Machine Learning in Multimedia Data: Vehicle Tracker

Baou Evangelia

Zouros Michael

Problem Presentation

- Collect audio signals from different static places near a road
- Extract useful information from these signals
- Implement different types of Machine and Deep Learning algorithms, in order to train them to recognize, given a new audio signal, how many vehicles have passed during the signal's duration
- Roads can be of any type and size (eg. 1 lane, 2 lanes, boulevard, highway)
- Vehicles can be of any type and size (eg. cars, motorcycles, buses) as long as they have an engine (eg bicycles are not calculated)

Data Collection

- Took place on different locations near various types of roads
- The recordings were made via our smartphones' microphones and had a duration that lasted just over 30 second
- Over 200 different samples were collected (172 used)
- Each sample was named in accordance to it's label (eg. 15_recording93 indicates that 15 vehicles have passed in the duration of recording93)

Data Preparation

Data preparation took place on the Audacity application and included the following:

- Audio trimming to exactly 30 seconds
- Noise reduction (eg. reduce the volume of bird singing or people talking) where possible
- Stereo to Mono transformation (one of our smartphones was producing mono signals)
- Conversion to WAV, for better quality

Data Exploration & Future Extraction

We experimented with a plethora of features and algorithms in order to achieve the best possible results. We concluded in three different algorithmic approaches, each feeded with different features. Specifically:

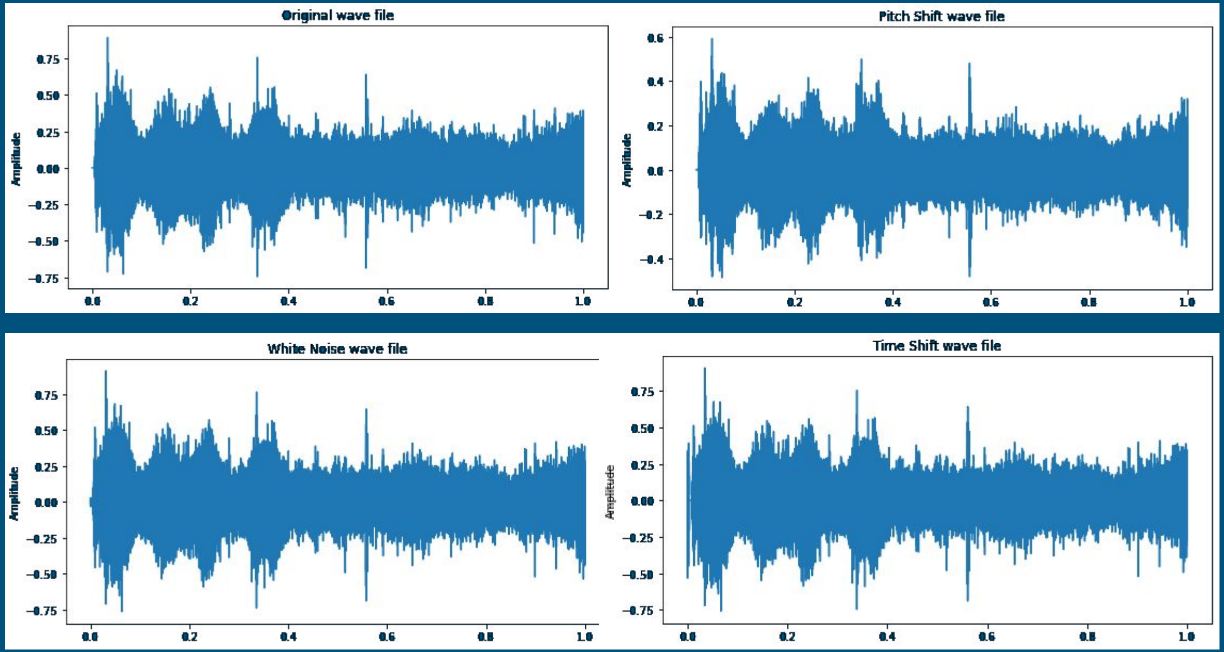
- We created the Mel Spectrograms of our audio signals and used them as input on a CNN model
- We extracted the Mel Frequency Cepstral Coefficients (MFCCs) and used them as input on an RNN-LSTM model
- We extracted various Time & Frequency Domain features (Spectral Centroid, Root Mean Square Energy, etc.) and used them as input on an SVC

