Robotic Inference Project Writeup

Abstract

The first project in term 2 of the Udacity Robotics Nano Degree program requires students to initiate their own inference project inclusive of data acquisition. The project builds on the initial reference project for digit image recognition inside the supplied Nvidia Digits environment.

The project ideas are the students own and must have at least 3 classification categories eg defective item vs normal item with classes (no item, defective item, normal item).

Introduction

Pedestrian and bicycle lanes are often crowded with many people not aware of or selectively ignoring the signage. It can lead, to an unsafe or hazardous environment, for all that use it with police officerw reluctant to enforce the rules via fines.

The concept selected, in this project, was to classify an image as either containing a pedestrian, not-pedestrian or background.

The goal being, that some sought of visual representation via a screen with a smile or a frown, could be given to act, as a robotic traffic controller. Other potential instantiation could include a torso using upper body movement to signal good or bad behaviour.

Background / Formulation

During the initial inference task, on supplied data, GoogLeNet was chosen as it had a good inference rate per image with reasonable accuracy. Using the Adam Optimiser with an initial learning rate of .001, it was able to meet the numerical requirements of inference time below 10 ms with accuracy > 75%. The input used for this reference model in Nvidia's DIGITS was 256x256 3 channel colour images.

Similar requirements of accuracy would be required for this inference project. It was not necessary to be a 100% accurate as a smiling or frowning face at least makes people think about what they are presently doing. It was not going to be used to issue fines or other enforcement notices. Video cameras would stream image data between 24-30fps which means that an Inception, VGG model and some ResNet models may be too slow for inference in real time.

With the additional perception that colour could also be useful in detecting pedestrian vs not-pedestrian, GoogLeNet was again chosen for this project using the Adam Optimiser with an initial learning rate of .001.

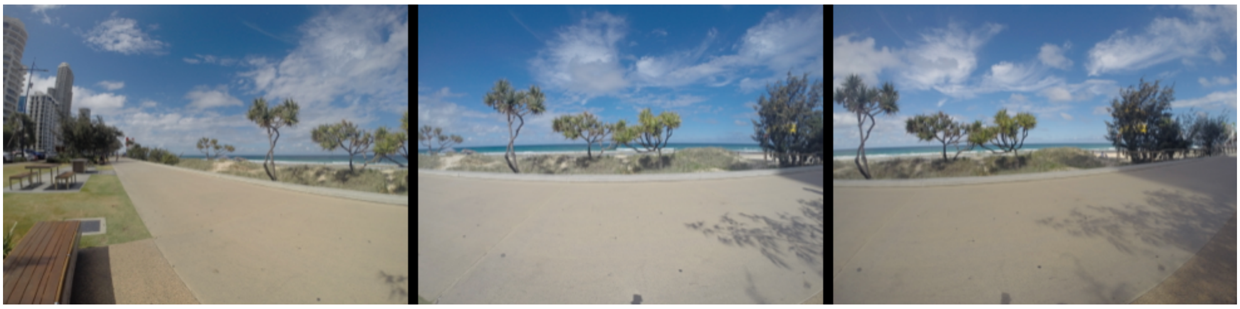
Other experimentation with using a higher initial learning rate of 0.01 with the above configuration over 5 epochs did not improve validation accuracy which remained around 50%. Similarly AlexNet over 5 epochs with the same Adam optimiser and learning rate did not increase accuracy. One experiment was performed using GoogLeNet with RMSProp optimiser and an initial learning rate 0.001 which did not result in improvement of validation accuracy but did have significantly higher training loss so was also not progressed.

Data Acquisition

A GoPro mounted on a tripod was used. It was positioned on the side of a pedestrian esplanade at Surfers Paradise, Gold Coast, Australia. As it was summer holidays a reasonable amount of varying traffic was expected. The background looked over the ocean to have a consistent image where there was not going to be movement (besides cloud) other then what was on the esplanade.

GoPro Data Acquisition

Three angles (facing left, centre and right) were used for capture per the following graphic:



Camera Background Angles

The GoPro was setup in wifi mode for time lapse capture, which was controlled via an iPhone. Initially 2 seconds elapsed was used, which eventually was dropped to 0.5. Using the GoPro time-lapse feature, meant that individual jpeg files were captured as opposed to a MP4 video.

