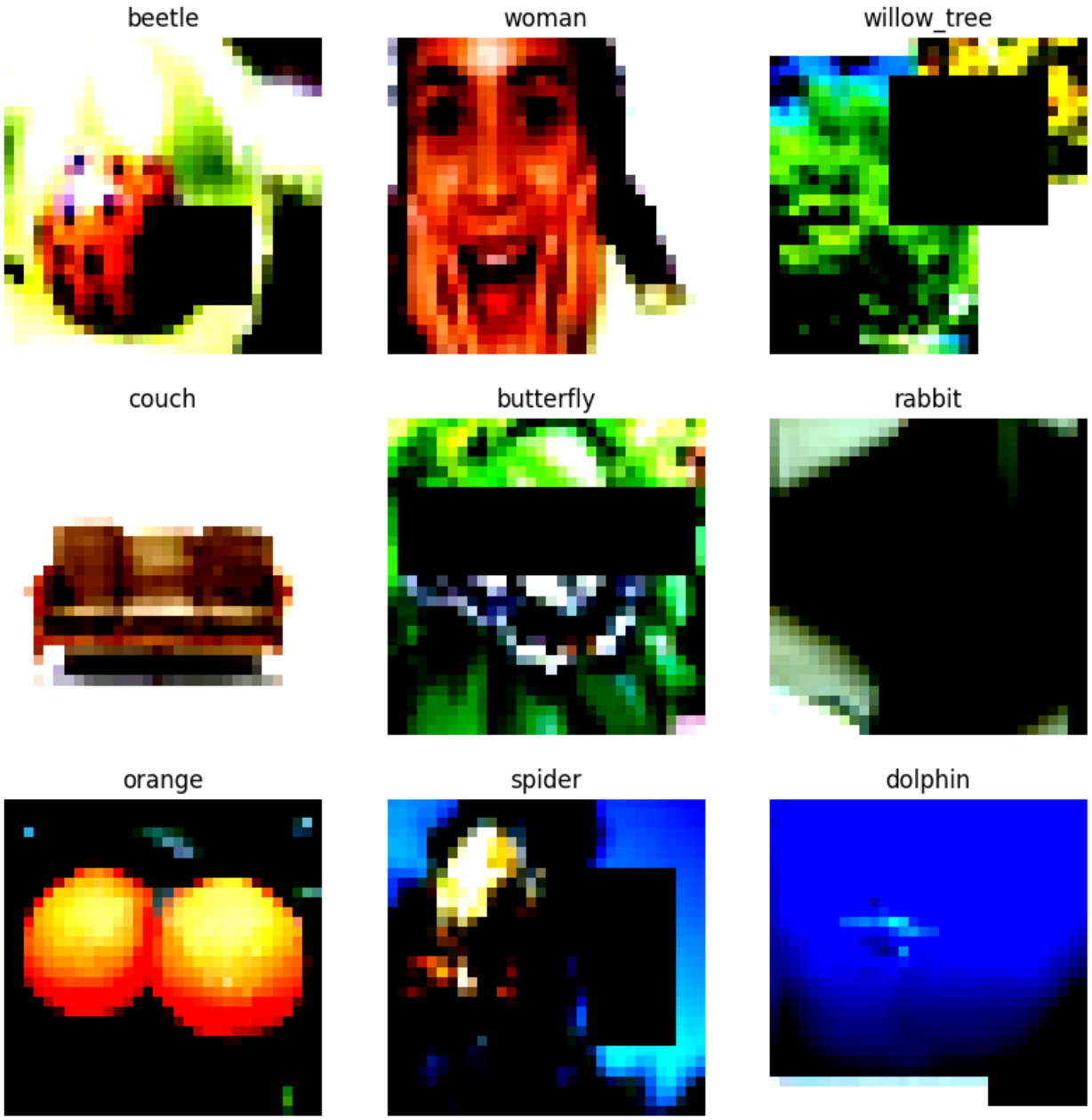
Note: Due to memory and GPU constraint model train with only 11 or 5 epochs and some point at less datasets.

