

An Audio-Based Vehicle Classifier Using Convolutional Neural Network

Ekim Bakirci¹

Aecom

Sunley House, 4 Bedford Park, Croydon London UK CR0 2AP

Haydar Aygun²

London South Bank University

103 Borough Road London UK SE1 0AA

1. INTRODUCTION

Different calculation methods are being used for road traffic noise calculation purposes. For the road noise maps, the Calculation of Road Traffic Noise (CRTN) method is preferred in the UK, on the other hand; in EU countries common noise assessment methods (CNOSSOS) are the officially accepted method. These methods may differ in terms of noise metrics, meteorological concerns, and vehicle classification, nonetheless; regardless of the method chosen, source levels are always dependent on the traffic load, i.e., quantity and speed of vehicles.

Piezoelectric cable traffic census systems and traffic camera systems are the most frequent methods used to collect traffic data. However, there are limitations about the applicability of these systems such as equipping all the roads with such devices.

The main objective of this work is to generate a traffic vehicle classifier that can count traffic load in real-life conditions for single or two-lane roads by using raw audio signals.

2. METHOD

Deep neural network-based models are efficient classifiers in handling complicated classification tasks. Deep learning has the potential to have a significant influence on a variety of industries, such as speech recognition, precision medicine, cancer diagnosis, self-driving cars, etc. (Hurwitz and Kirsch, 2018) (Shrestha and Mahmood, 2019). The convolutional neural network (CNN) is one of the most widely used deep learning architectures, and it may overcome temporal and spectral constraints. Therefore, deep learning methods and CNN algorithms were selected as a method in audio-based vehicle classification.

Convolution is a mathematical operation on two functions that results in a third function that indicates how the form of one is affected by the other. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers (Goodfellow et al., 2018). Convolutional neural networks are widely used to analyze visual images.

Algorithms of speech recognition have been used in sound classification problems widely. In addition, the Mel spectrogram method for feature selection has become a common practice (Medhat et al., 2020; Salamon et al., Nov 3, 2014; Su et al., 2019).

2.1. Data Sets Preparation

¹ ekim.bakirci@gmail.com

² aygunh@lsbu.ac.uk

As the great majority of the research on CNN's relied on image-based sources and databases, a traffic audio data collection is required for vehicle classification. For the utilization of a dataset, data collection from the relevant type of environment is essential. The accuracy of the network is strongly dependent on the dataset used for training the algorithm. The more closely the training dataset resembles real-world data, the higher the accuracy would be achieved. To maintain large dataset audio file extracted from YouTube videos and self-recording used for database preparation. The pass-by events are detected by searching the peaks of audio signals. For clear identification, the minimum duration between two pass-by occurrences and the maximum length of the vehicle pass by are considered carefully. This process depends on observations, and it is influenced by the speed and density of the road traffic being studied. For a highway, 1second minimum peak to peak distance and 2 seconds of peak width are considered. In dataset preparation, the aim is to gather distinctive and representative audio events to represent the class. These signals are used in training the algorithm and should represent the real-life experience. The dataset includes six different classes which are slow passing cars ($v < 70$ km/h), fast passing cars ($v > 70$ km/h), bus, slow trucks ($v < 50$ km/h), fast trucks ($v > 50$ km/h), where v is the speed of vehicles and random audio samples are used for the background class. Nearly 3500 audio samples are used in training the algorithm.

2.2. Feature Extraction

Mel spectrogram images are used in the extraction of features from audio recordings. This method was tested by Boddapati et al. and Medhat et al. and was successful with excellent accuracy (Boddapati et al., 2017; Medhat et al., 2020). Another feature parameter that must be determined is the segment length, which specifies the time interval of the Mel-spectrogram image. Rather than examining lengthy audio recordings, meaningful sound events must be examined. This reduces computing time while it eliminates irrelevant background data. The primary goal of this research is to assess traffic density within defined classifications. As a result, following peak detection, a segment representing the pass by the length of each type of vehicle must be considered. (See Figure 1.) In real-world situations, assigning such a parameter concerning average road speed is thought to enhance the efficiency of the code prepared. The optimum length of sound occurrences can be predicted using statistics generated from training data or by making broad assumptions about the target sounds (Mesaros et al., 2021). In this study, a 4-second segment length was selected due to the convenience of the prepared dataset and the observed pass-by durations.