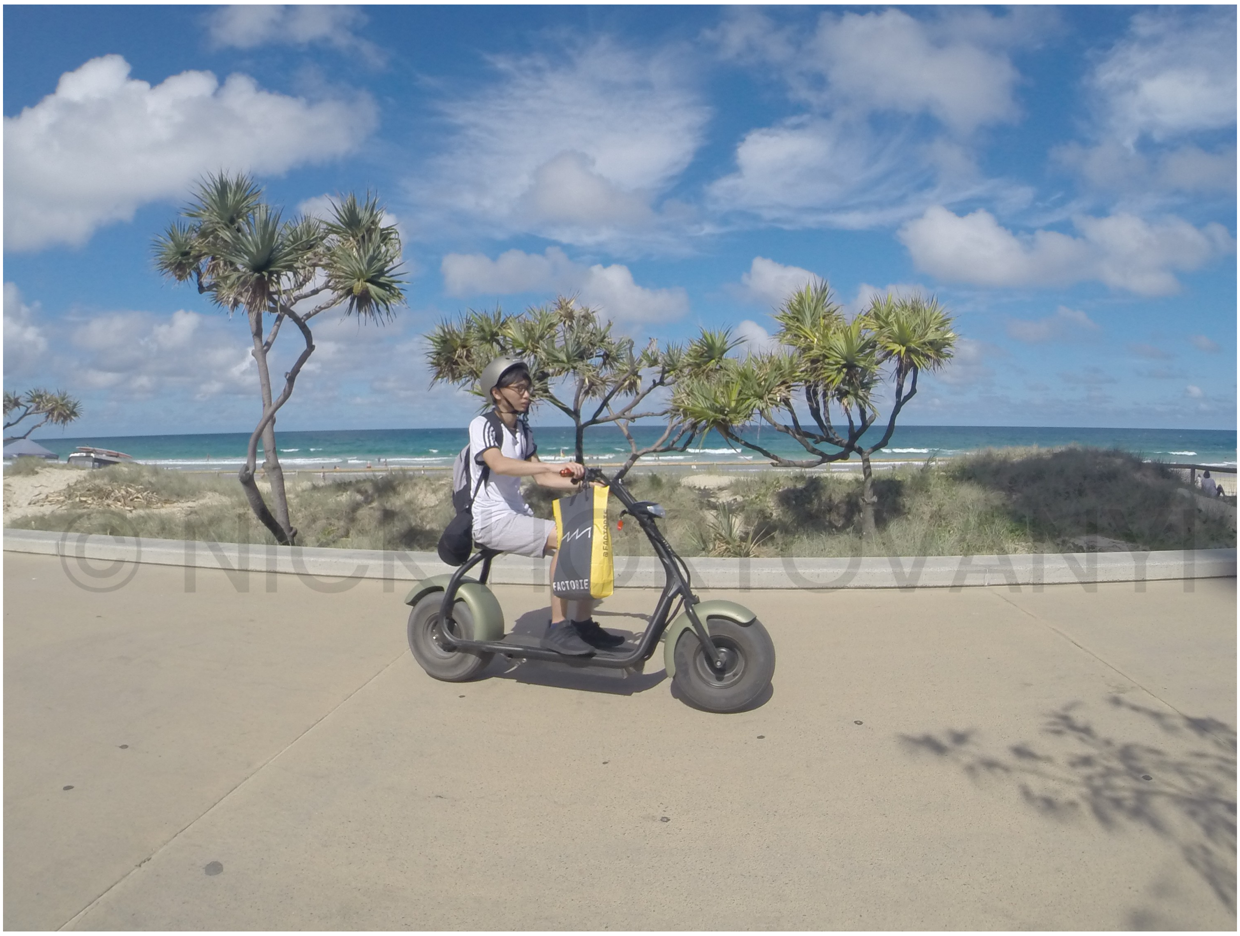
Using the iPhone to control the control, meant that I could visualise what was coming before starting the next capture batch.

After the capture the images were manually placed into a directory for each category.

