Visual scene real-time analysis

Lei WANG

November 2019

This code and this report can be found on my github[1]. The weight file can be found on the google drive[2].

1 Pretrain model

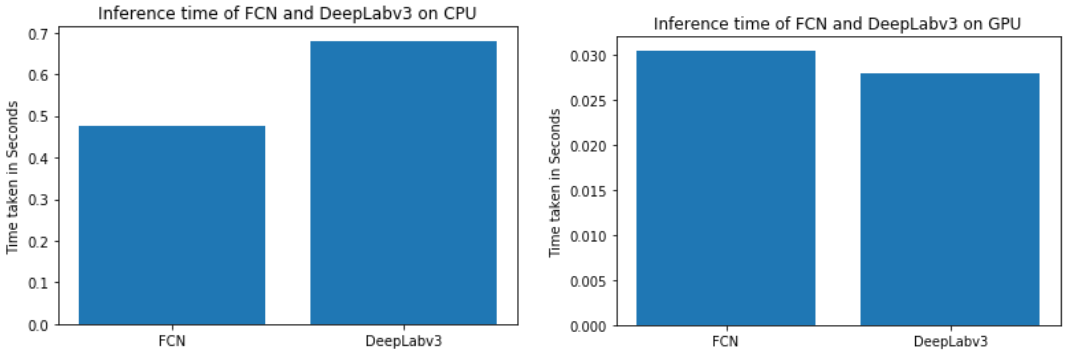
There are two models, deeplab3 and FCN, which can be imported the pretrained parameters. In these models, there are 21 default color labels.
We compare these two models from three aspects:

1. Inference time
2. Size of the model
3. GPU memory used by the model

1.1 Comparasion

1.1.1 Inference time

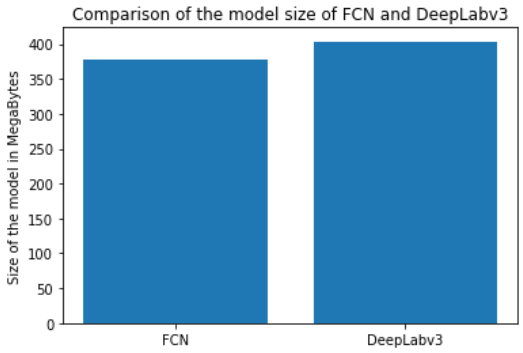
On CPU, the average inference time on FCN is 0.48 second, and the average inference time on DeepLab is 0.68 second. On GPU, the average inference time on FCN is 0.031 second, and the average inference time on DeepLab is 0.028 second.
There is a strange point. It takes longer for the FCN model on GPU, though DeepLab is much deeper than FCN.



1.1.2 Size of Model

DeepLab model is much deeper than FCN, thus it takes up much memory.

[1] https://github.com/ThorraySJTU/ParisTech-Course-Visual-scene-real-time-analysis
[2] https://drive.google.com/open?id=1P8_TZHiqwX0hr7CK47gjyY8yDRlQ7mgK



1.2 Test the pretrained model

We use the pretrained model to predict an image from a new dataset and a real image crawled from the website[3]. For these two cases, models can distinguish the car, person and others. The majority of the image is black, which is predicted as the background. However, we have to make a difference between the streets and the footpaths, etc, so these two pretrained model have to be retrained.

Dataset image Real image

2 Train model

At beginning of the train model part, the archive "Segmentation-dataset.rar" is unziped. As for CARLA images, the ground truth of this dataset is black, so the dataset is changed to "SYNTHIA_RAND_CVPR16_extract-100-images" for training our model.

