Homework II - Machine Learning for IoT 2021-2022

Khudayar Farmanli *s276310*
Fabio Tatti *s282383*
Fabiana Vinci *s279918*

I. Exercise 1

A. Version a

We tried to implement all the three different models from Lab03: *MLP, CNN, LSTM*. We first noticed that both *CNN* and *LSTM* were too much big in term of parameters, which meant that the final size of the model would result not conform to the given constrain. We focused mainly on the MLP, trying to improve the Multiple Output Metric value. In the *version a* we used the $batch\_size = 512$ in the training phase in order to decrease the error. We used the $learning\_rate = 0.01$ in order to have a proper gradient descent, since for bigger values could not converge, while for smaller converged in local minimum. We trained the model with $n\_epochs\_training = 100$ to guarantee the convergence. Respect to the original model we choosed to remove the central fully connected layer, in order to decrease the number of parameters of the final resulting network. Moreover we also used a scaling factor for all the models, to decrease of a given factor the number of units of the layer with respect to the original one. In this case we used $scaling\_factor = 0.025$. We then tried to implement the Magnitude-Based Pruning, with $n\_epochs\_pruning = 50$ and $final\_sparsity = 0.9$, but it did not result to be relevant for the purpose of those constraints. Finally, we used the weight quantization in order to decrease the size of the model, reducing it of a factor four approximately. The last step was to apply the compression of the file. The final results for version a are:

- Size of the file: 1.168 KB
- Temp MAE: 0.258
- Hum MAE: 1.176

B. Version b

The model we used is the original version of MLP with three fully connected layers so it can converge in a good way also with a smaller number of epochs and batchsize. To reduce the number of neurons, we used a $scaling\_factor = 0.03$. We used the weight quantization and the compression of the file as we did in the previous version in order to satisfy the constraints. The final results for version b are:

- Size of the file: 2.395 KB
- Temp MAE: 0.604
- Hum MAE: 2.395

Version	a	b
ouput_width	3	9
batch_size	512	32
learning_rate	0.01	0.01
n_epochs_training	100	20
scaling_factor	0.025	0.03

TABLE I
First exercise's parameters for versions a and b

Version	a	b	c
mel_bins	20	16	16
batch_size	64	32	32
learning_rate_training	0.02	0.03	0.03
scaling_factor	1	0.8	0.5
n_epochs_training	30	30	30
n_epochs_pruning	30	30	30
final_sparsity_pruning	0.7	0.8	0.8

TABLE II
Second exercise's parameters for versions a, b and c

II. Exercise 2

We tried the three different models shown in Lab3, *MLP, CNN-2D, DS-CNN* to start the analysis. The best performing one is the DS-CNN which seem to be the most efficient and the smallest in terms of size so we focused on it in order to satisfy the constraints of the different versions. The optimization method applied are: weight quantization, magnitued-based pruning and node-based pruning. The main challenge was to understand the best hyper-parameters to use in order to have good results. In particular we focused on the learning rate, scaling factor, batch size, number of mel bins. The scaling factor is on of the crucial parameter in order to reduce the size of the model instead the learning rate, the batch size and the number of mel bins are important for the accuracy.

For version a we could respect the constraints without applying the compression of the file, instead for version b and c we compressed the final pruned and quantized file.

This exercise involves the computation of the mfcc and stft which we had analyzed in the previous work so we decided to start with the best parameters found in the last Homework. For this reason to check the latency of the models the command line to use is:

python kwslatency.py –mfcc –length 256 –stride 128 –bins 16

The final results for the different versions are:

- A) Accuracy: 94.37 File Size: 53 KB
- B) Accuracy: 93.0 File Size: 29.62 KB Latency: 30 ms
- C) Accuracy: 92.37 File Size: 14.84 KB Latency: 25.41 ms