A article for Quantization of Tensorflow Model

Ritu Verma

June 8, 2021

1 What is Quantization and how it helps in converting the model into a lite weight model?

Quantization is a conversion technique that can reduce model size while also improving CPU and hardware accelerator latency, with little degradation in model accuracy.

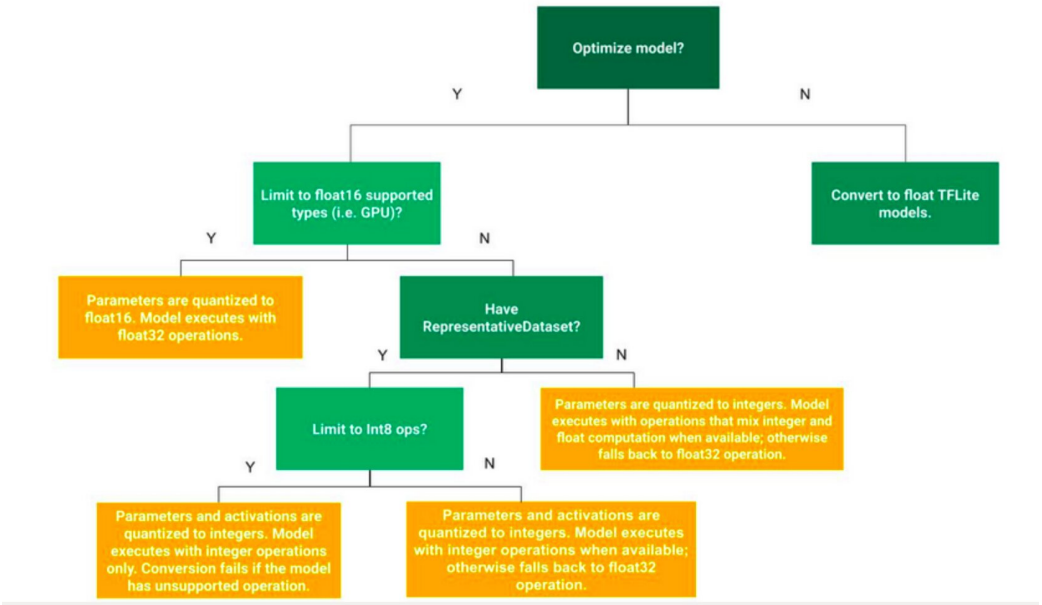

Figure 1: Quantization method use case

Quantization can be applied to a model in two flavors-
i) Post-training quantization is applied to a model after it is trained.
ii) Quantization-aware training where a model is typically trained to compensate for the loss in precision that might be introduced due to quantization. When you reduce the precision of the parameters of your model, it can result in information loss and you might see some reduction in the accuracy of your model. In these situations, quantization-aware training can be really helpful.

2 Why is it required?

Machine learning models are often bulky which often makes them inefficient for deployment in resource-constrained environments, like mobile phones, Android, iOS, Raspberry Pi's, microcontrollers and so on.
Even if you think that you might get around this problem by hosting your model on the Cloud and using an API to serve results think of constrained environments where internet bandwidths might not be always high, or where data must not leave a particular device.
We need a set of tools that make the transition to on-device machine learning seamless. In this report, I will show you how TensorFlow Lite (TF Lite) can really shine in situations like this. Well cover model optimization strategies and quantization techniques supported by TensorFlow.

3 Difference observed before and after applying Quantization?

Generally, our machine learning models operate in float32 precision format. All the model parameters are stored in this precision format, which often leads to heavier models. The heaviness of a model has a direct correlation to the speed at which the model makes predictions.
So, it might occur to you naturally that what if we could reduce the precision in which our models would operate, we could cut down on prediction times. That is what quantization does - it reduces the precision to lower forms like float16, int8, etc to represent the parameters of a model.

4 Final observation and conclusion?

Conversion of Tensorflow Model into Tensorflow Lite Model Quantization brings improvements via model compression and latency reduction. With the API defaults, the model size shrinks by 4x, and we typically see between 2 - 4x improvements in CPU latency in the tested backends. Eventually, latency improvements can be seen on compatible machine learning accelerators.

5 References:

$https://www.tensorflow.org/lite/performance/post_training_quantization$
https://en.wikipedia.org/wiki/Quantization