Lab Assignment: 6

- Import require libraries
- Set device agnostic code
- Take cifar100 as data set and perform augmentation and normalization on it
- Visualize the data

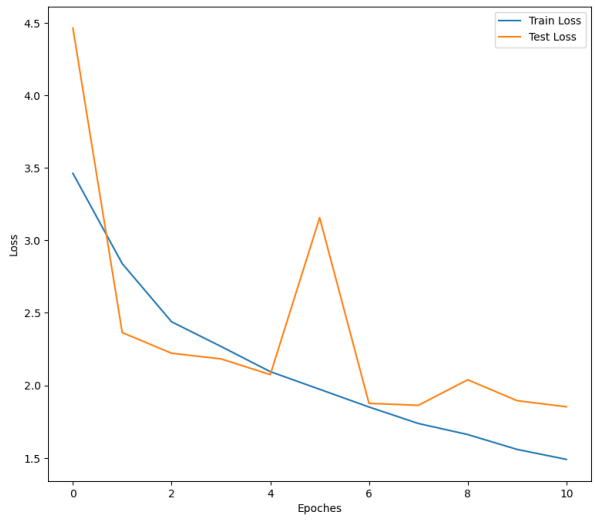


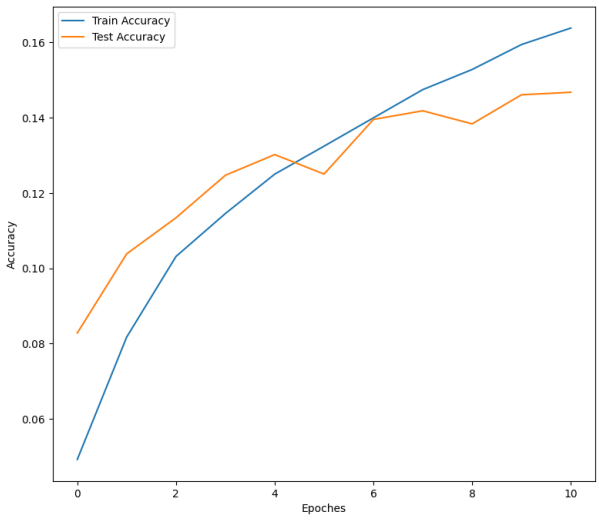
- Now train ResNet34, DenseNet121, Efficientnet-b0 and ConvneXt-Tiny and perform TorchScript, ONNX and ONNX quantization
- Result are below

ResNet34:

Normal Train:

Epoch: 1 Train Loss: 3.4614 | Test Loss: 4.4630 | Train Accuray: 0.0492 | Test Accuracy: 0.0828
Epoch: 2 Train Loss: 2.8392 | Test Loss: 2.3638 | Train Accuray: 0.0817 | Test Accuracy: 0.1038
Epoch: 3 Train Loss: 2.4381 | Test Loss: 2.2211 | Train Accuray: 0.1031 | Test Accuracy: 0.1134
Epoch: 4 Train Loss: 2.2683 | Test Loss: 2.1824 | Train Accuray: 0.1145 | Test Accuracy: 0.1247
Epoch: 5 Train Loss: 2.0944 | Test Loss: 2.0740 | Train Accuray: 0.1250 | Test Accuracy: 0.1302
Epoch: 6 Train Loss: 1.9728 | Test Loss: 3.1553 | Train Accuray: 0.1324 | Test Accuracy: 0.1250
Epoch: 7 Train Loss: 1.8507 | Test Loss: 1.8758 | Train Accuray: 0.1400 | Test Accuracy: 0.1395
Epoch: 8 Train Loss: 1.7371 | Test Loss: 1.8624 | Train Accuray: 0.1475 | Test Accuracy: 0.1418
Epoch: 9 Train Loss: 1.6610 | Test Loss: 2.0385 | Train Accuray: 0.1528 | Test Accuracy: 0.1384
Epoch: 10 Train Loss: 1.5584 | Test Loss: 1.8944 | Train Accuray: 0.1594 | Test Accuracy: 0.1461
Epoch: 11 Train Loss: 1.4895 | Test Loss: 1.8521 | Train Accuray: 0.1638 | Test Accuracy: 0.1468