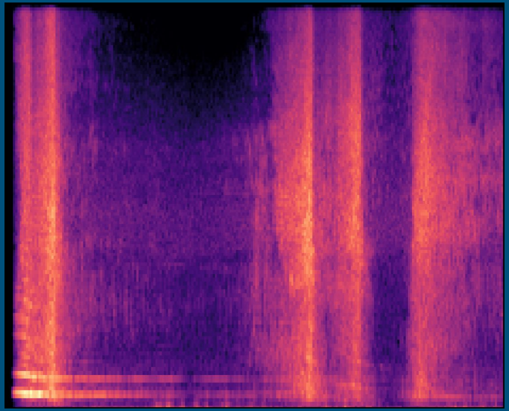
Data Revision and Augmentation

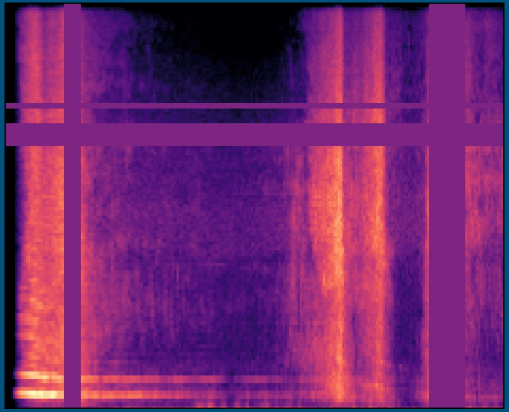
After experimenting with our initial dataset, we proceeded and achieved data augmentation for all of our three approaches. Specifically:

- For our CNN model, we used used masking/filtering on our Mel Spectrograms, managing to double the size of our initial dataset to 344 audio signals
- For our RNN-LSTM model, we generated new audio files from the original using various sound augmentation techniques, like White Noise Addition, Time and Pitch Shifting
- For our SVC, we also used Pitch Shifting (up and down), thus tripling our initial dataset

