LAB5 Report

Environment setup

GPU: GeForce Titan RTX

CUDA version: 10.1

TensorRT version: 7.0.0

Baseline

Training setup:

batch size:32

Transforms function :

transforms.Compose([

transforms.RandomResizedCrop(224),

transforms.RandomHorizontalFlip(),

transforms.ToTensor(),

transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225])

Result:

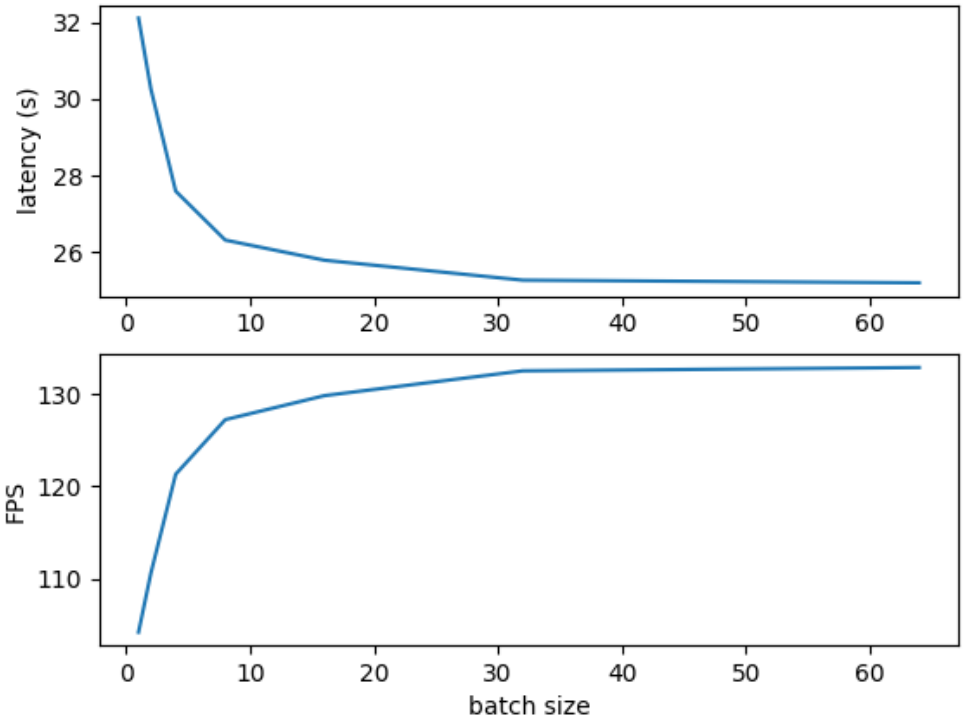
```
Test acc: 94.3830, time: 36.3968
```

Onnx:

```
torch.onnx.export(torch_model,               # model being run
                  x,                         # model input (or a tuple for multiple inputs)
                  "lab5_model.onnx",   # where to save the model (can be a file or file-like object)
                  export_params=True,        # store the trained parameter weights inside the model file
                  opset_version=10,          # the ONNX version to export the model to
                  do_constant_folding=True,  # whether to execute constant folding for optimization
                  input_names = ['input'],   # the model's input names
                  output_names = ['output'], # the model's output names
                  dynamic_axes={'input' : {0 : 'batch_size'},    # variable lenght axes
                                'output' : {0 : 'batch_size'}})
```

TensorRT:

準確率並沒有太大的改變，FPS及latency的曲線如下:



可以發現隨著batch size增加，latency減少而FPS增加。

Conclusion:

整體來說我覺得Inference time沒有減少想像中的多，如果是我會傾向在研究減少training的時間。另外tensorRT的相關資料較少，需要花比較多時間研究，