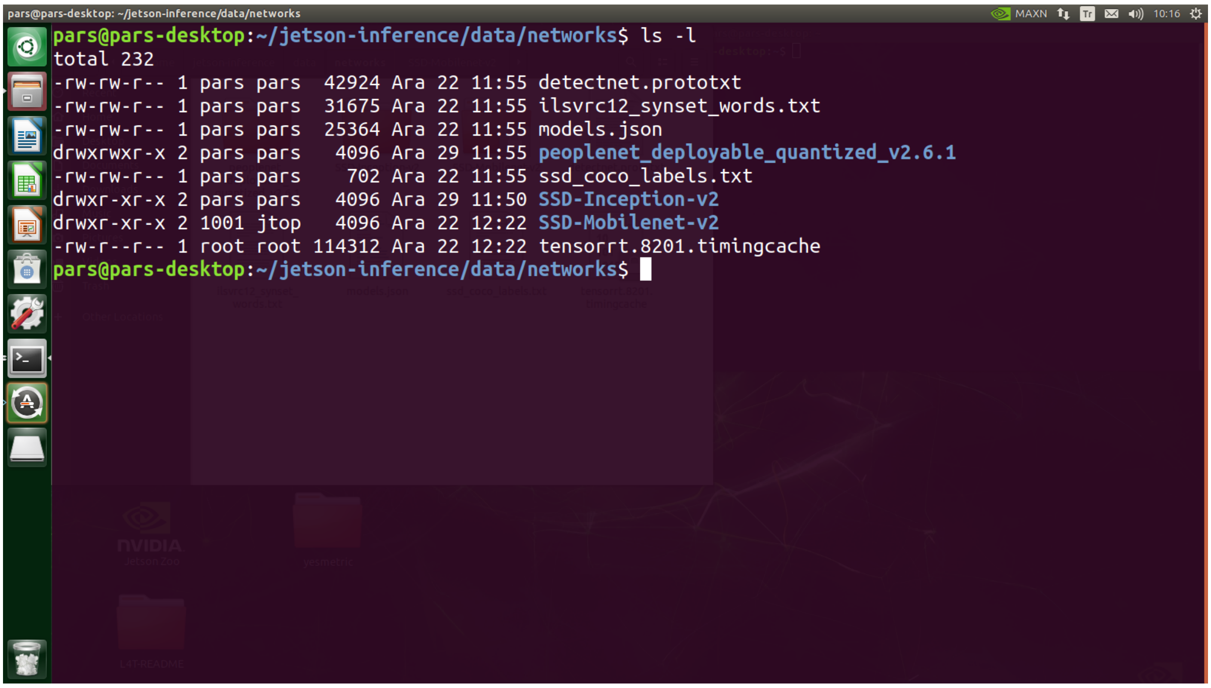
# Programming Assignment 4

Group number: 9

Group members:

- Deniz Dönmez
- Yiğit Eren Durmaz
- Gökay Gülsoy

Contents of the data/networks folder after running ssd-mobilenet-v2, ssd-inception-v2, and peoplenet networks with one sample image is as follows:



```
pars@pars-desktop:~/jetson-inference/data/networks$ ls -l
total 232
-rw-rw-r-- 1 pars pars  42924 Ara 22 11:55 detectnet.prototxt
-rw-rw-r-- 1 pars pars  31675 Ara 22 11:55 ilsvrc12_synset_words.txt
-rw-rw-r-- 1 pars pars  25364 Ara 22 11:55 models.json
drwxrwxr-x 2 pars pars   4096 Ara 29 11:55 peoplenet_deployable_quantized_v2.6.1
-rw-rw-r-- 1 pars pars    702 Ara 22 11:55 ssd_coco_labels.txt
drwxr-xr-x 2 pars pars   4096 Ara 29 11:50 SSD-Inception-v2
drwxr-xr-x 2 1001 jtop   4096 Ara 22 12:22 SSD-Mobilenet-v2
-rw-r--r-- 1 root root 114312 Ara 22 12:22 tensorrt.8201.timingcache
pars@pars-desktop:~/jetson-inference/data/networks$ 
```