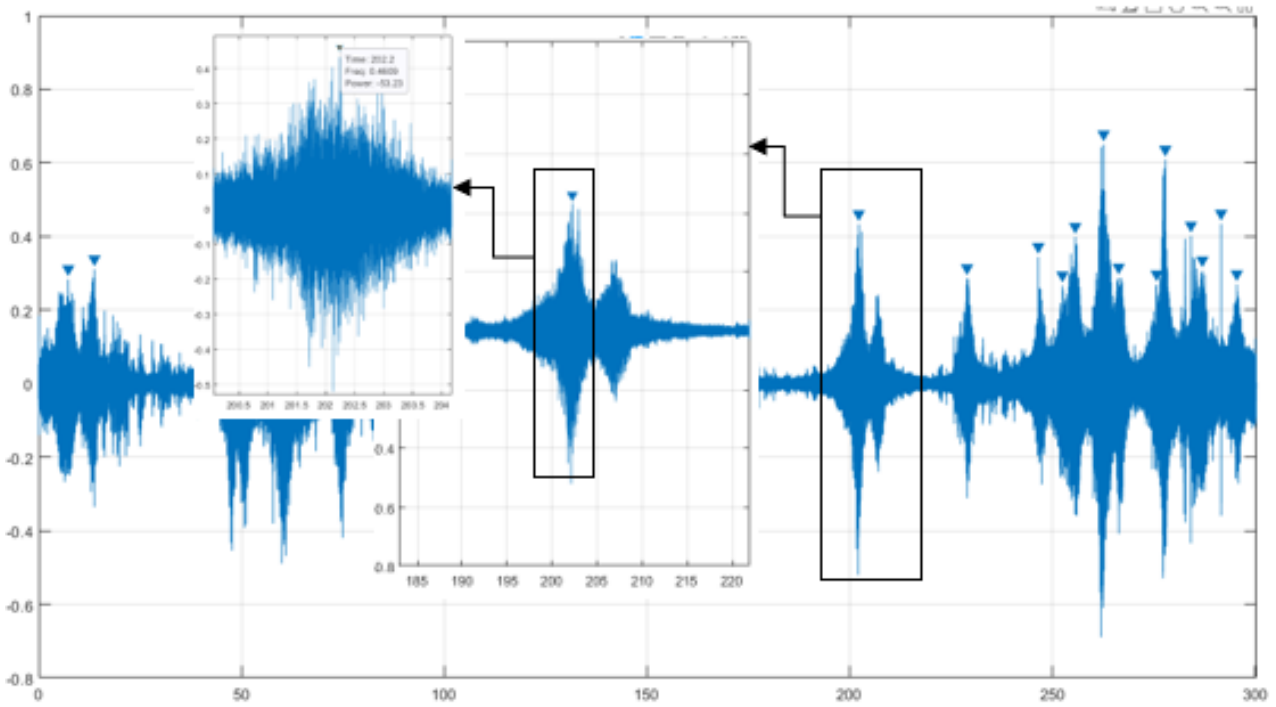
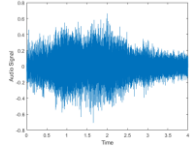
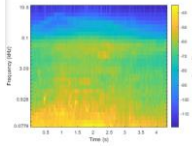
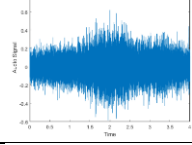
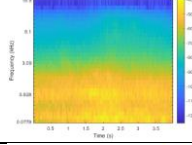
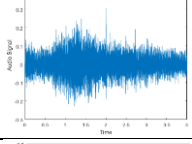
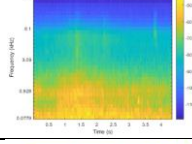
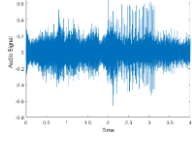
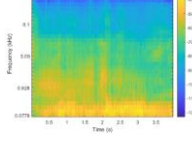
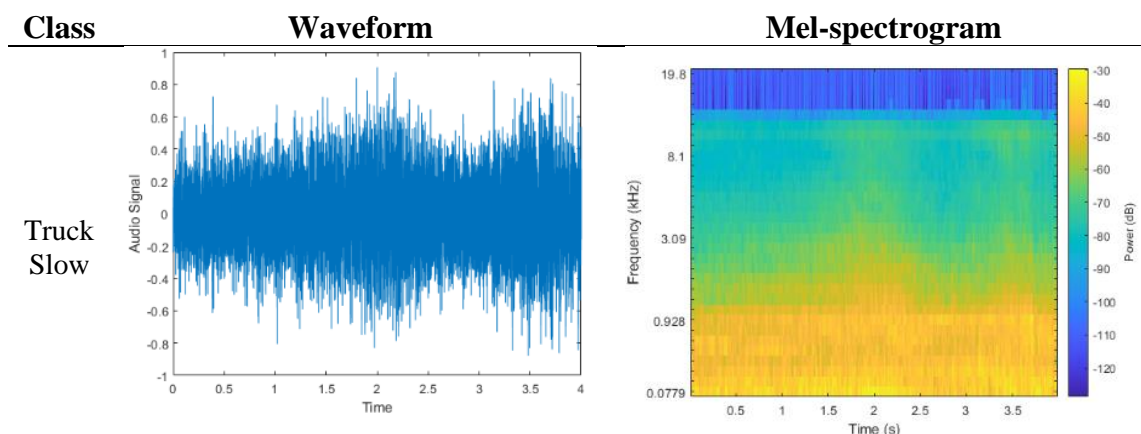


