Proposal

In this document we will find a list of proposed models, optimization techniques and platforms to launch the models.

Models

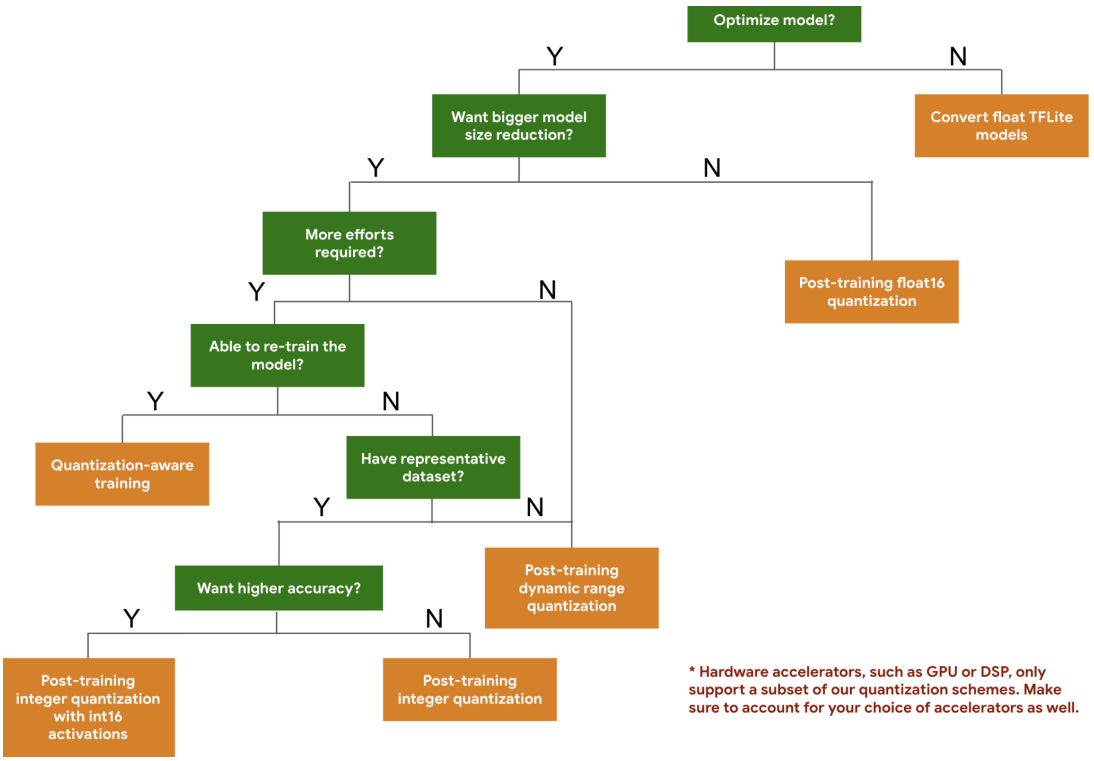
- BERT (Masked language modeling)
- GPT-2 (Text generation)
- ConvNeXT (Image classification)
- T5 (Translation)
- OWL-ViT (Object detection)
- MaskFormer (Image segmentation)
- BART (Summarization)
- MiniLM (Question answering)
- DialoGPT (Conversational)
- CodeParrot (code)

Optimization techniques

Quantization

Quantization reduces the accuracy for the numbers used in the model parameter representation. This achieves a smaller model with a faster computation.

The table below shows which type of quantization is needed in each specific case:



Pruning

Pruning removes the parameters of a model that have a smaller impact in the predictions. This just makes it so that the model is easier to compress, so it only affects the model download size.

Clustering

Clustering groups the weights on each layer in a predefined number of clusters and assigns the centroid value to all the weights in that cluster. This reduces the number of unique weights, thus reducing the model complexity.
As a result of this, the model can also compress easier, achieving similar benefits to pruning.

Platforms

- Google Cloud
- Amazon Web Services
- Microsoft Azure