ECE 528 Homework Assignment 3

Model Compression

1) a) The baseline network model built for the MNIST American Sign Language dataset gives a test accuracy of 92.512%

```
In [ ]:  test_loss, test_acc = model.evaluate(x_test,  y_test, verbose=2)
         print('\nTest accuracy:', test_acc)
```

```
225/225 - 1s - loss: 0.5518 - accuracy: 0.9251

Test accuracy: 0.9251254796981812
```

b) The baseline keras model built for the MNIST American Sign Language dataset gives a test accuracy of 94.562184%

```
In [ ]:  test_loss, test_acc = model.evaluate(x_test,  y_test, verbose=2)
         print('\nTest accuracy:', test_acc)
```

```
225/225 - 1s - loss: 0.5006 - accuracy: 0.9456

Test accuracy: 0.9456218481063843
```

The test accuracy of the dynamically quantized keras model for the MNIST American Sign Language dataset is 94.562186%

```
In [ ]:  print(evaluate_model(interpreter_quant))
```

```
0.9456218627997769
```

The baseline keras model and the dynamically quantized keras model sizes in the order of bytes are displayed below:

```
In [ ]:  print("Size of gzipped baseline Keras model: %.2f bytes" % (get_gzipped_model_size(tflite_model_file)))
         print("Size of gzipped dynamically quantized Keras model: %.2f bytes" % (get_gzipped_model_size(tflite_model_quant_file)))
```

```
Size of gzipped baseline Keras model: 48938774.00 bytes
Size of gzipped dynamically quantized Keras model: 9197623.00 bytes
```

c) The baseline test accuracy for the baseline network is 93.36%, whereas the test accuracy post quantization in the network model is less than 5%.

```
test_loss, test_acc = model.evaluate(x_test,  y_test, verbose=2)
print('\nTest accuracy:', test_acc)
```

```
225/225 - 1s - loss: 0.4053 - accuracy: 0.9336

Test accuracy: 0.9336307644844055
```

d) The baseline test accuracy for the baseline network is 92.4%, whereas the test accuracy of the quantization-aware training model is 88.26%. Hence, there is drop in the test accuracy rate post quantization-aware training.

```
print('Baseline test accuracy:', baseline_model_accuracy)
print('Quant test accuracy:', q_aware_model_accuracy)
```

```
Baseline test accuracy: 0.9240100383758545
Quant test accuracy: 0.8825989961624146
```

A model conversion from TensorFlow to TensorFlow Lite has been made in the code that results in the following accuracy rates:

i) TensorFlow quantization-aware training model accuracy = 88.26%

ii) TensorFlow Lite quantization-aware training model accuracy = 88.3%

```
print('Quant TFLite test_accuracy:', test_accuracy)
print('Quant TF test accuracy:', q_aware_model_accuracy)
```

```
Evaluated on 0 results so far.
Evaluated on 1000 results so far.
Evaluated on 2000 results so far.
Evaluated on 3000 results so far.
Evaluated on 4000 results so far.
Evaluated on 5000 results so far.
Evaluated on 6000 results so far.
Evaluated on 7000 results so far.


Quant TFLite test_accuracy: 0.8830172894590073
Quant TF test accuracy: 0.8825989961624146
```

2) The different values of accuracy rates and the model sizes in the order of bytes on using Clustering or Weight sharing model compression techniques in the MNIST American Sign Language dataset are displayed below:

```
Size of gzipped baseline Keras model: 48779626.00 bytes
Size of gzipped clustered and quantized TFlite model: 2445806.00 bytes
```

```
Size of gzipped baseline Keras model: 48779626.00 bytes
Size of gzipped clustered Keras model: 3192896.00 bytes
Size of gzipped clustered TFlite model: 3317693.00 bytes
```

```
print('Baseline test accuracy:', baseline_model_accuracy)
print('Clustered test accuracy:', clustered_model_accuracy)
```

```
Baseline test accuracy: 0.9445064067840576
Clustered test accuracy: 0.945343017578125
```

```
print('Clustered and quantized TFLite test_accuracy:', test_accuracy)
print('Clustered TF test accuracy:', clustered_model_accuracy)
```

```
Evaluated on 0 results so far.
Evaluated on 1000 results so far.
Evaluated on 2000 results so far.
Evaluated on 3000 results so far.
Evaluated on 4000 results so far.
Evaluated on 5000 results so far.
Evaluated on 6000 results so far.
Evaluated on 7000 results so far.


Clustered and quantized TFLite test_accuracy: 0.9452035694366983
Clustered TF test accuracy: 0.945343017578125
```

3) The different values of accuracy rates and the model sizes in the order of bytes on using weight pruning model compression technique in the MNIST American Sign Language dataset are displayed below:

```
print('Pruned and quantized TFLite test_accuracy:', test_accuracy)
print('Pruned TF test accuracy:', model_for_pruning_accuracy)
```

```
Evaluated on 0 results so far.
Evaluated on 1000 results so far.
Evaluated on 2000 results so far.
Evaluated on 3000 results so far.
Evaluated on 4000 results so far.
Evaluated on 5000 results so far.
Evaluated on 6000 results so far.
Evaluated on 7000 results so far.


Pruned and quantized TFLite test_accuracy: 0.935164528722811
Pruned TF test accuracy: 0.9355828166007996

Size of gzipped baseline Keras model: 48749076.00 bytes
Size of gzipped pruned Keras model: 15641374.00 bytes
Size of gzipped pruned TFlite model: 15617293.00 bytes
```