Figure 1: Segment Lengths

The Mel-spectrogram, in contrast to the STFT, is based on a non-linear frequency scale inspired by human auditory perception, and so provides a more compact spectrum representation of sounds (Abeßer, 2020). The Mel-scale is more discriminative at lower frequencies and less discriminative at higher frequencies to reflect the non-linear human ear perception of sound.

Table 1: Samples of Waveforms and Mel-spectrogram

| Class | Waveform | Mel-spectrogram |
|------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Bus |  |  |
| Car Fast |  |  |
| Car Slow |  |  |
| Truck Fast |  |  |



2.3. Audio Classification Networks

Four different single branch and feed-forward structured convolutional neural network structures are tried for the audio classification process. The networks consist of convolutional layers pool-max layers and activation functions batch normalization and *ReLU*. Number of convolutional layers, filter size and number of filters are altered to measure the performance in classification.

3. RESULTS

3.1. Training and Evaluation

The parameters of the classification networks, their training and accuracies are presented in Table 2.

Table 2: Training Accuracies of Networks

| Networks | Training Accuracy | Validation Accuracy | Layer | Filter Size | Number of Filters | Precision | F1 score |
|----------|-------------------|---------------------|-------|------------------------------|-------------------|-----------|----------|
| 1 | 95.14% | 95.98% | 6 | 3×3 | 16 to 64 | 96.77% | 96.78% |
| 2 | 96.57% | 97.41% | 6 | 3×3 | 16 to 256 | 98.45% | 98.01% |
| 3 | 96.29% | 96.55% | 4 | 3×3 | 16 to 128 | 97.78% | 97.28% |
| 4 | 96.00% | 96.55% | 4 | 7×7 to 3×3 | 16 to 128 | 97.49% | 97.24% |

The minor concern that stands up, when the confusion matrix of the first two networks is evaluated, was the mixing of fast cars and fast trucks. 11.1% of fast trucks are expected to be fast cars in the first network. Similarly, 12.7% of the fast trucks miss classified as fast cars by the second network. On the other hand, the total precision performance of the second network is 1.68% better than the first one.