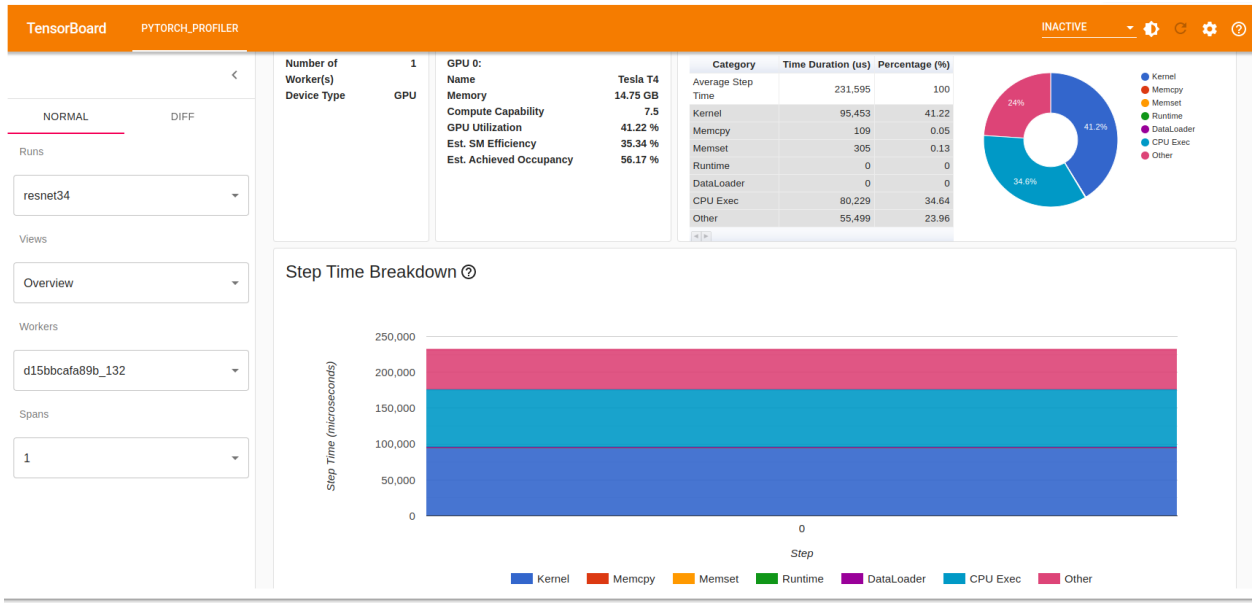




Torch Script average time:

```
Torch ResNet34: 5885.804243818179
TorchScript ResNet34: 6293.558985000036
FrozenScript ResNet34: 7047.93822900001
```

Tensor Board



ONNX and ONNX quantization

ONNX

```
AssertionError:
Not equal to tolerance rtol=0.001, atol=1e-05

Mismatched elements: 26 / 1000000 (0.0026%)
Max absolute difference: 0.00061035
Max relative difference: 2.4972825
  x:  array([[-11.016734,    -7.357694,    -2.490472,  ...,    -5.484066,
-2.787735,
        -3.247329],
            [  -8.735667,    -6.352623,    -4.557669,  ...,     0.297015,
-3.753641,...
  y:  array([[-11.016723,    -7.357697,    -2.490463,  ...,    -5.484061,
-2.787733,
        -3.247326],
            [  -8.735658,    -6.352619,    -4.55767 ,  ...,     0.29702 ,
-3.753641,...
```

ONNX quantization

Size

```
Original model size (MB): 81.52575778961182
ONNX full precision model size (MB): 81.37440967559814
ONNX quantized model size (MB): 20.43724822998047
```

