Machine Learning for IoT - Politecnico di Torino

Homework 3 report

Paolo Aberto	Lorenzo De Nisi	Carmine De Stefano
StudentID : s278098	Student ID: s276545	Student ID: s278176

1 Big/Little Inference

For the big/little inference, we chose to use perform it with the use of REST, since it was the best choice when working with a single client and a server. When training the big model we used MFCC to maximize the accuracy, while we switched to STFT in the little one, in order to reduce significantly the inference time.
For both networks, we decided to use a DS-CNN model, increasing the number of convolutional filters in the big model. For the small model, we also used magnitude based pruning, in order to respect thesize constraints while mantaining a good accuracy.
We also implemented learning rate policies for both the networks during training.

We tried different success checker policies, like a treshold on the entropy of the probabilities array. We finally settled on the first-second policy, where we used the difference between first and second biggest probabilities coming from the network prediction layer.

When it comes to the communication cost, we managed to reduce it a lot by switching from Float32 to Int16 when sending the data to the server. This did not impact the resulting accuracy, while reducing the cost of communication by a significant margin and allowing us to call the server for a prediction 113 times, resulting in a comunnication cost of about 4.47 MB and a final accuracy of 93.12 %.

Model	**Size**	**Compressed size**	**Epochs**	**MFCC**	**Total latency**	**Test set accuracy**
Big	20			YES		95.12%
Little	20			NO		90.62%

Table 1: Models used for the big/little inference

2 Cooperative Inference

For the cooperative inference, the chosen number of models is 4.
The communication is performed through MQTT, considering that it is relatively easy to publish the recording once for all the, possibly many, models with respect to create a web service for each of them. We managed to keep the number of models as low as possible while preserving accuracy. As expected all the models have an individual accuracy that is lower than the cooperative final accuracy that is 95.13% .
Two of them are derived from the proposed DS-CNN, while the remaining two are derived from the proposed CNN, with some modification on the BatchNorm layer and filters and biases of the Conv2D.
To tackle the big amount of time needed to complete the 800 inferences, during testing the quality of service has been lowered to 0, avoiding the four-step handshake that was time consuming. To ensure the correct communication in the final commit of the homework the QOS is again 2 (but using the optional parameter `qos` on both `inference_client.py` and `cooperative_client` is possible to change it)
The cooperative policy consists of averaging the logits (output of the last layer of the models) and taking the argmax of them.

Model	**Modification w.r.t. proposed models**	**Epochs**	**lr**	**Test set accuracy**
CNN-0	-	20	0.01	94.25
CNN-1	Conv2D(filters=64, bias=True), BatchNormalization(momentum=0.2)	20	0.01	93.125
DS-CNN-0	-	20	0.01	93.625
DS-CNN-1	Conv2D(filters=128, bias=True), BatchNormalization(momentum=0.2)	20	0.01	92.50
Cooperative inference				**95.13**

Table 2: Models used for the cooperative inference