[3]https://image.baidu.com/search/detail?ct=503316480z=0ipn=dword=%E9%81%93%E8%B7%AF%E5%9B%BE%E7%89%
8oe=utf-8cl=2lm=-1cs=2662937988%2C369855721os=2844611853%2C2519032861simid=4166679794%2C553890336adpicid=0lp

2.1 Encode module

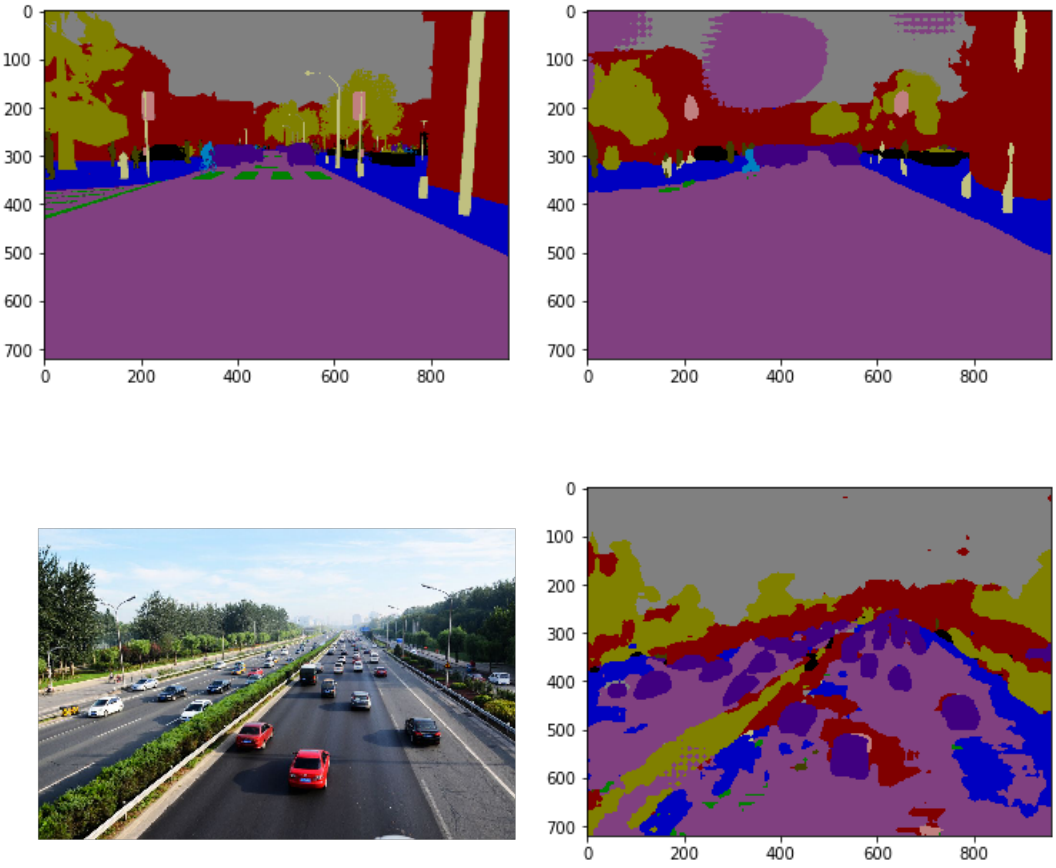
For the pretrain model, there are 21 default color labels. However, when I deal with the ground truth of new dataset, it exists 64 different colors [4]. For each pixel, they divide 64 and round up, then multiply by 64.

2.2 Model

The FCN model is chosen to predict the images. Because of the 64 color labels, the output channel of the last layer is changed to 64 channels. 64 color labels represent for 64 classes. The CrossEntropyLoss is chosen to measure the loss. After 10 epochs' training, we obtain a good model and its parameters.

3 Test model

The parameter of FCN model is loaded to predict the real image.
Training set Real image
The picture in the upper left corner is ground truth, and the picture in the upper right corner is our prediction. Except for the part of the sky, our predictions are basically correct.
From the prediction of the real image, the street can be distinguished from the footpaths. It can restore each part of the real image very well.



[4]There are 3 channels, for each channel it can be [0, 64, 128, 192]