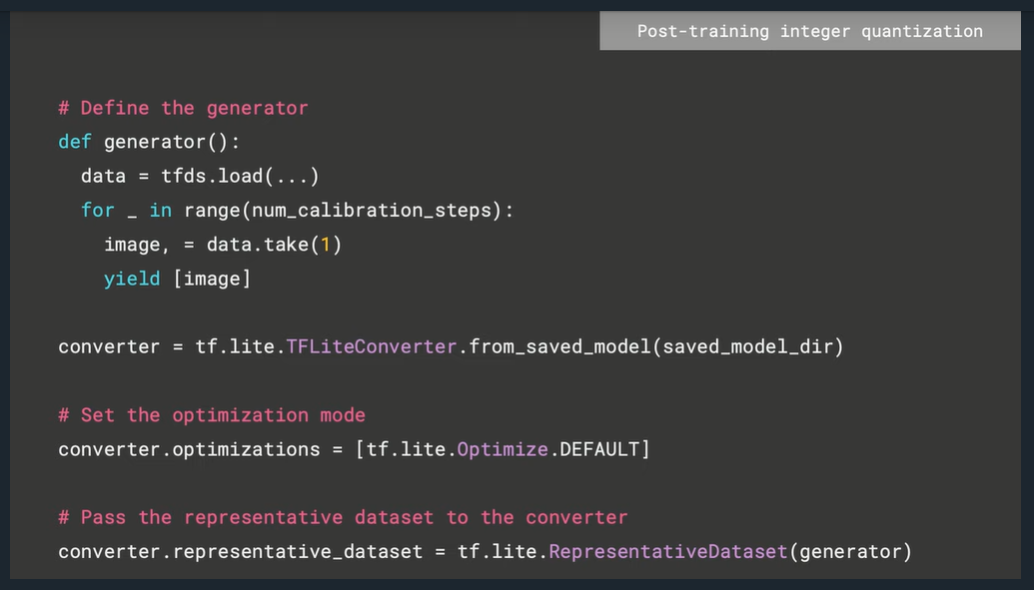
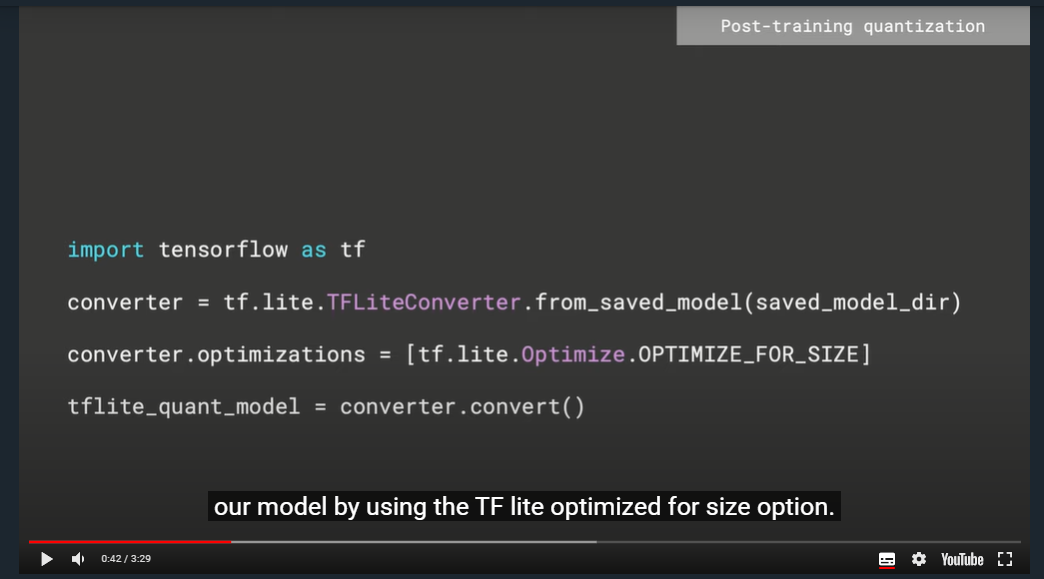
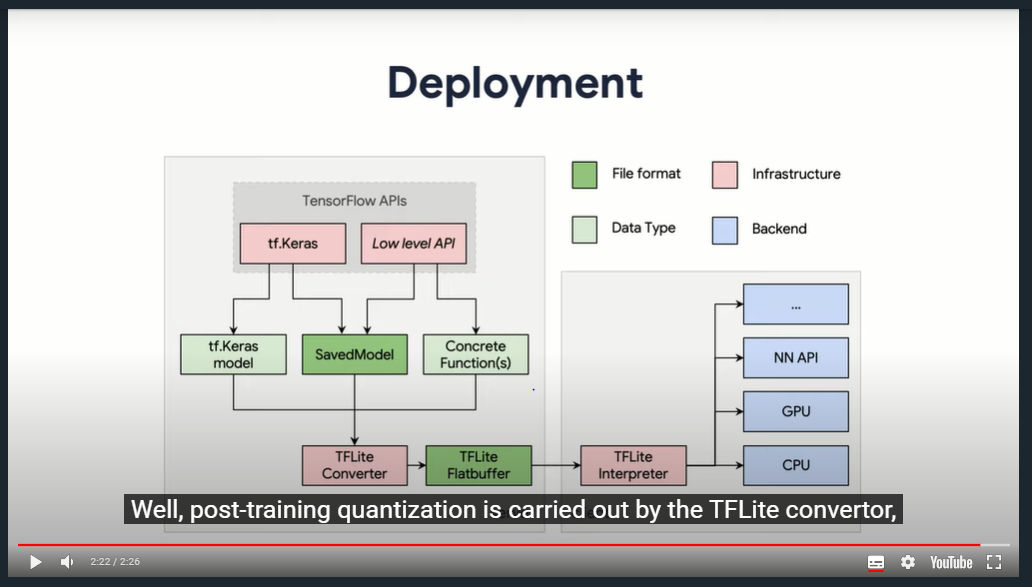
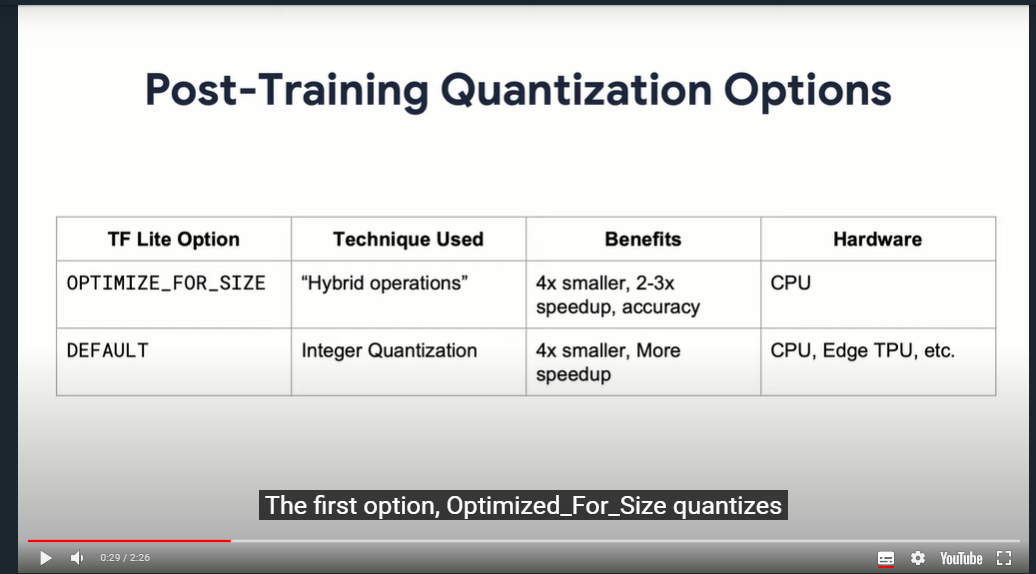
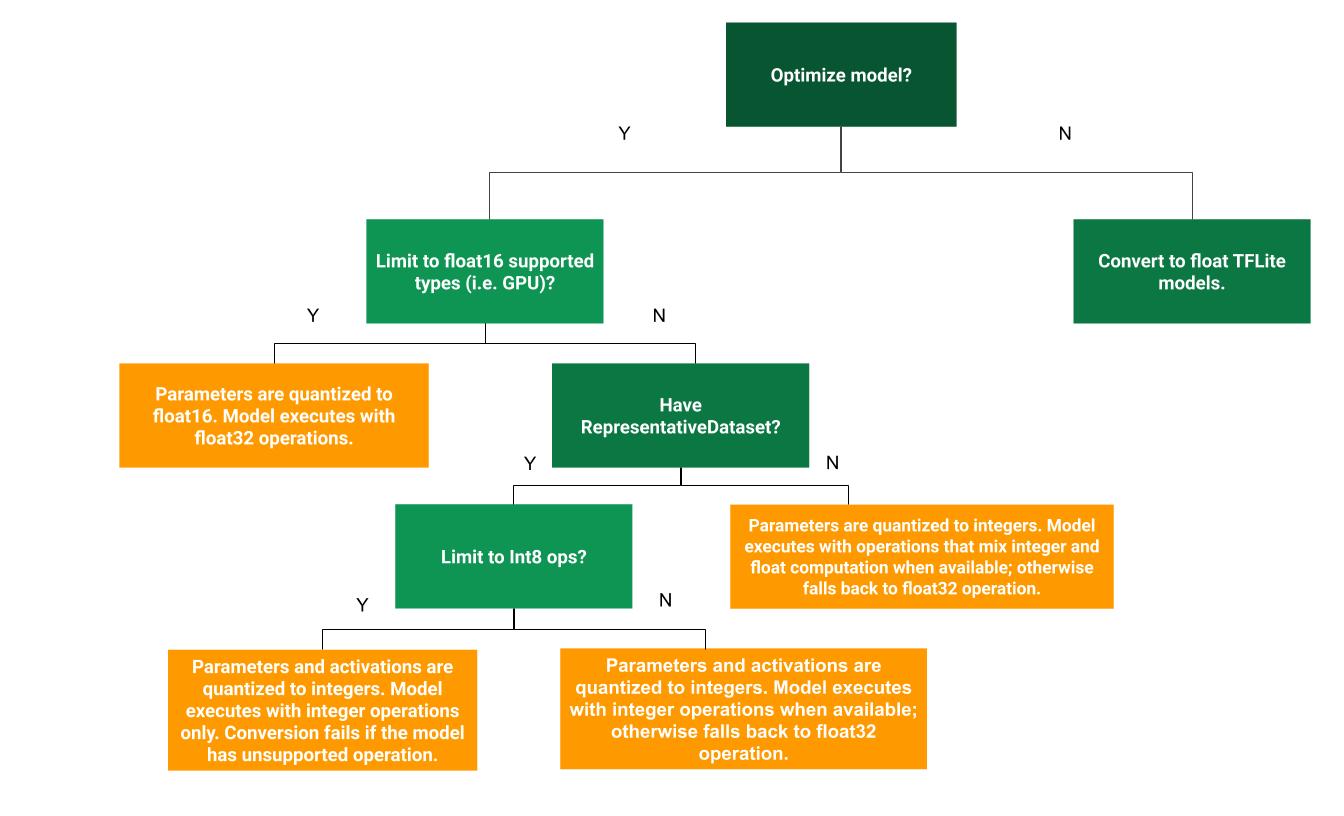
Quantization

To convert/reduce a model to run on mobile devices.



**Post-Training Decision Tree**

The decision tree depicted below can help you determine which post-training quantization method is best for your particular application



*Source: TF Lite Documentation*

If you don’t intend to quantize your model, you’ll end up with a floating point model. Also, remember that the converter will do its best to quantize all the operations (ops), but your model may still end up with a few floating point ops.

It is important to note that even though post-training quantization works really well, quantization-aware training generally results in a model with higher accuracy because it makes the model more tolerant to lower precision values. Therefore, quantization-aware training should be used in cases where the loss of accuracy brought by post-training quantization is beyond acceptable thresholds.

To learn more about Post-Training Quantization make sure to check out the [TF Lite Documentation](https://www.tensorflow.org/lite/performance/post_training_quantization).