Q1:

Some things that I do want to note prior to going into further detail on the question problems is that it's interesting that the Base model can have a varying degree of accuracy even though being trained on the exact same data set, and using the exact same model. I don’t know why this is but for this reason in my table containing the accuracy and size of each model I included the Base model accuracy to have a more accurate representation of the accuracy degradation from each technique.

When converting the model to a TFlite version the first thing that we can note from the conversion is that the size of the sign\_mnist.tflite file is significantly smaller than the base model seeing about a 3x size in compression while having no accuracy degradation. Now comparing the sign\_mnist\_quant\_dyn.tflite model we can see that adding quantization helps further compress the model. Compared to the base model we have now reached a compression ratio of 11.69 Finding a nearly 4x compression ratio in comparison to the regular Tflit model. Interestingly enough we don’t really face too much of a loss in accuracy, only losing 0.014% accuracy. Showing that the concept of quantization can drastically improve the size of the model without facing too many issues of accuracy degradation

Incorporating the 16 integer value and weights to 8 bit integer quantization model now we see that the quantization model can actually have the same accuracy as the base model while achieving nearly the same level of compression. In comparison to dynamically quantized models the 16 bit int values and weights to 8 bit int values sees an increase in size of 1.00064516129 in order to reach the same accuracy as the base model.

Interestingly enough when using quant aware training I found this to be the least impressive of the quantization models. While it is able to reach the same level of compression as the 18x8 model it does not reach the same level of accuracy. Notably this is also the model that took the longest to train as well. It also faced the largest percent decrease of accuracy compared to the other models.

| Model | Size | Accuracy |
| --- | --- | --- |
| Base | 145 MB | 0.93809-.956079 |
| sign\_mnist  .tflite | 49,548 Kb  49.548 Mb | ***0.95607***  0.95607 |
| sign  \_mnist  \_quant\_dyn.tflite | 12,400 Kb  12.4 Mb | ***0.95607***  0.95593 |
| sign\_mnist\_quant\_int16x8.tflite | 12,408 Kb  12.408 Mb | ***0.93837***  0.93837 |
| sign\_mnist  \_quant\_aware\_training  .tflite | 12,408 Kb  12.408 Mb | ***0.94562***  0.93488 |

Q2.

Clustering also gave a significantly smaller model than the base keras model in this case we were able to see that simply clustering the model and converting it into a tflite model we are able to decrease the overall size of the model by a factor of 9.767. However something that would be important to note here is that the model did take an overall hit to the accuracy decreasing the accuracy by 0.8%.

Looking at the clustered Tflite model that we also quantize we see that we can further extract more compression without having an overall compression ratio of 13x. Interestingly enough once the model was quantized it did gain some of its accuracy back, while negligible I imagine that using the 16x8 quantization instead would result in a higher accuracy as it did in comparison to dynamic quantization.

| Model | Size | Accuracy |
| --- | --- | --- |
| Base | 46.98 Mb | 0.9555 |
| Clustered TF | 4.81 Mb | ***0.9555***  0.9475 |
| Clustered and Quantized TFlite | 3.59 Mb | ***0.9555***  0.9478 |

Q3.

When creating the Pruned tflit model we see that compression ratio is 3.13x and from there We can see the accuracy actually drops quite a bit. Taking an accuracy reduction of 2.37% which is the greatest largest decrease in accuracy both relatively and in magnitude compared to any of the previous compression methods used in this assignment.

When we then Quantize the model after pruning and Convert it into a tflite model we see that the compression rate is now significantly better. We are able to achieve a total compression rate of 12 times. Though it is important to note that the accuracy was further negatively impacted. However something I believe to be extremely important in relation to the compression of the model is the speed of the model taking to run. When running to test for accuracy of the other models each on average took about 25 minutes to complete. Pruning the model seemed to drastically increase the speed of the model when running, cutting down the accuracy validation to about 6 minutes.

| Model | Size | Accuracy |
| --- | --- | --- |
| Base | 47.02 Mb | 0.9538 |
| Pruned TFlite | 15.01 Mb | ***0.9538***  0.9301 |
| Pruned and Quantized TFlite | 3.91 Mb | ***0.9538***  0.9293 |