Quality Control of Training Data

Chau Yuan Qi,
Intern@ST Autonomous Solutions

Abstract

We present a heuristic that can effectively flag mislabelled data to be removed from the training data. The heuristic has achieved an accuracy of 0.9 after manual verification of the flagged data.

1. Introduction

In most classification problems, the state-of-the-art model is good enough to generalize and learn the patterns. However, the performance can be badly affected when there are mislabellings and outliers in the training data. In this report, we will investigate a heuristic to effectively sample out most of the mislabelled data as well as outliers in any particular class.

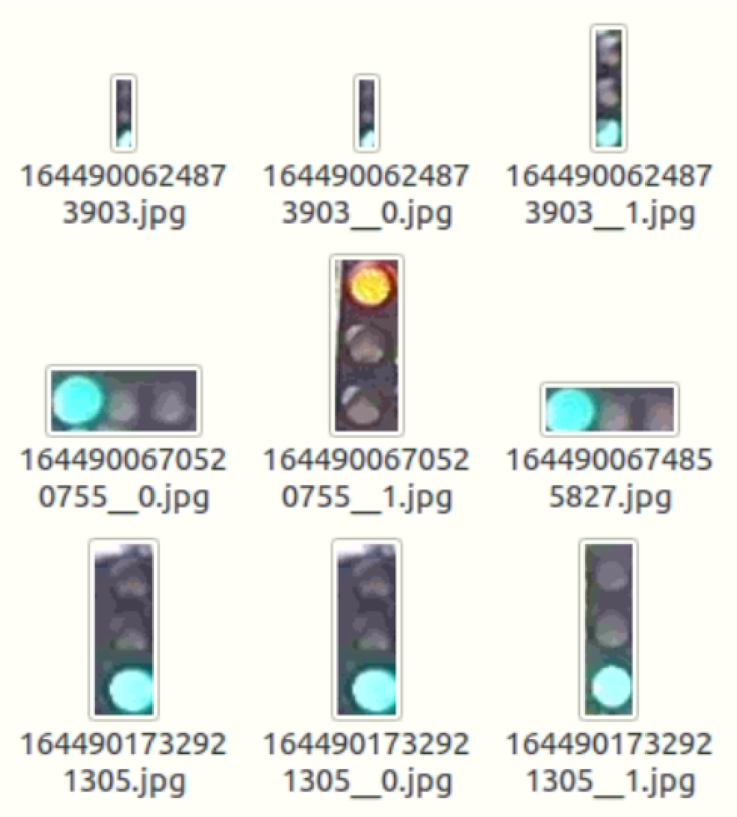


Figure 1: Example of a red light mislabelled as a green light

2. Implementation

Firstly, the training data is being trained on any model (resnet18 in this experiment) until convergence. After convergence, the entire training data is then passed through the model to find out the failed predictions. The heuristic is to gain insights from the failed predictions, which are what the model failed to learn from the data). From the failed predictions, there will be a distribution of confidences by the model (determined by a softmax layer) for each prediction as shown in Fig. 2. The confidence distribution can be interpreted as the probability of the object being in the particular predicted class. Ideally, the distribution should skew to the left, because the model should not be confident in something that they predicted wrongly.