There exists a relatively small decrease in total precision (0.67 percent) when the second and third networks are compared, however, no difference is observed for the classification of fast trucks and fast cars. Although the estimation error of fast trucks is decreased from 12.7 to 9.5 percent, the misclassification error for fast automobiles increased by 2.2 percent.

The major stumbling block was classifying fast cars and fast trucks in all networks. Trucks have a longer pass-by duration than automobiles in real life. Thus, one of the unique characteristics of these audio recordings can be considered as the pass-by duration of the vehicles. Even though a set of four-second-long segments were utilized in training, certain mismatches between fast passing trucks and vehicles are expected. Also, the most distinctive classes are slow cars, slow trucks, and bus pass-by.

All the networks have relatively similar scores. As predicted, the second network -which has six convolutional layers and up to 256 filters- is the most powerful network among the others. It should be underlined that the outcomes are obtained using a prepared dataset. On the other hand, Network 2 and Network 4 have closer outcomes while Network 4 has a simpler topology.

3.2. Training and Evaluation

The algorithms were tested using road traffic videos from YouTube and self-recorded audios. The results of this type of data-driven application are very reliant on the training data. As a result, the diversity, quantity, and quality of the training data are critical.

Table 3: Test Results

| Rec. 1 -Urban light Traffic | Car S. | Bus | Total | Class Estimation Error | | Total Error | | |
|-----------------------------|------------|-----------|-----------|------------------------|------------|-------------|----------|------|
| | | | | Car S. | Bus | | | |
| Actual Quantity | 26 | 1 | 27 | | | | | |
| N1 Prediction | 24 | 2 | 26 | 7.7 | 100.0 | 3.7 | | |
| N2 Prediction | 26 | 0 | 26 | 0.0 | 100.0 | 3.7 | | |
| N3 Prediction | 26 | 0 | 26 | 0.0 | 100.0 | 3.7 | | |
| N4 Prediction | 23 | 3 | 26 | 11.5 | 200.0 | 3.7 | | |
| Rec. 2 -Urban light Traffic | Car S. | Bus | Total | Car S. | Bus | | | |
| Actual Quantity | 29 | 1 | 30 | | | | | |
| N1 Prediction | 28 | 2 | 30 | 3.4 | 100 | 0.0 | | |
| N2 Prediction | 29 | 1 | 30 | 0.0 | 0.0 | 0.0 | | |
| N3 Prediction | 30 | 0 | 30 | 3.4 | 100 | 0.0 | | |
| N4 Prediction | 29 | 1 | 30 | 0.0 | 0.0 | 0.0 | | |
| Video 1 - Highway | Car F. | Truck F. | Bus | BG | Total | Car F. | Truck F. | |
| Actual Quantity | 53 | 41 | 0 | 0 | 94 | | | |
| N1 Prediction | 2 | 84 | 0 | 0 | 86 | 96.2 | 104.9 | 8.5 |
| N2 Prediction | 46 | 38 | 4 | 0 | 88 | 13.2 | 7.3 | 6.4 |
| N3 Prediction | 0 | 85 | 2 | 2 | 89 | 100.0 | 107.3 | 7.4 |
| N4 Prediction | 0 | 86 | 2 | 0 | 88 | 100.0 | 109.8 | 6.4 |
| Video 2 - Highway | Car F. | Truck F. | Car S. | BG | Total | Car F. | Truck F. | |
| Actual Quantity | 144 | 3 | 0 | 0 | 146 | | | |
| N1 Prediction | 2 | 1 | 135 | 0 | 138 | 98.6 | 50.0 | 5.5 |
| N2 Prediction | 18 | 1 | 119 | 1 | 139 | 87.5 | 50.0 | 4.8 |
| N3 Prediction | 118 | 0 | 20 | 0 | 138 | 18.1 | 100.0 | 5.5 |
| N4 Prediction | 38 | | 75 | 3 | 138 | 73.6 | 100.0 | 20.5 |