Image data was captured for the three categories background (322), pedestrian (349) and not-pedestrian (94). It became apparent at this time that not enough not-pedestrian image data had been captured. This was mainly due to the initial 2 seconds elapsed time used. Due to high heat and humidity of the Australian summer, in the following afternoons, it was not practical to capture more data from the same spot.

Skateboarders were placed in the pedestrian category.

An example of a pedestrian and not-pedestrian follows. Background examples are as above.



not-pedestrian example

pedestrian example

A jupyter notebook was used to create a generator to supplement the data by randomising the image brightness, randomly flipping the images vertically and jittering the images randomly in the x (by + or - 25 pixels),y (by + or - 50 pixels) planes to create supplemental image data.

The images were also resized to 256 x 256 and saved as PNGs.

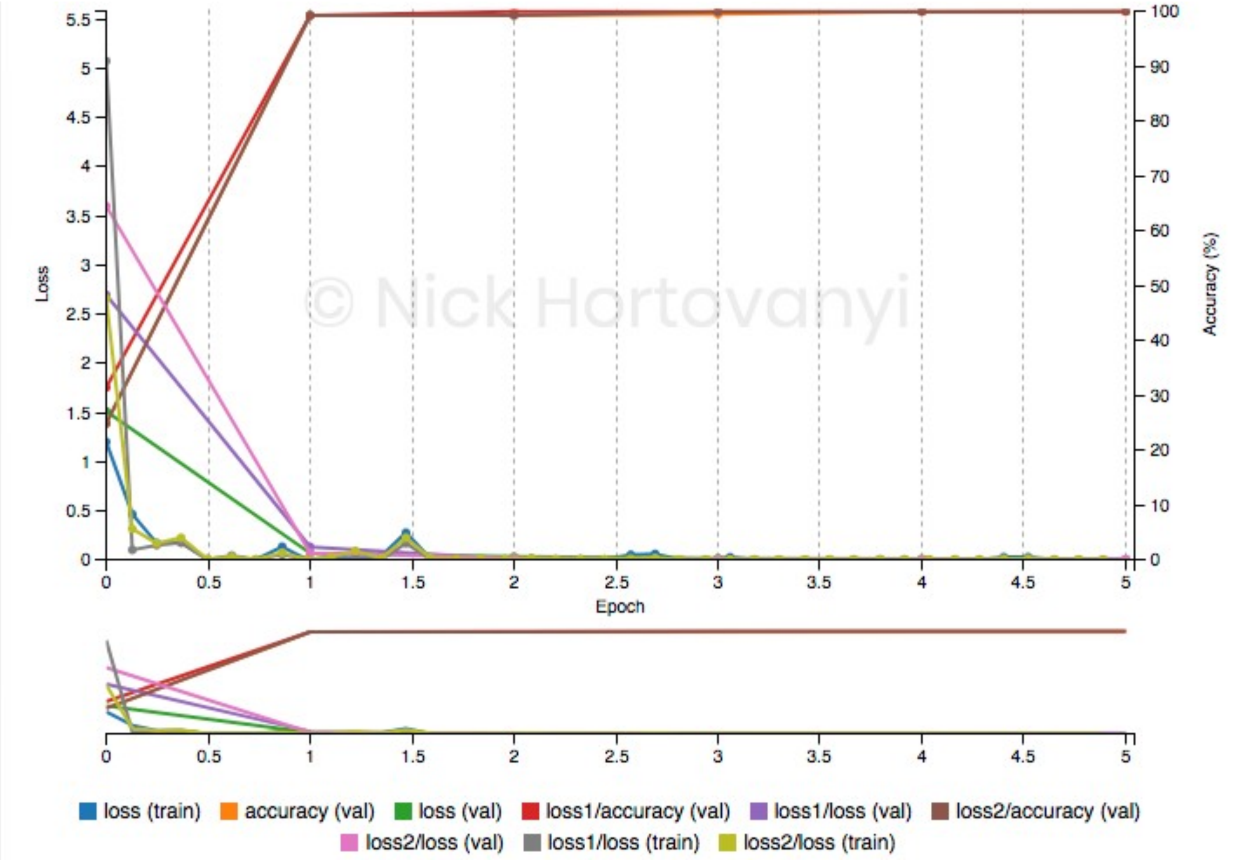
The final generated supplemental data had 2000 not-pedestrian, 1000 pedestrian with 1000 background images.

