clips within a video and sets the predicted label to the category that has the highest average predicted probability across the four categories. The second approach predicts labels per video clip and sums predicted labels across video clips within a video. The final predicted video label is set as the one that has the highest number of video clips with that label (similar to majority voting). Since the average approach and majority approach coincide in more than 90% of the cases, we present results with the average approach to avoid 50-50 problem which occurs sometimes with the majority approach.

[7] One may notice that the true distribution of labels within our dataset does not perfectly coincide with the true distribution from Table 1. This is because Table 1 uses all road segments present in the geospatial file while for training, we only use the road segments for which we have videos. It is reassuring however that the distribution based on the available videos is similar to the distribution based on the full set of roads.

clips, give better performance. Notice that this result was not clear ex-ante and that is another reason why we chose clip length as our main tuning parameter. Longer videos could have higher accuracy as they could contain more and better connected information about road activity while shorter videos might contain more disconnected information which might reduce performance. Indeed, we observe the opposite in our results which we interpret as the additional information contained in the longer video possibly being "lost on the way", i.e., this information getting averaged out across the network layers which mechanically reduces its richness. On the other hand, while more disconnected information comes from the shorter clips, less averaging out is occurring throughout the network such that more of the video clip information (albeit of lower content and noisier) is maintained.

Unfortunately due to time and financial constraints, we were unable to run our second approach which deals with class imbalance directly with all six models. Instead, we run the second approach only with the best model from the first approach, the ResNet-3D, 10FPS. The results are summarized below:

Table 5: Second Approach Accuracy Results (in %)

	Accuracy				
Set	**1**	**2**	**3**	**4**	**All**
Training	64.22	48.38	66.27	82.22	59.41
Validation + Test	52.94	48.48	36.36	28.57	47.05

Table 6: Label Distribution (Validation Set): True vs Predicted

	Category Distribution			
Model	**1**	**2**	**3**	**4**
True Distribution	40	38.82	12.94	8.23
Second Approach Predicted Distribution	27.05	36.47	20	16.47

Notice that we combine validation and test sets since we are training a single model hence we can evaluate the performance directly on all non-training sets. Notice that thanks to solving the class imbalance issues, the model is much better at recognizing category 3 and category 4 roads. Aggregate accuracy on the validation is now higher for this model compared to the first case and comes mainly from improving class-specific accuracy for categories 3 and 4. We observe, however, that bias is also higher. The model is worse at predicting categories 1 and 2 labels on the training set. To put everything in perspective, however, state-of-the-art video classification algorithms that we clearly could not have implemented due to time and financial constraints reach an accuracy of 80%-98% depending on dataset used. Simpler video classification models reach an accuracy of around 70%. However, note that these are videos classifying different activities (e.g., eating, swimming, dancing, etc.) rather than road danger, which is a very different and arguably more complex activity. Indeed, we tried to ask World Bank traffic specialists to classify a subset of videos manually in order to get an estimate of human error for this task, but unfortunately we were not able to get this estimate in time. Looking at previous papers that do road danger classification based on satellite images, those reach a top accuracy of about 73%, but have much larger datasets. Hence, overall our 50% accuracy on the validation set looks like an encouraging starting point given all constraints we were facing.

5. Conclusion and Future Work

We recognize that our work so far does not give any unambiguous answer on which model works best in the context of road danger classification. However, we believe that we can form reasonable priors on what kind of models might improve performance:

- Deep 3D pre-trained models in which some of the earlier 3D layers are re-trained as well to account for the novelty of classification task.
- For longer video-clips, possibly come up with a structure that reduces the loss of information due to averaging and pooling across the network before reaching the fully connected layers.
- Explore the ordinality of our classes. During training a true label belonging to category 1 should have higher mass of assigned predicted probability to categories 1 and 2 rather than 3 and 4 since 3 and 4 are increasingly more dangerous than 2.

6. References

- World Health Organization:

  apps.who.int/gho/data/node.main.A997

- Video Classification Blog Post:

  https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5

- Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675. 2016.

- Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

- Chen, Quanjun, et al. "Learning deep representation from big and heterogeneous data for traffic accident inference." Thirtieth AAAI Conference on Artificial Intelligence. 2016.

- Diba A, Fayyaz M, Sharma V, Karami AH, Arzani MM, Yousefzadeh R, Van Gool L. "Temporal 3d convnets: New architecture and transfer learning for video classification." arXiv preprint arXiv:1711.08200. 2017.

- Hébert, Antoine, et al. "High-Resolution Road Vehicle Collision Prediction for the City of Montreal." arXiv preprint arXiv:1905.08770 (2019).

- Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. "Large-scale video classification with convolutional neural networks." InProceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014. pp. 1725-1732.

- Lin, Lei, Qian Wang, and Adel W. Sadek. "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction." Transportation Research Part C: Emerging Technologies 55 (2015): 444-459.

- Najjar, Alameen, Shun'ichi Kaneko, and Yoshikazu Miyanaga. "Combining satellite imagery and open data to map road safety." Thirty-First AAAI Conference on Artificial Intelligence. 2017.

- Theofilatos, Athanasios. "Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials." Journal of safety research 61 (2017): 9-21.

- Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.

- Yuan, Zhuoning, Xun Zhou, and Tianbao Yang. "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

- PyTorch Libraries.