4. CONCLUSIONS

Four convolutional networks were, trained, evaluated, and tested with the same dataset. Although variations were proposed to networks, the training accuracies were close to each other and changing in between 97.41% to 95.98. The precision achieved in these tests are satisfactory. Especially for the slow passing cars, two of the networks made estimations exact. On the other hand, the same precision could not achieve in bus classification. When reason behind this error examines, it is found that the driving characteristics like acceleration and stops is the recordings cause multi peaks and this situation explains double or triple counts. The proposed technique is effective under fluent traffic circumstances. Moreover, each vehicle simply should have only one peak during pass-by to avoid double counting. The duration of the audio segments in the dataset can also be varied. This might help in decreasing misclassifications

When the number of layer units in the networks is considered, the precision of feed-forward structures increases as the networks become deeper (more layers). However, training time gets longer similarly. Sophisticated network architectures with larger layers and a greater number of narrow filters perform better. However, this also involves identifying a good enough challenge to achieve a balance between feasibility efficiency and effectiveness. With the dataset occupied and real-life recordings, the utilized networks continue to perform in line with the purpose of the study.

Since deep learning mostly relies on data-driven techniques, the dataset is the most important part of the classification problem. The audio slices used in training should comply with the purpose. The algorithms mostly succeeded at similar locations where the data was gathered for the data set. As the quantity and variety of pass-by events in the dataset increases, the precision of the algorithm would improve. The used dataset does not contain an audio sample representing the pass-by of a muscle car; most probably it would be misclassified as a truck. Similarly, motorcycles, scooters, vans, minibuses, delivery trucks were the other missing vehicle types in the dataset.

5. ACKNOWLEDGEMENTS

This work was supported by the London South Bank University and LSBU Acoustics research group. The authors wish to extend their gratitude to Frekans Akustik for helping to collect data.

6. REFERENCES

1. Abeßer, J. (2020) A review of deep learning based methods for acoustic scene classification, *Applied Sciences*, 10 (6), pp. 2020. DOI: 10.3390/app10062020.
2. Boddapati, V., Petef, A., Rasmusson, J. and Lundberg, L. (2017) Classifying environmental sounds using image recognition networks, *Procedia Computer Science*, 112 , pp. 2048-2056. DOI: 10.1016/j.procs.2017.08.250.
3. Goodfellow, I., Bengio, Y., Courville, A. and Lenz, G. (2018) *Deep Learning*. 1. Auflage ed. Frechen: mitp. Available from: http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=030119978&sequence=000002&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.
4. Hurwitz, J. and Kirsch, D. (2018) *Machine Learning For Dummies, IBM Limited Edition*. John Wiley & Sons,.
5. Medhat, F., Chesmore, D. and Robinson, J. (2020) Masked conditional neural networks for sound classification, *Applied Soft Computing*, 90 , pp. 106073. DOI: 10.1016/j.asoc.2020.106073.
6. Mesaros, A., Heittola, T., Virtanen, T. and Plumbley, M. D. (2021) Sound Event Detection: A tutorial, *IEEE Signal Processing Magazine*, 38 (5), pp. 67-83.
7. Salamon, J., Jacoby, C. and Bello, J. P. (Nov 3, 2014) A Dataset and Taxonomy for Urban Sound Research, in: ACM, pp. 1041-1044.
8. Shrestha, A. and Mahmood, A. (2019) Review of deep learning algorithms and architectures, *IEEE Access*, 7 , pp. 1. DOI: 10.1109/ACCESS.2019.2912200.
9. Su, Y., Zhang, K., Wang, J. and Madani, K. (2019) Environment sound classification using a two-stream CNN based on decision-level fusion, *Sensors (Basel, Switzerland)*, 19 (7), pp. 1733. DOI: 10.3390/s19071733.

