CS/ECE 528: Embedded Systems and Machine Learning
Fall 2021

Homework/Lab 3: Software Model Optimization

Assigned: 30 September 2021
Due: 7 October 2021

Instructions:

- Submit your solutions via Canvas.
- Submissions should include your jupyter notebooks in a zip file, with notebooks names q1.ipynb, q2.ipynb, etc. in a single folder. You can include comments in your notebooks to explain your design choices.
- **"Save and checkpoint" your notebook after running your notebook, so that cell outputs are preserved.** If you are using Colab, make sure to 'Save' your notebook after running it, before downloading it and submitting.

Q1. (180 points) Embedded and IoT devices often have limited memory or computational power. Various optimizations can be applied to models so that they can be run within these constraints. Optimizations can potentially result in changes in model accuracy, which must be considered during the application development process. The accuracy changes depend on the individual model being optimized, and are difficult to predict ahead of time. Generally, models that are optimized for size or latency will lose a small amount of accuracy. Depending on your application, this may or may not impact your users' experience. In rare cases, certain models may gain some accuracy as a result of the optimization process. Tensorflow supports many different types of post-training quantization optimizations (https://www.tensorflow.org/lite/performance/post_training_quantization) as well as quantization during training (https://www.tensorflow.org/model_optimization/guide/quantization/training). In this question you will explore the impacts of these optimizations on model size and accuracy.

The target of the models will be the MNIST American Sign Language dataset that can be found (and downloaded from) here: https://www.kaggle.com/datamunge/sign-language-mnist. The American Sign Language letter database of hand gestures represent a multi-class problem with 24 classes of letters (excluding J and Z which require motion). It is a drop-in replacement for the conventional MNIST model but is more challenging to achieve high accuracy on.

(a) Your first task is to create a model for the MNIST American Sign Language dataset. Aim for an accuracy of at least 90% on the test images with your baseline model. Include your notebook file. Note: do not use any customized layers in your model, as these are not supported by some of the optimizations you will explore in this question. Also if you use BatchNormalization, make sure that it comes immediately before an activation layer, otherwise some of the optimizations may create issues.

(b) TensorFlow Lite supports dynamic range quantization, where it statically quantizes only the weights from 32-bit floating point to 8-bits of precision. At inference, weights are converted from 8-bits of precision to 32-bit floating point and computed using floating-point kernels. This conversion is done once and cached to reduce latency. To further improve latency, "dynamic-range" operators dynamically quantize activations (inputs) based on their range to 8-bits and perform computations with 8-bit weights and activations. For the model from (a), use the TFLiteConverter to convert your trained model into a TensorFlow Lite model, and save it as a "sign_mnist.tflite" file. Then perform post training dynamic quantization, and save the resulting file as "sign_mnist_quant_dyn.tflite". Determine the accuracy of both models on the test set, and comment on the model sizes of the files. Include your notebook and tflite files.

(c) TensorFlow Lite additionally supports converting activations to 16-bit integer values and weights to 8-bit integer values during model conversion from TensorFlow to TensorFlow Lite's flat buffer format. This mode is called the "16x8 quantization mode". It can improve accuracy of the quantized model significantly, when activations are sensitive to the quantization, while still achieving reduction in model size. Moreover, this fully quantized model can be consumed by integer-only hardware accelerators. For the model from (a), now perform 16x8 quantization and save the resulting file as "sign_mnist_quant_int16x8.tflite". Comment on the model size and accuracy, relative to the original model, as well as the dynamic quantized models. Include your notebook and tflite files.

(d) To improve accuracy over post-training quantization, Tensorflow also supports quantization-aware training. For the model from (a), now perform quantization-aware training and save the resulting file as "sign_mnist_quant_aware_training.tflite". Comment on the model size and accuracy, relative to the original model, as well as the dynamic and int16x8 quantized models. Include your notebook.

Q2. (75 points) Clustering, or weight sharing, is another model compression technique. This technique reduces the number of unique weight values in a model, leading to benefits for deployment. It first groups the weights of each layer into N clusters, then shares the cluster's centroid value for all the weights belonging to the cluster. By applying a compression algorithm to the clustered weights, it is possible to significantly reduce memory footprint. The approach can further be combined with quantization to achieve even greater improvements. For more details, refer to the tutorial here: https://www.tensorflow.org/model_optimization/guide/clustering

Starting with your baseline MNIST sign language model from Q1, use weight sharing and quantization to minimize the size of the generated Tensorflow Lite model while still maintaining at least 90% test accuracy for the model. Your score will depend on the size of your final model. You can use any of the Tensorflow Lite quantization approaches from Q1 together with weight sharing to achieve your goal.

Q3. (75 points) Another approach for model optimization involves weight pruning. Tensorflow provides a prune_low_magnitude() API to train models with removed connections. At a high level, the technique works by iteratively removing (i.e. zeroing out) connections between layers, given a schedule and a target sparsity. The Keras-based API can be applied at the level of individual layers, or the entire model. The approach can further be combined with quantization to achieve even greater improvements. For more details, refer to the tutorial here: https://www.tensorflow.org/model_optimization/guide/pruning.

Starting with your baseline MNIST sign language model from Q1, use weight pruning and quantization to minimize the size of the generated Tensorflow Lite model while still maintaining at least 90% test accuracy for the model. Your score will depend on the size of your final model. You can use any of the Tensorflow Lite quantization approaches from Q1 together with weight pruning to achieve your goal.

Q4. (2.5% course grade EXTRA CREDIT) A more aggressive form of quantization is to reduce all weights to just a few bits, e.g., ternary weights (3 values that can be stored in 2 bits). Obviously the accuracy of such models will be much lower than your baseline model, but careful quantization aware training can achieve a good compromise between model size and accuracy. Unfortunately, Tensorflow does not natively support such aggressive quantization. However, the qkeras library (https://github.com/google/qkeras) that works with Tensorflow/Keras allows exploring a much more comprehensive quantization of model layer parameters, with support for quantization aware training. Take a look at https://github.com/google/qkeras/blob/master/notebook/QKerasTutorial.ipynb and other examples on the qkeras github repo to understand how to use qkeras to explore a variety of quantization configurations. To run notebooks that

use qkeras, you will need to include the qkeras folder (from github) in the same directory as the notebook. If you want to run notebooks that use qkeras in Colab you will need to add the following to your notebooks:

```
!git clone https://github.com/google/qkeras.git
import sys
sys.path.append('qkeras')

!pip install git+https://github.com/keras-team/keras-tuner.git
!pip install tensorflow_model_optimization
```

Starting with your baseline MNIST sign language model from Q1, create a quantized model where all weights and biases in convolution and dense layers are 4 bits. You can ignore quantization of any batch normalization layers. Your goal is to achieve at least 85% accuracy on the test dataset. Qkeras provides various options for 4-bit quantization, so explore them carefully and be prepared to train for a large number of epochs to allow the model accuracy to improve. Use print_qstats() to calculate the size (in bits) of your model and contrast it with the model size of the baseline model (assuming 32-bit parameters). Your score will depend on the accuracy you are able to achieve. Include your notebook and a word/pdf file describing model size, accuracy, and comparison of sizes.