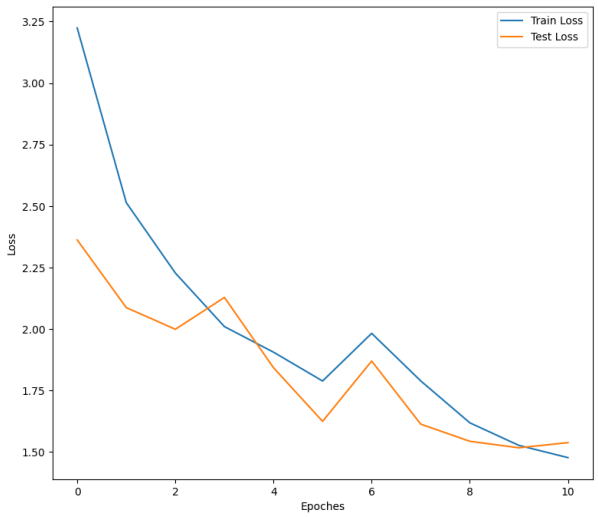
Time

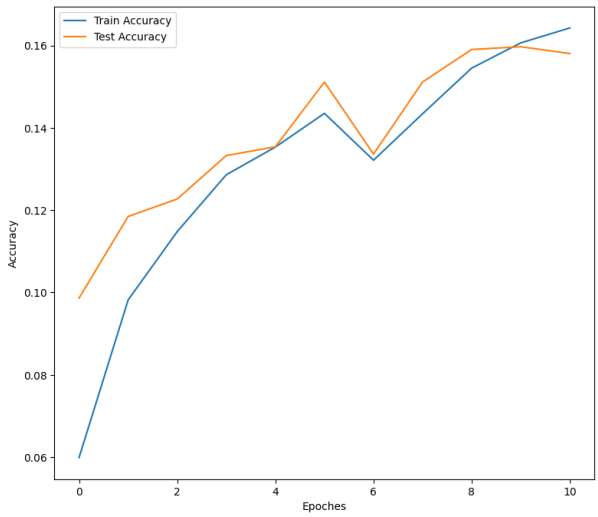
```
Average runtime of ONNX Model in GPU: 29573.097154399868
Average runtime of ONNX Quantized Model in GPU: 97380.8883776
```

DenseNet121:

Normal Train:

Epoch: 1 Train Loss: 3.2242 | Test Loss: 2.3619 | Train Accuray: 0.0599 | Test Accuracy: 0.0986
Epoch: 2 Train Loss: 2.5137 | Test Loss: 2.0867 | Train Accuray: 0.0982 | Test Accuracy: 0.1185
Epoch: 3 Train Loss: 2.2273 | Test Loss: 1.9985 | Train Accuray: 0.1148 | Test Accuracy: 0.1227
Epoch: 4 Train Loss: 2.0095 | Test Loss: 2.1280 | Train Accuray: 0.1286 | Test Accuracy: 0.1333
Epoch: 5 Train Loss: 1.9054 | Test Loss: 1.8421 | Train Accuray: 0.1353 | Test Accuracy: 0.1354
Epoch: 6 Train Loss: 1.7882 | Test Loss: 1.6241 | Train Accuray: 0.1435 | Test Accuracy: 0.1511
Epoch: 7 Train Loss: 1.9820 | Test Loss: 1.8694 | Train Accuray: 0.1321 | Test Accuracy: 0.1336
Epoch: 8 Train Loss: 1.7886 | Test Loss: 1.6127 | Train Accuray: 0.1434 | Test Accuracy: 0.1511
Epoch: 9 Train Loss: 1.6181 | Test Loss: 1.5430 | Train Accuray: 0.1545 | Test Accuracy: 0.1590
Epoch: 10 Train Loss: 1.5263 | Test Loss: 1.5165 | Train Accuray: 0.1606 | Test Accuracy: 0.1597
Epoch: 11 Train Loss: 1.4765 | Test Loss: 1.5375 | Train Accuray: 0.1643 | Test Accuracy: 0.1581