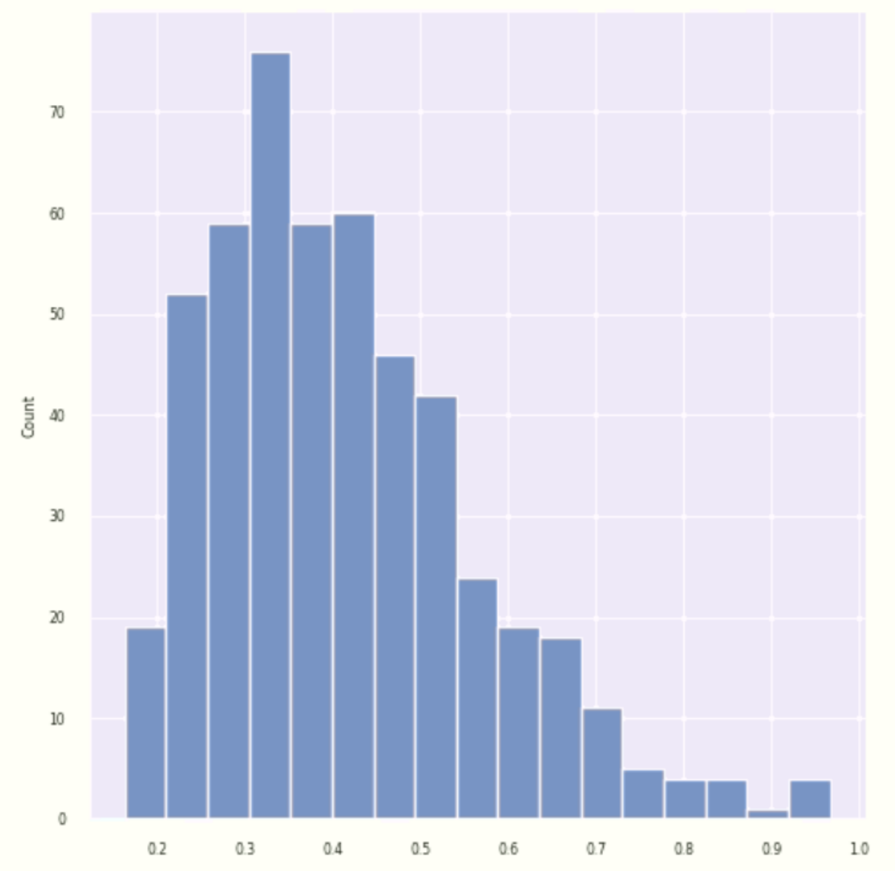
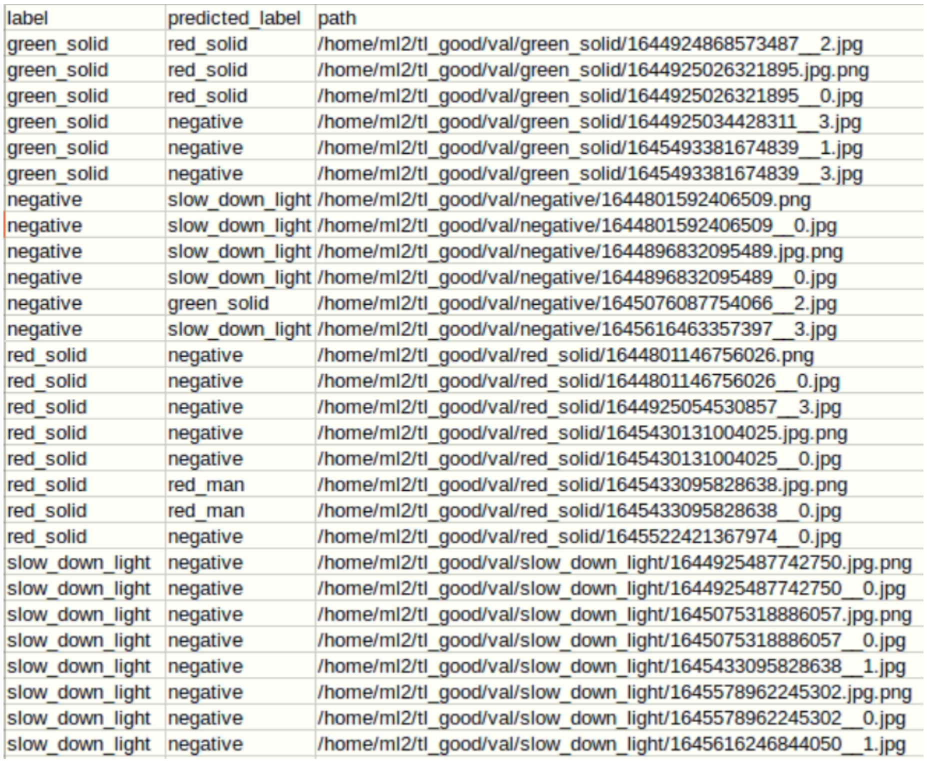


Figure 2: Example of a confidence distribution

2.1 Flagging the Mislabelled Data

The problem with Fig. 2 is that there are some samples that the model is extremely confident to be true but in fact is wrong. The threshold is arbitrarily set at 0.7 in this experiment, in which any predictions done by the model at 0.7 or higher confidence but is not of the same class as the ground truth is flagged as a mislabel. The hypothesis is that if this particular data from class A is mislabelled to be class B, the model will have seen similar data in class A than class B, and therefore have a higher confidence in the predictions being in class A than class B. This is done under the assumption that the model will be able to predict well for both class A and class B due to the majority of the classes being correctly labeled and therefore still allowing the model to generalize.

