Quantization aware training

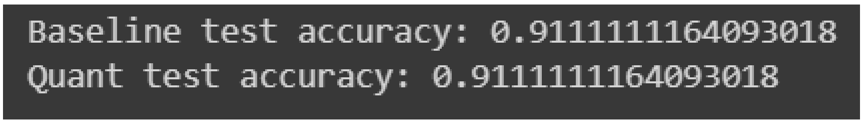
The experiments thus conducted proved one thing, either implementing the same model in the board or converting a trained model to a tf-lite format heavily affects the performance. This can be attributed to the fact that during quantization there is a lot of internal compression and loss of precision. Because of that, the model performs poorly when implemented on board. The drop in performance was as drastic as a 20+% drop in accuracy on the board. Quantization-aware training was implemented expecting no drop or lesser drop in performance while executing the model on the Sony Spresense board.

Quantization aware training was implemented using the high-level library provided by TensorFlow. The main steps involved in this process are as follows

1. Define the model
2. Fit the training data to the model
3. Train the model using the training dataset and validate using the validation dataset
4. Using tensorflow_model_optimization, quantize the trained model
5. Recompile the quantized model
6. Train and evaluate the quantized model
7. The resulting quantized model can then be converted to the tf-lite model, which can be deployed in the Sony Spresense board

The expected performance will be either there will be no drop in the baseline model and the quantization aware training model (idea scenario) or there will be a minor drop in the accuracy of the quantization aware training model.

In our case



```
Baseline test accuracy: 0.9111111164093018
Quant test accuracy: 0.9111111164093018
```

Both the baseline and quant models give the same accuracy.

Experimenting with a new dataset

Created a new dataset merging all the available ppm files. Used RGB, LAB, and HSV color spaces and extracted relevant statistical values. The data collection process spanned for 4 days hence logically added 4 classes to the dataset.

Without normalization, the model gave a 43% Accuracy trained over 65 epochs

After normalization, the model gave a 95% Accuracy over 65 epochs

After quantization-aware training

The normalized model on quantization aware training and converting to tf-lite model gave 96%. This was suspected because of class imbalance. Using oversampling the class imbalance was fixed and the resultant model gave an accuracy of 87.5%.