Results

The initial inference task, on supplied data, GoogLeNet was chosen as it had a good inference rate per image with reasonable accuracy. Using the Adam

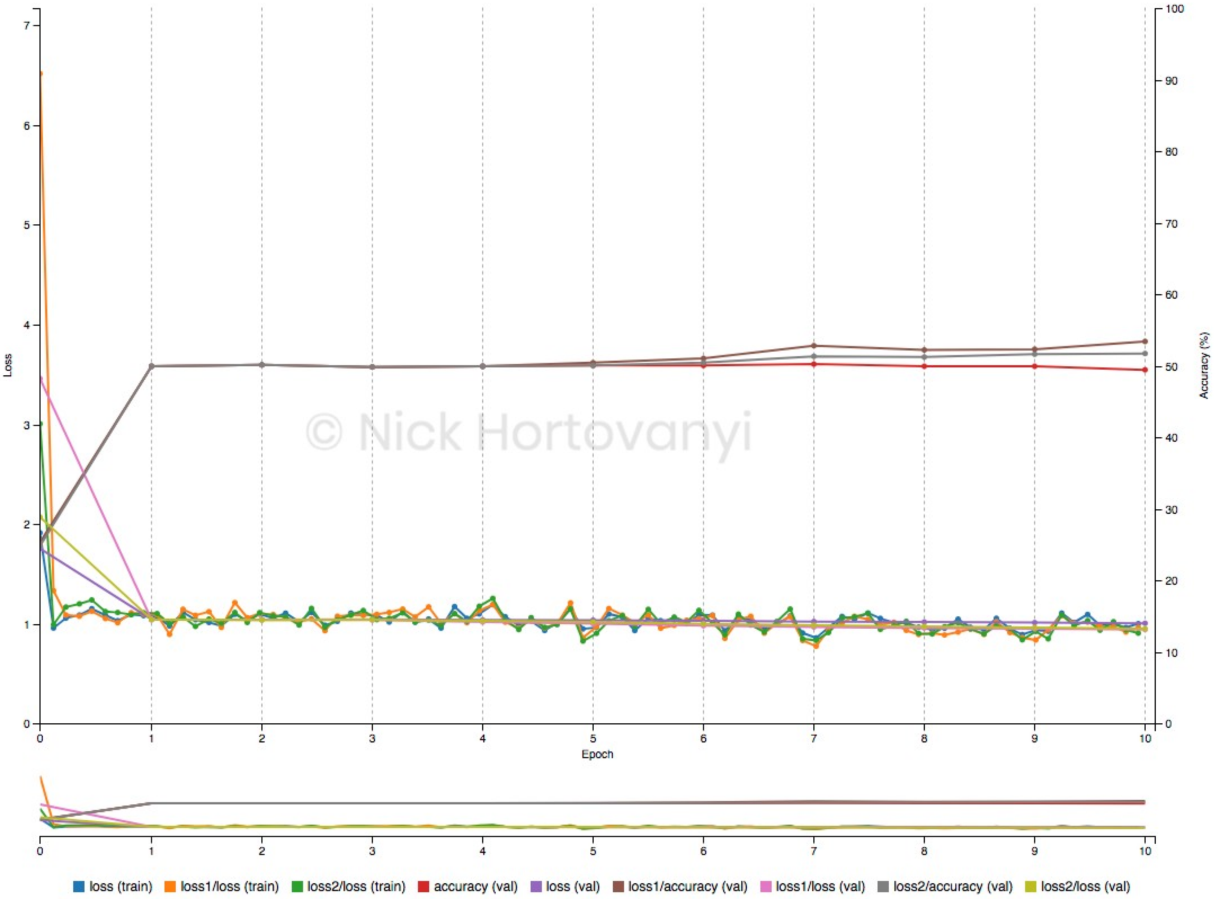
Optimiser with an initial learning rate of .001, it was able to meet the numerical requirements of inference time below 10 ms (~5 ms actual) with accuracy > 75% (75.40984% actual).

