Machine Learning for IOTs – 2021/2022

Homework 2

Mustafa Elshaigi s289815
Joana da Orada Gonçalves s293583
Abdul Hadi Saeed s290480

1 Exercise 1

The aim of the first task of the homework is to infer multi-step, multi-output predictions for humidity and temperature on time series windows. The models chosen for the purpose were ['MLP', 'LSTM', 'CNN'], trained for 20 epochs and evaluated on test data using Multi-Output Mean Squared Error (MAE) as a metric. Each model was converted to TFlite and evaluated on the size and performance. Evaluating the results obtain, the model LSTM, was the model that had the worst performance and the biggest size.

In order to decrease the model complexity, structured pruning was applied with the width multiplier-α-by assigning the following values [0.5,0.3,0.1, 0.05,0.03]. To fulfil the size constraint, post-training quantization with weight only and weight plus activation were applied. The Weights plus activation quantization approximates both weights and activations were approximated, which leads to a increase in MAE. Since the weight only quantization had better performance providing the best balance between MAE reduction and file size reduction, with no restrictions of latency.

The constrains were met using MLP with the combination of structured pruning with value of α=0.03 and weight quantization for either version *a* and *b* combined with Zlib compression. The best results turned out to be ≈59x reduction regarding the size and ≈15% reduction in MAE values for both version. The following 1 represents the results obtained.

2 Exercise 2

For the second task, the sound is represented as Mel-frequency cepstrum. The previous studies refer that the CNN models have better performance because it can discriminate spectro-temporal patterns. The models used for this exercise are the following: ['CNN' , 'Depth_wise_CNN'].

Version-A: Structure pruning with α=0.5 combined with weight only quantization was applied in this version. Both of the models had good performances. The selected model was depth_wise CNN model because, although the same accuracy as CNN was obtained, a better reduction in the tflite model size was acquired ≈ 15x with 92.75% accuracy.

Version-B : CNN was chosen for this version. Since it was observed that the inference latency is contributing less than the MFCC prepossessing pipeline in the total latency. In order to reduce the latency, the mel bins were reduced to 16 and, frame length and frame steps increased by factor of ≈2x. This leads to a reduction of dimensions of the spectrogram and Mel-frequency cepstrum but also reduces the quality of the extracted features, lowering the accuracy. Structured pruning [α=0.4] combined with quantization aware training plus full integer quantization were applied. This approach was taken to compensate the reduction of the accuracy caused by the quantization, and also, by the MFCC pipeline modifications. The results were satisfied with ≈25x reduction of the size, only 0.75% drop in accuracy from the original Tf_lite model obtaining 92.5% and total latency of 23.72 ms

Version-C : Depth_wise CNN with Structured pruning [α=0.3] and weight only quantization was used. The values obtained are as follows: TF_lite model size after compression with Zlib of 23.4 kB, accuracy of 92.38% and total latency of 23.04 ms.

3 Conclusion:

From both exercises, it is possible to conclude that Neural Network Models are often over designed which leads to over-fitting and waste of memory and energy and increase latency. In sum, the first exercise has a conclusion that applying pruning and quantization made a significant reduction in the model size while enhanced the accuracy. The second exercise, it is possible to conclude that model inference time contributes less in the total latency, in our experiment the prepossessing pipeline time was ≈39x the inference time of the model. Adjusting the pipeline resulted in ≈2x reduction in latency. It was also observed that the depthwise convlutions gives very good approximations for the normal convolutions with fewer parameters, faster inference time and sometimes better accuracy obtained.

	VERSION A				
model	structured pruning	Q	size (kB)	MAE	
				TEMP	HUM
mlp	-	-	72.197 kb	0.296	1.138
	$\alpha = 0.03$	weight only	1.225 kb	0.256	1.189
cnn	-	-	66.453	0.257	1.136
	$\alpha = 0.05$	weight only	1.392	0.281	1.181
	VERSION B				
model	structured pruning	Q	size (kB)	MAE	
				TEMP	HUM
mlp	-	-	78.147	0.553	2.286
	$\alpha = 0.03$	weight only	1.414	0.552	2.385
cnn	-	-	69.356	0.574	2.237
	$\alpha = 0.05$	weight only	1.59	0.613	2.356

Table 1: Results from the first exercise.

	model	structured prun	quantization	mfcc	size (kB)	acc (%)	total latency (ms)
Tf lite	ds_cnn	-	-	A	567	91.2	-
	cnn	-	-	A	1192	93.25	-
a	ds_cnn	$\alpha = 0.5$	weight only	A	37	92.75	-
b	cnn	$\alpha = 0.4$	quant aware train..	B	47.3	92.62	23.72
c	ds_cnn	$\alpha = 0.3$	weight only	B	23.4	92.38	23.04

Table 2: Results from the second exercise

	frame length	frame step	lower freq	upper freq	mel bins	coeff
A	640	320	20	4000	40	10
B	1024	400	20	4000	16	10

Table 3: MFCC Options