Interview Question:

Data management and analysis

Introduction

Training machine learning models an interesting and important process in our work, but knowing how to train and fine tune a deep learning model does not help if you have inaccurate data as input, and testing does not mean anything if ground truth you compare to is not right. That's why public dataset like GTSDB are so useful, but sometimes you need to label your own data in order to train for your particular use-case. Annotation is a time consuming and boring job, so sometimes mistakes are made in the process. Error like omission or noisy bounding boxes may confuse the neural network and reduce it's accuracy.

Illustration 1: Examples of red round sign

Managing the data and constantly trying to improve their quality is something we spend a lot of effort on. The goal of this set of question is to make you more familiar with this process and to discover it's usefulness. For that we will stay in the continuity of the red-round-traffic-sign detection question set (If you didn't do it, no worries, it's not a requirement). We are going to use the same GTSDB dataset, this dataset give us bounding boxes for traffic signs in Germany. The other questionnaire was using it to detect red-round-traffic-sign, example of which you can find on Illustration 1.

However, if you did the set of question and reviewed the data you may have noted that some sign in the picture above are not annotated in GTSDB. More precisely, the no stop and no parking traffic sign that are on Illustration 2.

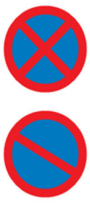


Illustration 2: Signs not included in GTSDB

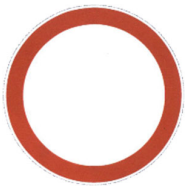
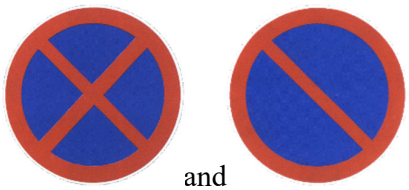
This question set will then have two main focus, first explore how we can improve the existing dataset include this new sign, second, how to asses the performance of a model given it's results and a ground truth that may contain errors.

Question 1

Update the gt.txt file included in the GTSDB dataset to include the two missing signs, using the following class labels:

	
43	44

To help you in this task we provide you the detection done by one of our trained deep learning model (yolo3). This model was trained using image from china including TT100K dataset as well as some data we collected and annotated in the lab. This model is trained to detect five different kind of signs:

				 and
RedRoundSign	pg	ps	pne	pn

As you can see this signs are not exactly the same than the one in Germany but are still similar and generalization is still good enough to use be used on German data.

The json file containing the prediction can be found at: https://www.dropbox.com/s/b9aootj3z9gl1vs/GTSDB.json?dl=0

The format of this file is straightforward. You can see bellow what it look like. Basically the “frames” field is a list of frames, each frame having a name (“frame_number”) and a list of bounding boxes (“signs”). Each bounding boxes as some “coordinates” as a list of x coordinate of the top left corner, y coordinate of the top left corner, width and height. "detection_confidence" is the confidence given by the network for this prediction (threshold was 0.5) and “class” is the predicted sign class.


```
{
  "output": {
    "video_cfg": { [...] },
    "framework": { [...] },
    "frames": [
      {
        "frame_number": "00001.ppm",
        "RoIs": "pg,976,334,49,54;",
        "signs": [
          {"coordinates": [976,334,49,54],
           "detection_confidence": 0.917523,
           "class": "pg"
          }]
```

```
        },
        {
          "frame_number": "00002.ppm",
          "RoIs": "",
          "signs": []
        },
[…]
```


Question 2

In this second stage, you will write a script that compare your updated ground truth and the prediction outcome. You are free to use whatever metrics make the more sense to you: Map, Recall/Precision, TP/FP/FN… Then using the results from your script, you will extract some problematic cases from the prediction, such as false negative detection, false positive, wrong classes…

How confident are you that the number of pn signs that are not detected given by your script is the actual number?

If you had to improve (fine tune) the model used for detection to get better results on this dataset, what would you do?

Do you thinks that your new annotations are accurate enough for training? What kind of bias may you have introduced in the dataset and how may it affect the training outcome?

Reports

For each part you will send us (james.tsai@ce.gatech.edu, zyy@gatech.edu and pyu68@gatech.edu) a report of around 2 or 3 pages in pdf format with your analysis and comments. You will also send us your code by giving us the link to your github / gitlab repository.

If you have any question about this document, feel free to send an email to Lucas Yu at pyu68@gatech.edu.

Deadline:

Sunday Aug 15, 2021 11:59 PM