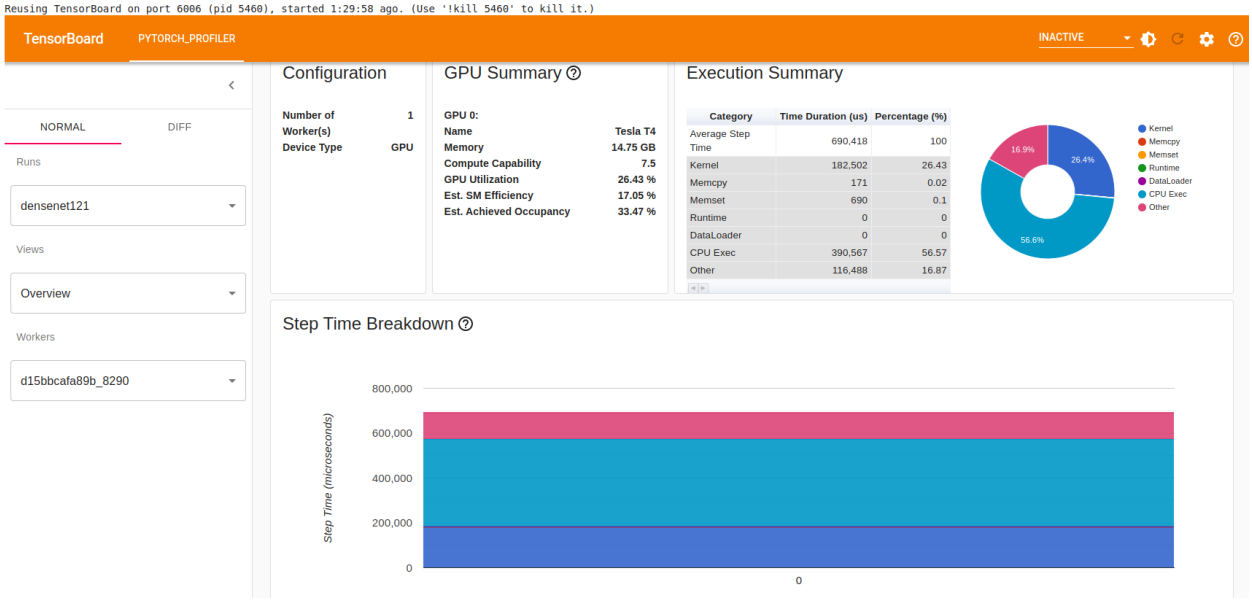




Torch Script average time:

```
Torch DenseNet121: 11922.605186818186
TorchScript DenseNet121: 20971.345630818258
FrozenScript DenseNet121: 9594.243977181659
```

Tensor Board



ONNX and ONNX quantization

ONNX

```
No AssertionError
```

ONNX quantization

Size

```
Original model size (MB): 27.503769874572754
ONNX full precision model size (MB): 27.292320251464844
ONNX quantized model size (MB): 7.562499046325684
```