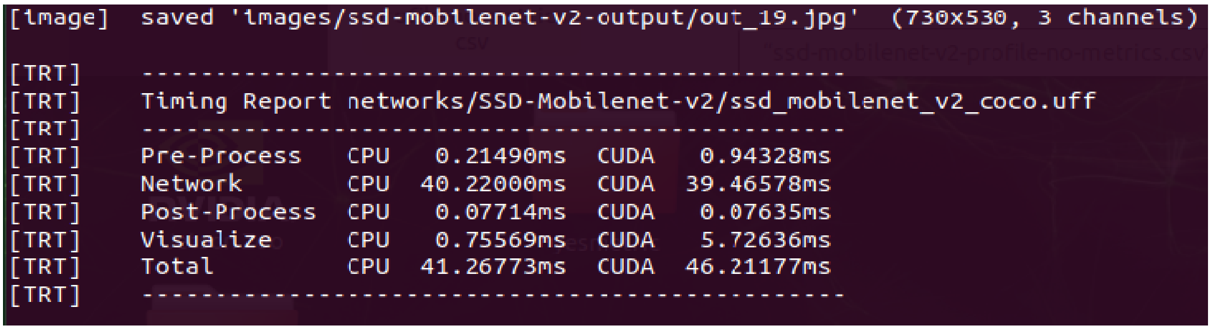
**Fig 1:** Contents of the data/networks folder after executing each network with one sample image

Profiling commands for ssd-mobilenet-v2, inference time for the last input image, and multiprocessor efficiency metric values for most and least time consuming kernels are as follows:

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/ssd-mobilnet-v2-output /ssd-moiblenet-v2-profile-no-metrics.csv ./detectnet “images/myimages/*.jpg” “images/ssd-mobilenet-v2-output_%i.jpg”

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/ssd-mobilnet-v2-output /ssd-mobilenet-v2-profile-metrics.csv ./detectnet “images/myimages/*.jpg” “images/ssd-mobilenet-v2-output_%i.jpg”

Inference time for the last input image for ssd-mobilenet-v2 is as follows:



```
[image]  saved 'images/ssd-mobilenet-v2-output/out_19.jpg'  (730x530, 3 channels)

[TRT]    ---------------------------------------------
[TRT]    Timing Report networks/SSD-Mobilenet-v2/ssd_mobilenet_v2_coco.uff
[TRT]    ---------------------------------------------
[TRT]    Pre-Process   CPU   0.21490ms  CUDA   0.94328ms
[TRT]    Network       CPU  40.22000ms  CUDA  39.46578ms
[TRT]    Post-Process  CPU   0.07714ms  CUDA   0.07635ms
[TRT]    Visualize     CPU   0.75569ms  CUDA   5.72636ms
[TRT]    Total         CPU  41.26773ms  CUDA  46.21177ms
[TRT]    ---------------------------------------------
```

**Fig 2:** Inference time for the last input image fed to ssd-mobilenet-v2 network

Multiprocessor efficiency metric values for least and most time consuming kernels is as follows:

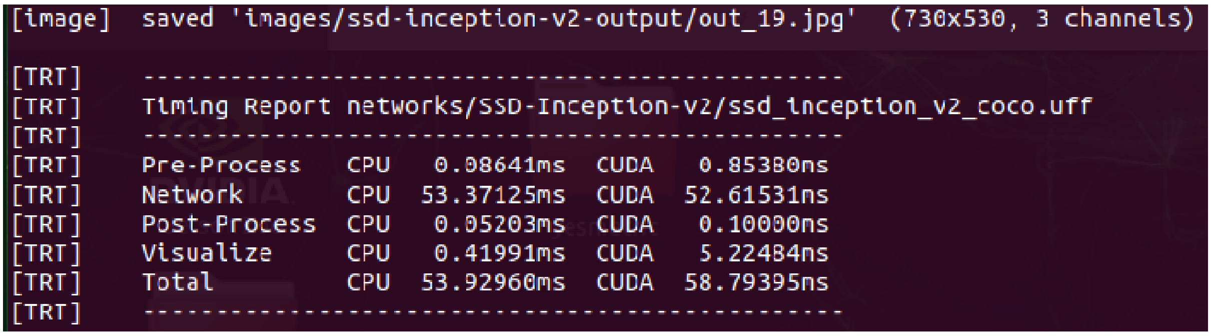
- Least time consuming kernel: setUniformOffsets => 31.531338%
- Most time consuming kernel: cuDepthWise => 98.481452%

Profiling commands for ssd-mobilenet-v2, inference time for the last input image, and multiprocessor efficiency metric values for the least and most time consuming kernels are as follows:

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/ssd-inception-v2-output /ssd-inception-v2-profile-no-metrics.csv ./detectnet “images/myimages/*.jpg” “images/ssd-inception-v2-output_%i.jpg”

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/ssd-inception-v2-output /ssd-inception-v2-profile-metrics.csv ./detectnet “images/myimages/*.jpg” “images/ssd-inception-v2-output_%i.jpg”