During training of the initial inference task 100% validation was achieved per the following training graph after 5 epochs.



Training Graph

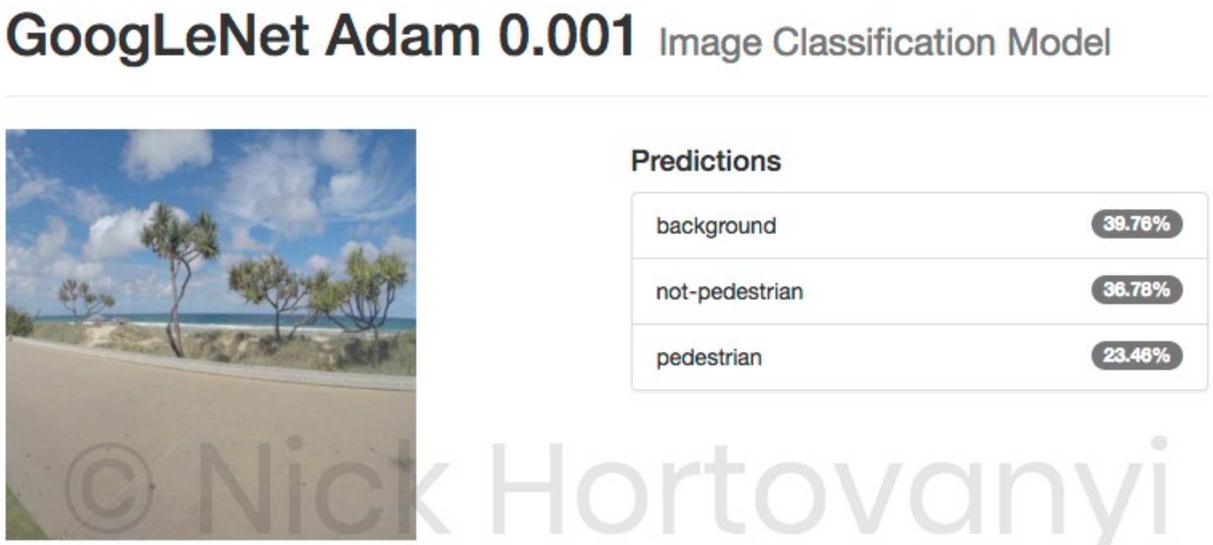
However similar training results were not achieved for this inference project on captured data. The following training graph after 10 epochs follows



Project Training Graph

This had a validation accuracy of ~50%.

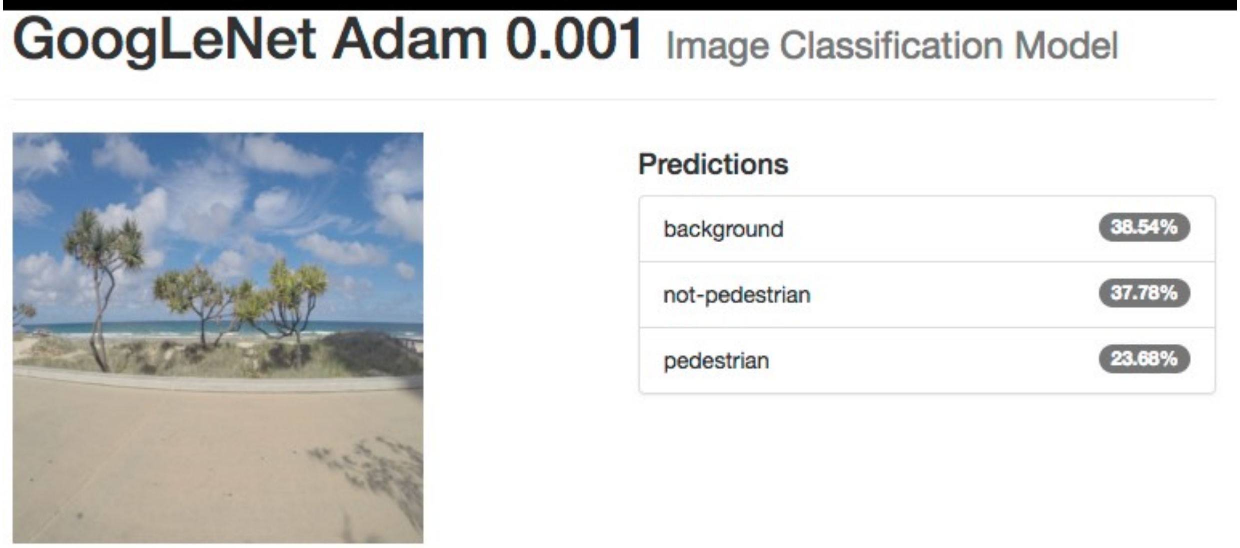
The following are the results of two randomly selected images per classification category uploaded as original high-res jpeg images from the GoPro.