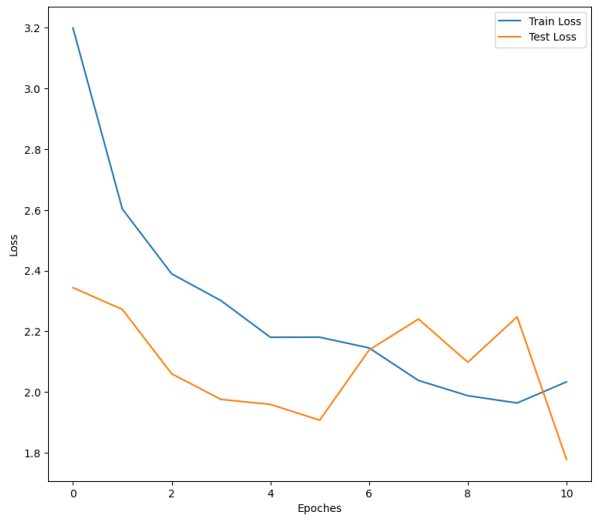
Time

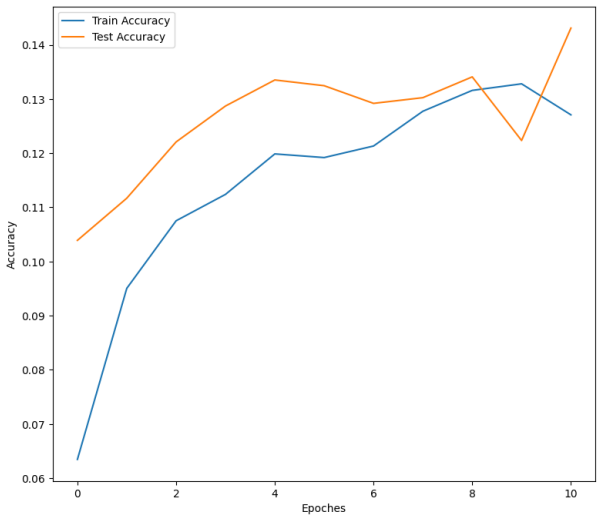
```
Average runtime of ONNX Model in GPU: 2419.164396899987
Average runtime of ONNX Quantized Model in GPU: 7287.094726399982
```

EfficientNet-b0:

Normal Train:

Epoch: 1 Train Loss: 3.1993 | Test Loss: 2.3436 | Train Accuray: 0.0635 | Test Accuracy: 0.1039
Epoch: 2 Train Loss: 2.6031 | Test Loss: 2.2720 | Train Accuray: 0.0950 | Test Accuracy: 0.1117
Epoch: 3 Train Loss: 2.3892 | Test Loss: 2.0598 | Train Accuray: 0.1075 | Test Accuracy: 0.1221
Epoch: 4 Train Loss: 2.3008 | Test Loss: 1.9753 | Train Accuray: 0.1124 | Test Accuracy: 0.1287
Epoch: 5 Train Loss: 2.1802 | Test Loss: 1.9590 | Train Accuray: 0.1199 | Test Accuracy: 0.1335
Epoch: 6 Train Loss: 2.1803 | Test Loss: 1.9069 | Train Accuray: 0.1192 | Test Accuracy: 0.1325
Epoch: 7 Train Loss: 2.1450 | Test Loss: 2.1380 | Train Accuray: 0.1213 | Test Accuracy: 0.1292
Epoch: 8 Train Loss: 2.0378 | Test Loss: 2.2402 | Train Accuray: 0.1277 | Test Accuracy: 0.1302
Epoch: 9 Train Loss: 1.9876 | Test Loss: 2.0980 | Train Accuray: 0.1316 | Test Accuracy: 0.1341
Epoch: 10 Train Loss: 1.9636 | Test Loss: 2.2475 | Train Accuray: 0.1328 | Test Accuracy: 0.1223
Epoch: 11 Train Loss: 2.0331 | Test Loss: 1.7776 | Train Accuray: 0.1271 | Test Accuracy: 0.1431