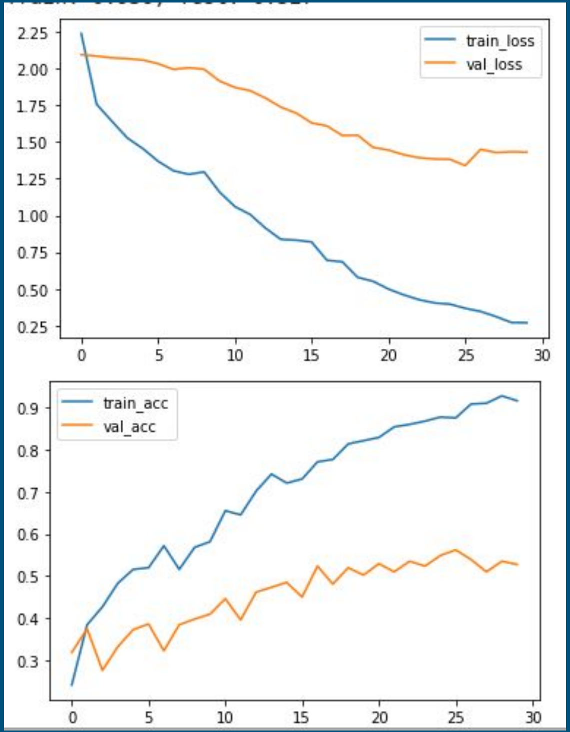
Augmented Data







1st Approach - CNN Results

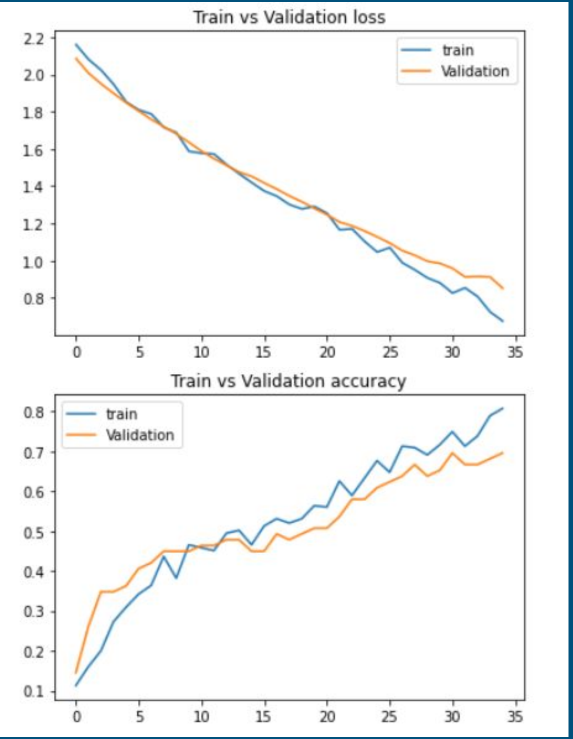


Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 254, 254, 32)	896
module_wrapper_1 (ModuleWrap	(None, 254, 254, 32)	128
activation_3 (Activation)	(None, 254, 254, 32)	0
max_pooling2d_3 (MaxPooling2	(None, 127, 127, 32)	0
dropout_2 (Dropout)	(None, 127, 127, 32)	0
conv2d_4 (Conv2D)	(None, 125, 125, 64)	18496
activation_4 (Activation)	(None, 125, 125, 64)	0
max_pooling2d_4 (MaxPooling2	(None, 62, 62, 64)	0
conv2d_5 (Conv2D)	(None, 60, 60, 128)	73856
activation_5 (Activation)	(None, 60, 60, 128)	0
max_pooling2d_5 (MaxPooling2	(None, 30, 30, 128)	0
dropout_3 (Dropout)	(None, 30, 30, 128)	0
flatten_1 (Flatten)	(None, 115200)	0
dense_2 (Dense)	(None, 32)	3686432
dense_3 (Dense)	(None, 8)	264

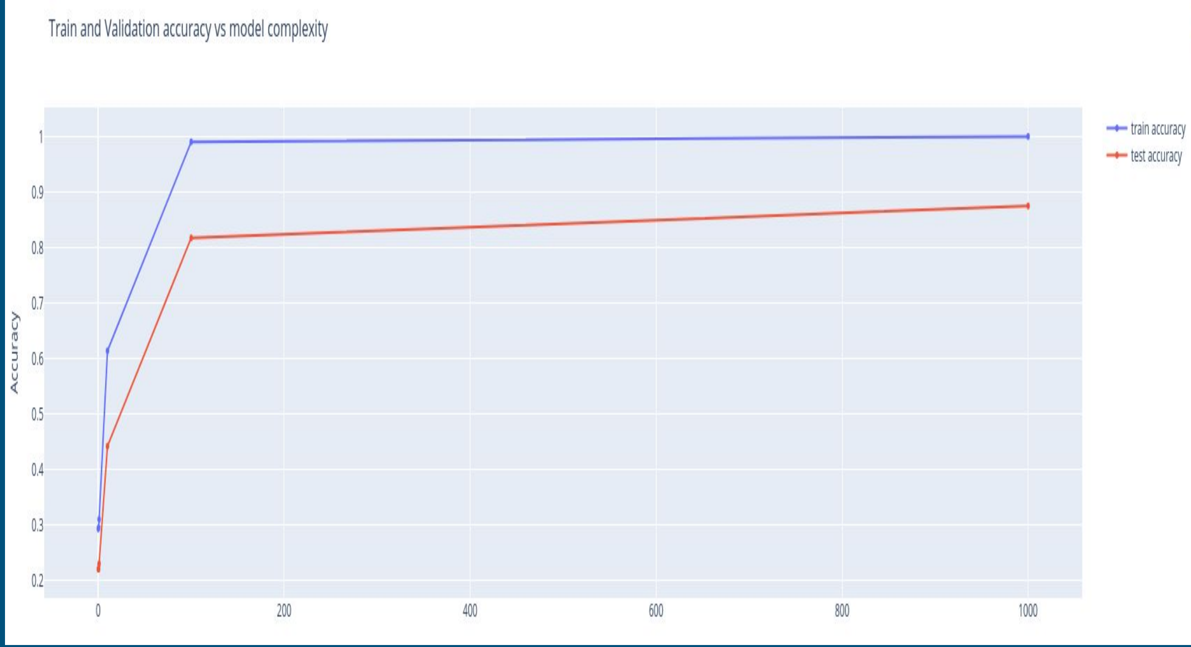
Total params: 3,780,072
Trainable params: 3,780,008
Non-trainable params: 64