The threshold can be set higher if a higher accuracy in capturing mislabellings is needed, and can be set lower if there is a lower tolerance for badly labeled data (note that correctly labeled data might be flagged as threshold go lower). By using this method @ 0.7 confidence on the traffic_light dataset, we found out that the flagged mislabelled data have an accuracy of 0.9 after being manually verified. A mislabels.csv file as shown in Fig. 3 will be automatically generated in order for further investigation or easy removal of flagged data during training.



label	predicted_label	path
green_solid	red_solid	/home/ml2/tl_good/val/green_solid/1644924868573487__2.jpg
green_solid	red_solid	/home/ml2/tl_good/val/green_solid/1644925026321895.jpg.png
green_solid	red_solid	/home/ml2/tl_good/val/green_solid/1644925026321895__0.jpg
green_solid	negative	/home/ml2/tl_good/val/green_solid/1644925034428311__3.jpg
green_solid	negative	/home/ml2/tl_good/val/green_solid/1645493381674839__1.jpg
green_solid	negative	/home/ml2/tl_good/val/green_solid/1645493381674839__3.jpg
negative	slow_down_light	/home/ml2/tl_good/val/negative/1644801592406509.png
negative	slow_down_light	/home/ml2/tl_good/val/negative/1644801592406509__0.jpg
negative	slow_down_light	/home/ml2/tl_good/val/negative/1644896832095489.jpg.png
negative	slow_down_light	/home/ml2/tl_good/val/negative/1644896832095489__0.jpg
negative	green_solid	/home/ml2/tl_good/val/negative/1645076087754066__2.jpg
negative	slow_down_light	/home/ml2/tl_good/val/negative/1645616463357397__3.jpg
red_solid	negative	/home/ml2/tl_good/val/red_solid/1644801146756026.png
red_solid	negative	/home/ml2/tl_good/val/red_solid/1644801146756026__0.jpg
red_solid	negative	/home/ml2/tl_good/val/red_solid/1644925054530857__3.jpg
red_solid	negative	/home/ml2/tl_good/val/red_solid/1645430131004025.jpg.png
red_solid	negative	/home/ml2/tl_good/val/red_solid/1645430131004025__0.jpg
red_solid	red_man	/home/ml2/tl_good/val/red_solid/1645433095828638.jpg.png
red_solid	red_man	/home/ml2/tl_good/val/red_solid/1645433095828638__0.jpg
red_solid	negative	/home/ml2/tl_good/val/red_solid/1645522421367974__0.jpg
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1644925487742750.jpg.png
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1644925487742750__0.jpg
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645075318886057.jpg.png
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645075318886057__0.jpg
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645433095828638__1.jpg
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645578962245302.jpg.png
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645578962245302__0.jpg
slow_down_light	negative	/home/ml2/tl_good/val/slow_down_light/1645616246844050__1.jpg

Figure 3: Mislabels.csv

2.2 Flagging the Outliers

By using the same heuristic, we can look at the lower spectrum of the confidence distribution and that predictions that have too low confidence should be flagged as outliers. The hypothesis is such that if the maximum confidence that the model has on a particular class is still very low, it means that the model couldn’t differentiate the input data into any particular class, therefore making the data an outlier. However, from the experiments, the results are not very convincing and therefore this heuristic will not be recommended.

3. Conclusion

The presented heuristic is shown to effectively flag out mislabelled data using with high accuracy. I believe that by filtering out the flagged data, the performance of the model will increase greatly. Not only that, this method can also be used for quality control of our dataset annotation.