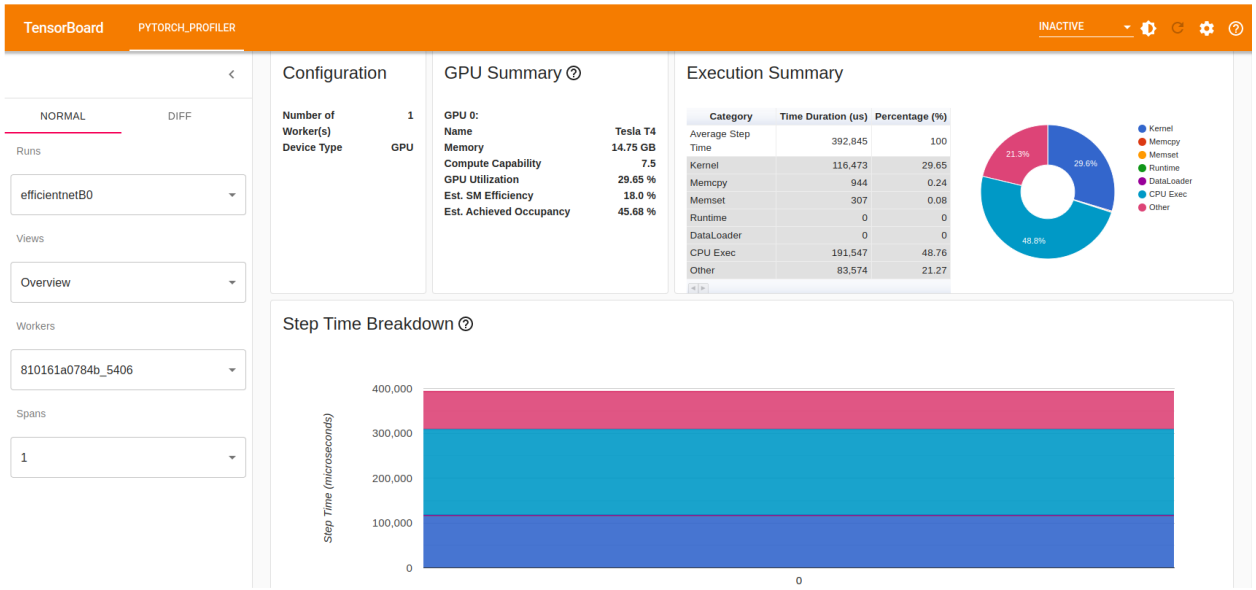




Torch Script average time:

```
Torch EfficientnetB0: 7603.004892909088
TorchScript EfficientnetB0: 9949.372756818071
FrozenScript EfficientnetB0: 5565.760808636407
```

Tensor Board



ONNX and ONNX quantization

ONNX

```
AssertionError:
Not equal to tolerance rtol=0.001, atol=1e-05

Mismatched elements: 100000 / 100000 (100%)
Max absolute difference: 361.20773
Max relative difference: 25111.96
 x: array([[-2.349335, -0.478157,  2.092996, ...,  2.045079,  3.711552,
         2.051082],
       [ 3.434702,  3.736609,  6.950093, ...,  8.285063,  8.208593,...
  y:  array([[-17.744719,  -13.114738,   -4.648403,  ...,   -8.932738,
-4.651407,
         -8.13747 ],
            [  -8.682911,   -8.194717,   -7.174371,  ...,   -1.421207,
-6.708521,...
```

ONNX quantization

Size

```
Original model size (MB): 16.066096305847168
ONNX full precision model size (MB): 27.292320251464844
ONNX quantized model size (MB): 7.562499046325684
```

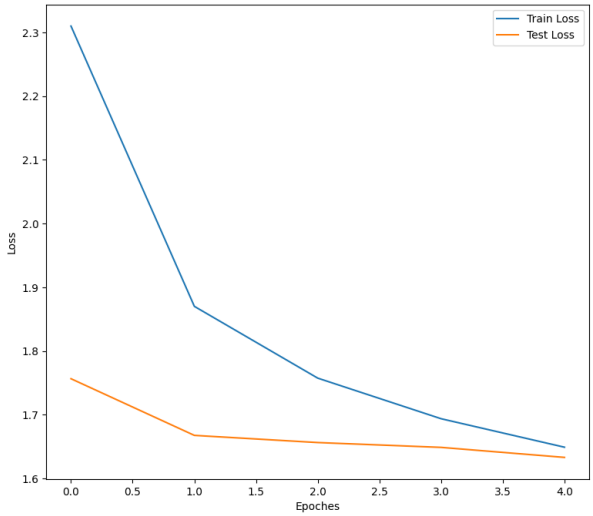
Time

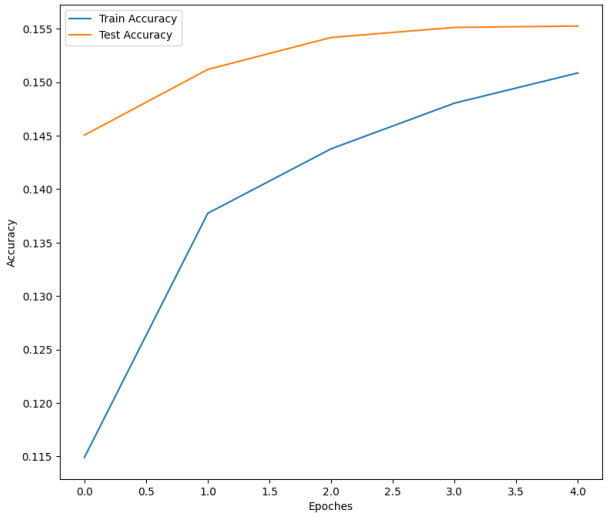
```
Average runtime of ONNX Model in GPU: 2526.3720583999657
Average runtime of ONNX Quantized Model in GPU: 7139.441247900095
```

