HW 2 Group 14

Scorca Francesco s288876 Dicataldo Michele s290091 Padovano Dario s291475

1. Temperature and Humidity Forecasting

The architectures trained are two-layers MLPs and CNNs, exploiting "Early Stopping", namely training until the MAEs of the validation set reaches some threshold values, for a minimum of 16 epochs[1]. The optimizations considered are **Weight quantization**, **Structured pruning**, with α for the first layer, β for the second, **Magnitude-based pruning**. To avoid grid-searching we adopted a greedy approach, consisting in subsequent finetunings.

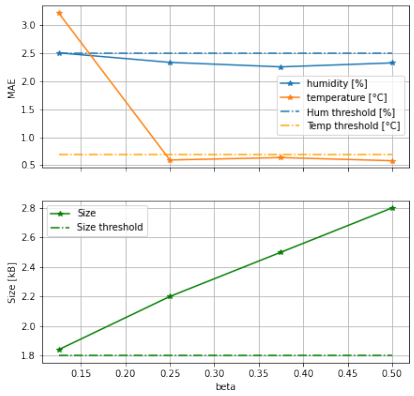

Figure 1

Considering version b and the CNN, the fully-connected affects the most the size, bringing probably redundant information. Thus, first we conducted few tests to fix sparsity = 0.9 and alpha = 0.5. Then, the behavior in function of β is shown in Figure 1, where we can see how reducing it is possible to keep stable results in terms of MAE, while having a linear reduction of size, finding $\beta = 0.125$. Then, we finetuned sparsity and α, noticing that not only reduced size, but also allowed the temperature MAE to not explode, highlighting a necessity of magnitude coherence among different layers. This workflow, applied to both architectures, led to the following results respectively for version a and b.

[1] since the final sparsity is reached at the 15-th epoch.

Mod.	alpha	beta	Sp.	Tmae	Hmae	Size
MLP	0.125	0.0625	0.78	0.263°C	1.16%	1.42kB
CNN	0.125	0.125	0.80	0.617°C	2.36%	1.76kB

Table 1

2. Keyword Spotting

The architectures trained are a MLP with 4 dense layers, a CNN having 3 blocks formed by convolutional layer, batch normalization and a dense layer, and a DS-CNN with an additional convolutional layer in the first two blocks (Depthwise or Separable). Although quality and time constraints are not orthogonal we can assume they will depend more respectively on the architecture (optimizations) and pre-processing.

2.1. Preprocessing

The principal test involved the number of mel-filters, the resize method and the frame length (fixing the frame step to half this). We found no difference among the bilinear, bicubic, nearest neighbour resizing methods, hence we kept the default bilinear. The results of the other two parameters are shown in Figure 2, showing that the number of filters needs to be a power of 2, and to prefer smaller windows. Thus, we found the number of mel-bins of 32, and a frame length of 256ms, later increased to 512ms.

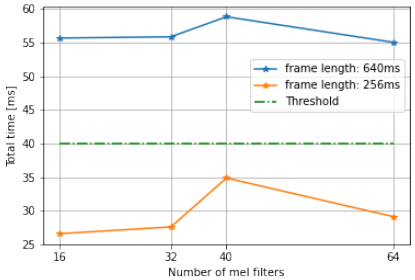

Figure 2

2.2. Optimization

The optimizations exploited are the same of Section 1, through which the DS-CNN outperformed the remaining

architectures both in terms of accuracy, size and latency time. Differently from the previous study, for magnitude-base pruning we used the same parameter α for all the layers, not being necessary to finetune a specifical multiplier for each of these.

2.3. Results

Table 2 shows the results. Pruning was necessary only in version c, where it collaterally brought inference time to a quarter of the initial value. A final observation can be done about the differences between the models used for *a* and *b*, that use simple depthwise convolution, and the one for *c* that go further using separable depthwise, achieving lower accuracy but lower size too.

alpha	**sp.**	**acc.**	**size**	**inf. lat.**	**tot. lat.**
1	//	93.80%	124.64kB	//	//
1	0.86	92.75%	45.67kB	6.97ms	28.03ms
0.25	0.65	91.30%	17.64kB	1.67ms	22.97ms

Table 2