Homework #2 Report

Group #18
Eduard Ciprian Ilas s277901
Gabriele Cuni s277957
Valeria Sorrenti s276146

1. **Multi-Step Temperature and Humidity Forecasting**

   The first optimisation script aims at building, training and deploying an optimized multi-output model able to perform time series forecasting using the the Jena Climate Dataset as requested. The model takes 6 consecutive values of temperature and humidity, and outputs the next 6 values. The model of choice was a simple multilayer perceptron with two dense layers, which has been trained for 22 epochs using the ADAM optimizer. Regarding the optimization, we employed structured pruning via width scaling, using α=0.1 and so reducing the number of units to 12 per dense layer which has greatly helped with the reduction in size without overly affecting accuracy. Furthermore, a magnitude-based pruning with polynomial decay enforced final sparsity of 85% in order to further reduce size of the `tflite` model once compressed. Finally, weights-only PTQ has been applied to the trained model to store values in `int8`, further reducing the final size of the model. The resulting `tflite` model resulted in having a temperature MAE of 0.47°C, a pressure MAE of 1.75%, and a final size of 1.67KB, satisfying both versions constraints, thus the two required model files coincide.

2. **Keyword Spotting**

   The second exercise requires 3 different models for keywords spotting, each satisfying different constraints. It is to be noted that the order of the labels we trained and tested the models is $['down','right','left','up','yes','no','go','stop']$. We ended up using the exact models we had used in Lab3 and Lab 4, with different parameters for each of them, as follows:

   (a) DS-CNN model, trained for 20 epochs with MFCC features, width scaling $\alpha = 0.5$, polynomial decay with final sparsity 0.8, and weights-only PTQ with `int8` values. The resulting `tflite` model had an accuracy of 90.13% and a size of 17.29 KB once compressed.

   (b) CNN model, trained for 12 epochs with MFCC features with a learning rate of 0.01 (diminished by a factor of 25 by the 10th epoch), width scaling $\alpha = 0.35$ and weights-only PTQ with `int8` values. The resulting `tflite` model had an accuracy of 90.25%, a size of 32.56 KB once compressed and an inference latency of 0.55ms.

   (c) DS-CNN model, trained for 35 epochs with STFT features with a learning rate of 0.01, width scaling $\alpha = 0.4$ and weights-only PTQ with `int8` values. The resulting `tflite` model had an accuracy of 90.38%, a size of 33.68 KB once compressed and total latency of 20.85ms.