Inference time for the last input image for the ssd-inception-v2 is as follows:



```
[image]  saved 'images/ssd-inception-v2-output/out_19.jpg'  (730x530, 3 channels)

[TRT]    ---------------------------------------------
[TRT]    Timing Report networks/SSD-Inception-v2/ssd_inception_v2_coco.uff
[TRT]    ---------------------------------------------
[TRT]    Pre-Process   CPU   0.08641ms  CUDA   0.85380ms
[TRT]    Network       CPU  53.37125ms  CUDA  52.61531ms
[TRT]    Post-Process  CPU   0.05203ms  CUDA   0.10000ms
[TRT]    Visualize     CPU   0.41991ms  CUDA   5.22484ms
[TRT]    Total         CPU  53.92960ms  CUDA  58.79395ms
[TRT]    ---------------------------------------------
```

**Fig 3:** Inference time for the last input image fed to ssd-inception-v2 network

Multiprocessor efficiency metric values for least and most time consuming kernels is as follows:

- Least time consuming kernel: setUniformOffsets => 35.078278%
- Most time consuming kernel:

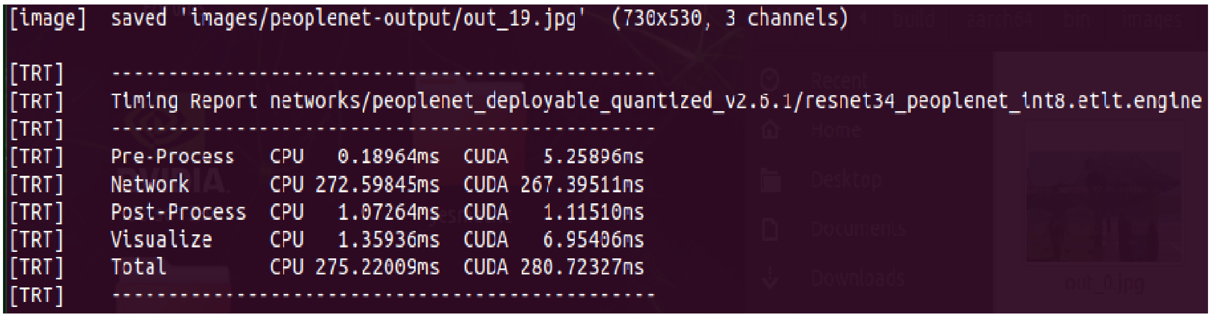
  trt_maxwell_fp16x2_hcudnn_fp16x2_128x64_relu_small_nn_v1 => 99.951964%

Profiling commands for peoplenet, inference time for the last input image, and multiprocessor efficiency metric values for the last and most time consuming kernels are as follows:

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/peoplenet-output /peoplenet-profile-no-metrics.csv ./detectnet “images/myimages/*.jpg” “images/peoplenet-output_%i.jpg”

- sudo /usr/local/bin/cuda/nvprof –csv –log-file images/peoplenet-output /peoplenet-profile-metrics.csv ./detectnet “images/myimages/*.jpg” “images/peoplenet-output_%i.jpg”

Inference time for the last input image for the peoplenet is as follows:



```
[image]  saved 'images/peoplenet-output/out_19.jpg'  (730x530, 3 channels)

[TRT]    ------------------------------------------------
[TRT]    Timing Report networks/peoplenet_deployable_quantized_v2.6.1/resnet34_peoplenet_int8.etlt.engine
[TRT]    ------------------------------------------------
[TRT]    Pre-Process   CPU   0.18964ms  CUDA   5.25896ms
[TRT]    Network       CPU 272.59845ms  CUDA 267.39511ms
[TRT]    Post-Process  CPU   1.07264ms  CUDA   1.11510ms
[TRT]    Visualize     CPU   1.35936ms  CUDA   6.95406ms
[TRT]    Total         CPU 275.22009ms  CUDA 280.72327ms
[TRT]    ------------------------------------------------
```

**Fig 4:** Inference time for the last input image fed to peoplenet network

Multiprocessor efficiency metric values for least and most time consuming kernels are as follows:

- Least time consuming kernel: generatedNativePointWise => 80.758906%
- Most time consuming kernel: trt_maxwell_fp16x2_hcudnn_winograd_fp16x2_128x128_ldg1_ldg4_relu_tile148m_nt_v1 => 99.543529%

In general for three of these network models multiprocessor efficiency of most time consuming kernels are greater than least time consuming kernels which is plausible because we expect most time consuming kernels to perform highly parallel tasks so that they are expected to utilize all the streaming multiprocessor cores nearly at maximum level. When we compare the network models among themselves highest multiprocessor efficiency for the least time consuming kernel belongs to peoplenet, which indicates that it better utilized the streaming multiprocessor cores for the least time consuming kernels compared to least time consuming kernels of ssd-mobilenet-v2 and ssd-inception-v2 network models. For the most time consuming kernels, multiprocessor efficiencies are close to maximum. Peoplenet has slightly better multiprocessor efficiency compared to ssd-mobilenet-v2 and ssd-inception-v2 according to above analysis results.