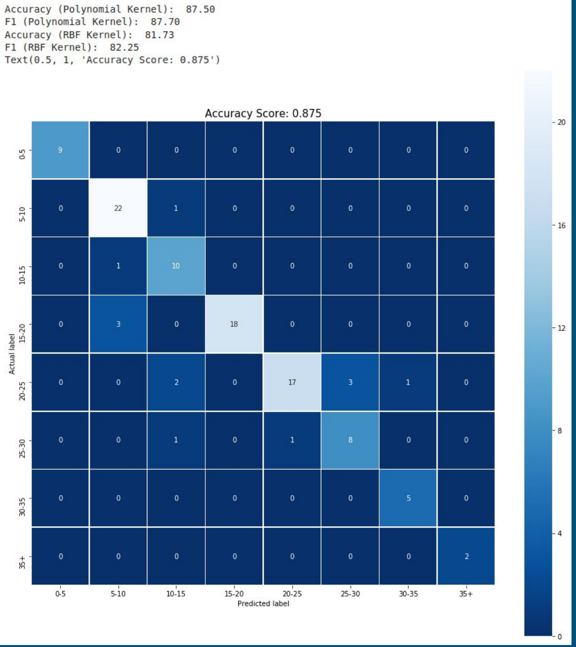
2nd Approach - RNN-LSTM Results

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_8 (LSTM)                (None, 469, 128)          86528
_________________________________________________________________
lstm_9 (LSTM)                (None, 128)               131584
_________________________________________________________________
dropout_8 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_8 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_9 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_9 (Dense)              (None, 8)                 520
=================================================================
Total params: 226,888
Trainable params: 226,888
Non-trainable params: 0
_________________________________________________________________
```



3rd Approach - SVC Results





SVC Predictions & Live Presentation

Table 1: *Vehicle Number Predictions*

Passed	Label Predicted	Classified
11	2 (11-15)	Correct
42	7 (35+)	Correct
29	3 (15-20)	Wrong
9	1 (5-10)	Correct
10	1 (5-10)	Correct
4	0 (1-5)	Correct
33	4 (20-25)	Wrong
17	4 (20-25)	Wrong

Discussion

- The results suggest that our two NN models don’t perform as well as our SVC algorithm
- The RNN-LSTM model seems to perform much better than the CNN during training, but still faces problems during the prediction stage
- Our SVC algorithm seems to have the best prediction results
- Prediction seems way harder in samples with lots of vehicles passing by
- Small and noisy datasets are better to be approached via the traditional machine learning techniques and algorithms

Future Work

The study can be extended in a number of different ways:

- Better recording devices with noise reduction, different types of roads
- It can be implemented using a Regression approach
- It can be extended from vehicle detection to vehicle detection and classification
- A combination of both acoustic and image/video data could yield far more better prediction results

References

[1]Dalir, Ali, Ali Asghar Beheshti, and Morteza Hoseini Masoom. "Classification of vehicles based on audio signals using quadratic discriminant analysis and high energy feature vectors." arXiv preprint arXiv:1804.01212 (2018).

[2]George, Jobin, et al. "Exploring sound signature for vehicle detection and classification using ANN." International Journal on Soft Computing 4.2 (2013): 29.

[3]Wieczorkowska, Alicja, et al. "Spectral features for audio based vehicle and engine classification." Journal of Intelligent Information Systems 50.2 (2018): 265-290.

[4]Johnstone, Michael N., and Andrew Woodward. "Automated detection of vehicles with machine learning." (2013).

[5]Chellappa, Rama, Gang Qian, and Qin-fen Zheng. "Vehicle detection and tracking using acoustic and video sensors." 2004IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 3.IEEE, 2004.

Thank you!

SVM

RNN

CNN