ConvneXt-Tiny:

Normal Train:

Epoch: 1 Train Loss: 2.3097 | Test Loss: 1.7565 | Train Accuray: 0.1149 | Test Accuracy: 0.1451
Epoch: 2 Train Loss: 1.8700 | Test Loss: 1.6676 | Train Accuray: 0.1377 | Test Accuracy: 0.1512
Epoch: 3 Train Loss: 1.7574 | Test Loss: 1.6564 | Train Accuray: 0.1438 | Test Accuracy: 0.1542
Epoch: 4 Train Loss: 1.6938 | Test Loss: 1.6488 | Train Accuray: 0.1480 | Test Accuracy: 0.1551
Epoch: 5 Train Loss: 1.6491 | Test Loss: 1.6331 | Train Accuray: 0.1509 | Test Accuracy: 0.1553

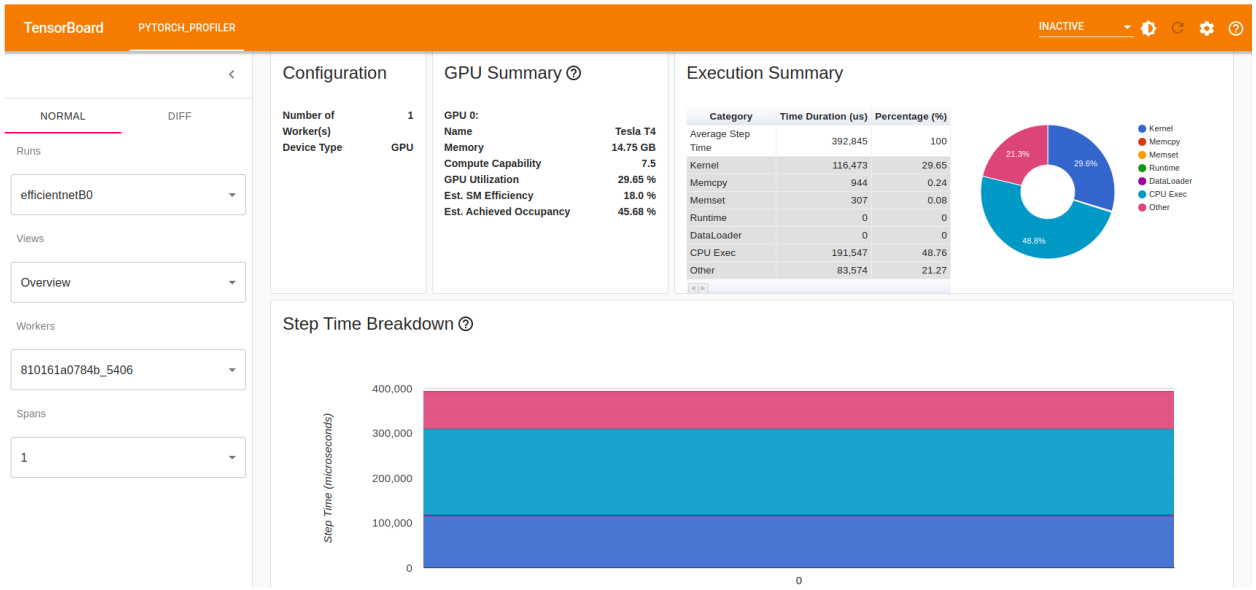




Torch Script average time:

```
Torch Convnext_tiny: 5711.639052400005
TorchScript Convnext_tiny: 5235.740192000003
FrozenScript Convnext_tiny: 7068.381913000007
```

Tensor Board



ONNX and ONNX quantization

ONNX

```
No AssertionError
```

ONNX quantization

Size

```
Original model size (MB): 106.47789859771729
ONNX full precision model size (MB): 106.51996421813965
ONNX quantized model size (MB): 27.01221752166748
```

Time

```
Average runtime of ONNX Model in GPU: 3369.5091919999977
Average runtime of ONNX Quantized Model in GPU: 14633.848032500053
```