GoogLeNet Adam 0.001 Image Classification Model

Predictions

background	39.76%
not-pedestrian	36.78%
pedestrian	23.46%

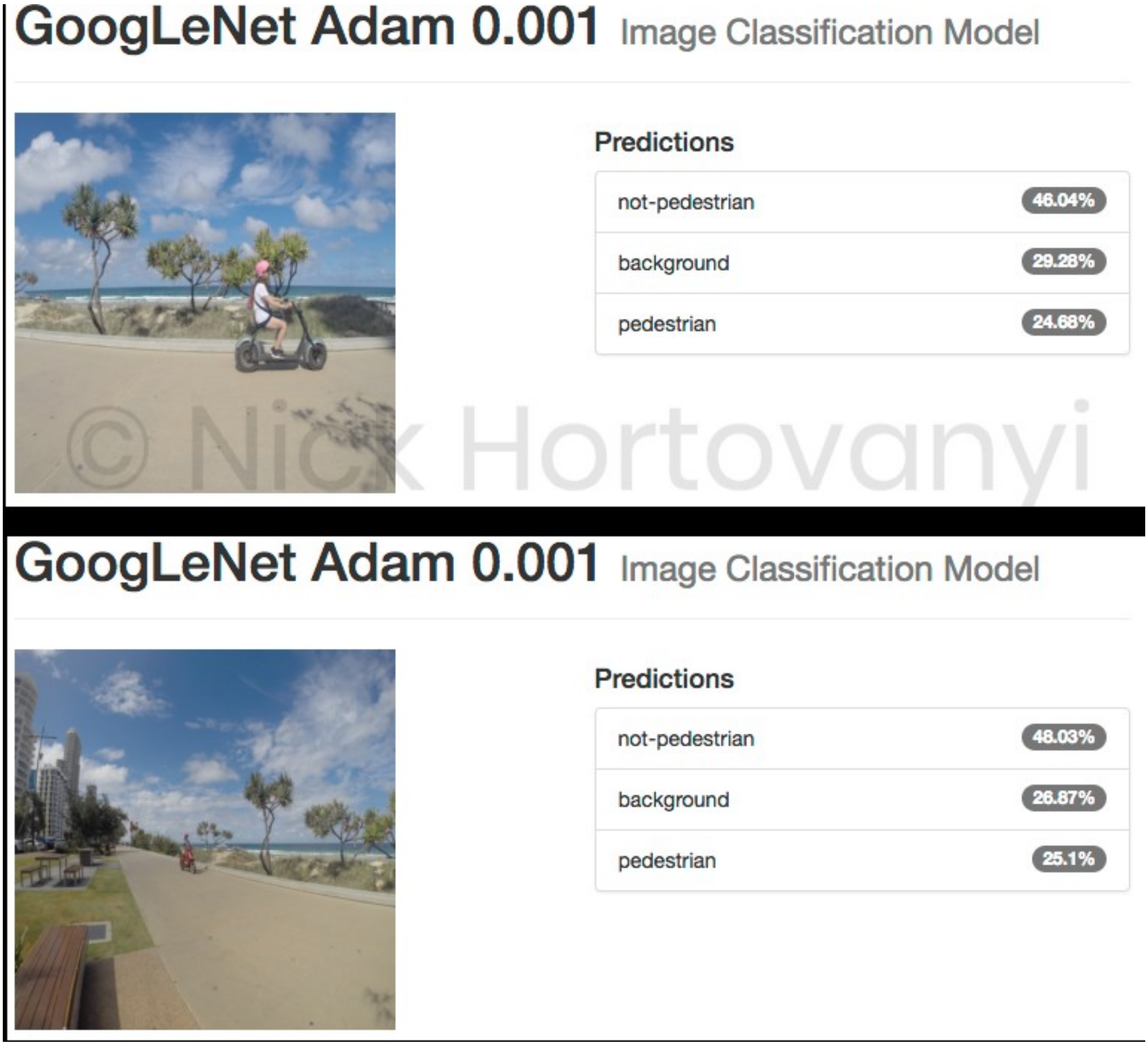


GoogLeNet Adam 0.001 Image Classification Model

Predictions

background	38.54%
