Machine Learning for IOT: Homework II, Group 16

1st Francesco Capobianco
Polytechnic of Turin
Turin, Italy
ID: *s281307*

2nd Pierluigi Compagnone
Polytechnic of Turin
Turin, Italy
ID: *s288301*

3rd De Cristofaro Carmine
Polytechnic of Turin
Turin, Italy
ID: *s291129*

I. Exercise I

In this exercise, we are asked to implement multi-output/multi-step models for temperature and humidity forecasting over the Jena Climate Dataset. The assignment provides some constraints on mean absolute error values and on model size which our networks must be complain with.
Starting from the baseline results, reported below, we decide to optimize the Multi Layer Perceptron model for version a and the Convolutional Neural Network for version b.

- MLP:
  - TMAE: 0.29 °C , RhMAE: 1.4 %
  - tflite Size: 74 kB
- CNN:
  - TMAE: 0.26 °C, RhMAE: 1.38 %
  - tflite Size: 70 kB

version a:

We apply a structured pruning with **width multiplier** equal to **0.03** on the MLP model in order to decrease the model size. On this way we implement a magnitude-base pruning too, whose **final sparsity** is equal to **0.4**. Moreover we use a weights only post training quantization that improve the efficiency of the computation phase and both the time and energy memory transfers.

version b:

As before, our aim is to drastically reduce the model size without loosing model precision and so we apply the same optimization steps except for the quantization that is applied on weights and activations. This time we use a **width multiplier** equal to **0.04** and a magnitude-base pruning with **final sparsity** equal to 0.25.

The results are reported in table I.

TABLE I
Results

Model	Optimization	TMAE (°C)	RhMAE (%)	Size (kB)
version a: MLP	PTQ weights structured 0.03 final sparsity 0.4	$0{,}26 < 0{,}3$	$1{,}19 < 1{,}2$	$1{,}32 < 1{,}5$
version b: CNN	PTQ w+a structured 0.04 final sparsity 0.25	$0{,}67 < 0{,}7$	$2{,}4 < 2{,}5$	$1{,}79 < 1{,}8$

II. Exercise II

The second exercise demands to train a model on original mini speech dataset for keyword spotting, a speech recognition problem that deals with the identification of keywords in utterances. As before we are asked to build the model in according to some given constraints on accuracy, model size and total latency.

We decide to implement a Depth-Wise Separable Convolutional Neural Network for all the three requested results grids, clearly changing the parameters for the Mel-frequency cepstral coefficients, computed in the preprocessing steps, and the optimization settings.

version a:

First of all we preprocess the audio signal of the dataset by means of MFCC whose parameters are:
frame lenght/step = 640/320 for the spectrogram computation, **lower/upper freq** = 20/4000 and **# mel-bin/coefficients** = 40/10.
We respect the given constraints applying, on our DS CNN, weights-only post training quantization that allows us to reduce the memory occupation by a factor of 4 since we cast the weights from float32 to int8.

TABLE II
Results

Model	Optimization	Accuracy (%)	Size (kB)	Latency (ms)
DS-CNN version a:	PTQ weights	$92{,}3 > 92$	$122{,}6 < 130$	
version b:	PTQ w+a structured 0.7 final sparsity 0.73	$93{,}1 > 91{,}5$	$49{,}55 < 50$	$36{,}65 < 40$
version c:	PTQ weights structured 0.3 final sparsity 0.35	$91{,}01 > 91$	$24{,}16 < 25$	$34.84 < 40$

version b:

For this configuration we start with a down-sampling factor of 2 and an MFCC's parameters are:
frame lenght/step = 320/160 for the spectrogram computation, **lower/upper freq** = 20/4000 and **# mel-bin/coefficients** = 16/10.
While for the optimization we apply the same pruning schedule and for the quantization we apply it on weights and activations. In table II we report the adopted numerical value of the parameters.

version c:

The applied preprocessing is the same of version b, whereas to gain a feasible trade-off between size and mae values we applied a structured and magnitude-based pruning in addiction to a weights quantization.
The summary and the results are reported in table II