not-pedestrian	37.78%
pedestrian	23.68%

Inference Background Sample



GoogLeNet Adam 0.001 Image Classification Model

Predictions

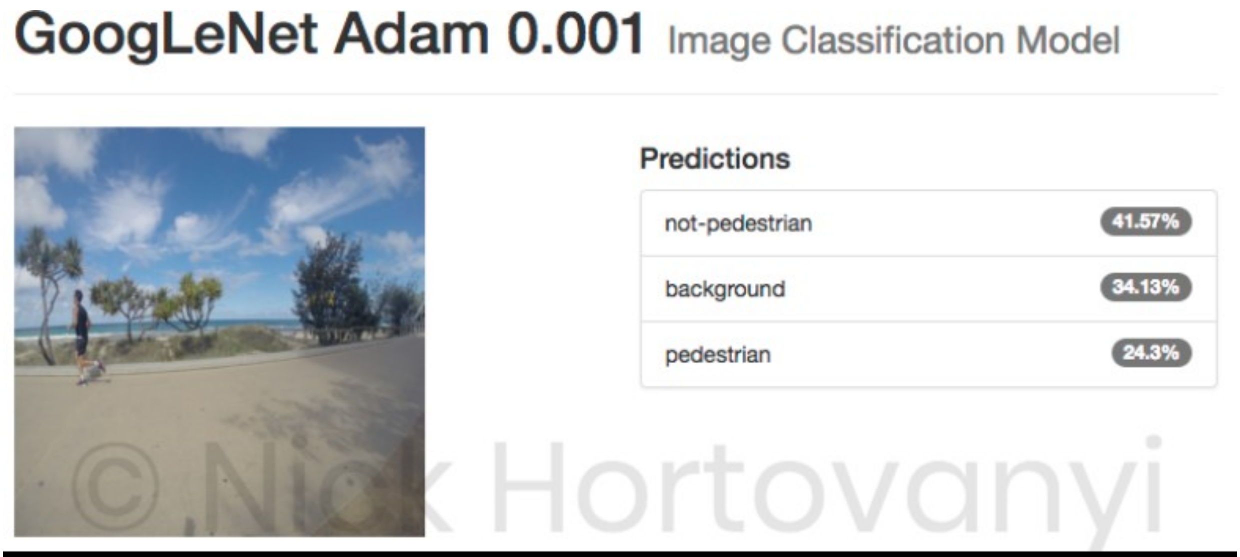
not-pedestrian	46.04%
background	29.28%
pedestrian	24.68%

GoogLeNet Adam 0.001 Image Classification Model

Predictions

not-pedestrian	48.03%
background	26.87%
pedestrian	25.1%

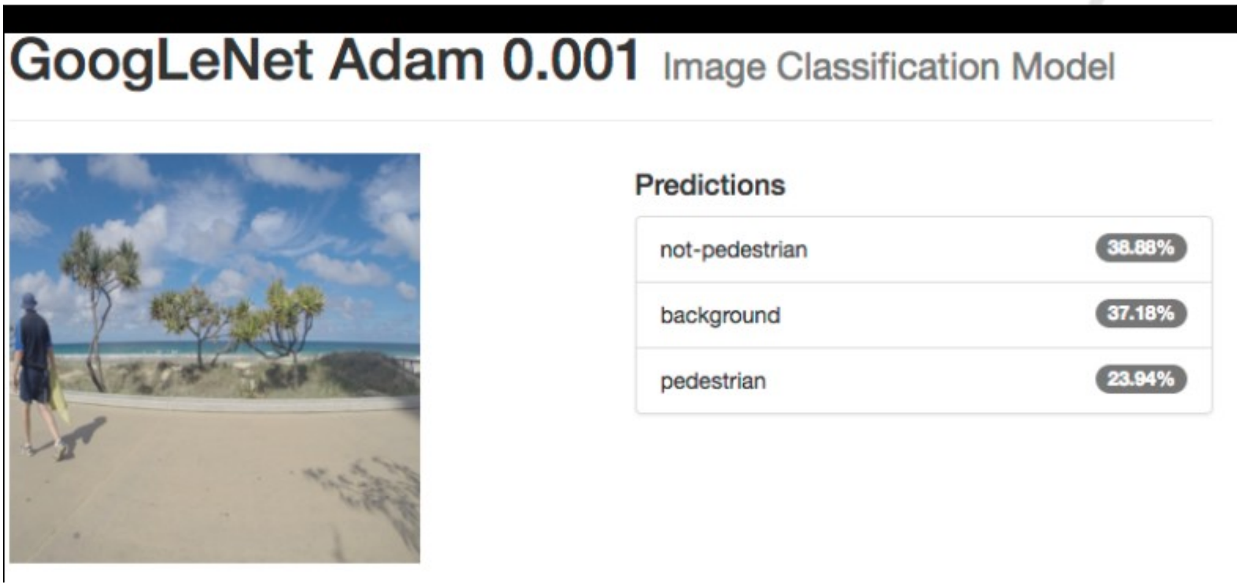
Inference Not-Pedestrian Sample



GoogLeNet Adam 0.001 Image Classification Model

Predictions

not-pedestrian	41.57%
background	34.13%
pedestrian	24.3%



GoogLeNet Adam 0.001 Image Classification Model

Predictions

not-pedestrian	38.88%
background	37.18%
pedestrian	23.94%

Inference Pedestrian Sample

The indicative inspection suggests that there was insufficient data to get a result > 75% for this project at this time. It also appears that the model as trained can not distinguish between pedestrian and not-pedestrian but can distinguish a background image.

Inference times were not tested separately as GoogLeNet is known to have a fast inference time which would be sufficient for this project.

Discussion

The dataset collected did not have enough sample images. This was as a result of using time-lapse with too high a value. In hind site, a combination of 0.5-1 sec time-lapse for slow moving pedestrians with >30 FPS video for higher speed moving non-pedestrians would of allowed for more data.

Of note is when there is a combination of non-pedestrian and pedestrian in the same frame led to the though of potentially using object detection first to find a window to classify. This was not implemented in this version however it would have led to a more accurate ability to classify as the background would be eliminated ie if no objects detected it must be a back ground image.

For this project the duration of a consistent display of say 2 to 3 seconds to the passing pedestrian and non-pedestrian traffic would drive the ultimate inference time required. It would suggest that it needs to be an average classification of a 1 second or two when leading up to where the video is captured.

Of note were skateboarders. There is only the skateboard (which has a low profile in the image) that distinguishes it from a pedestrian as velocity is not taken account with single images.

In addition the depth (away from the camera) of the traffic passing by changes the size of the object that needs to be classified. Pre-filtering and zooming these to a consistent size may improve accuracy.

Future Work

Whilst the project did not achieve a good train validation result, it has laid the foundation for future iterations. There is potential to capture more data and use object detection to refine the training and inference steps of the project.

Providing soft means to monitor and influence peoples decisions regarding signage and the associated rules for safe usage of (non-vehicle) transit paths would potentially be received positively by the community. Police forces lack the budget or people to enforce these rules and they are reluctant to issue minor infringement notices (with potential to destroy good will in the community). Hence a more subtle robotic person could improve the situation in a cost effective means.

The size of this market is unknown. However speed cameras using smiley and sad faces are used around the local area where I live. They are having a positive impact on driver behaviour in the areas deployed. Thus there is a market for a more community friendly and automated means to impact on behaviour.