POLITECNICO DI TORINO

MACHINE LEARNING FOR IoT

HOMEWORK 2

GROUP 14,

BERNAUDO Jacopo, s276228

ERFANMANESH Omid, s272519

VASSALLO Maurizio, s276961

REPORT HOMEWORK 2

Ex1) Multi-Step Temperature and Humidity Forecasting

The solution for the version **A** is:

- The model architecture is a **MLP** with 2 Dense layers, each with 128 neurons and a final Dense layer with 12 neurons, the prediction window length(6) x prediction width(2: temperature and humidity);
- 3 types of optimizations are used: Weight-Only Post-Training Quantization, Structured Pruning, Magnitude-Based Pruning:
    - Weight-Only Post-Training Quantization allowed to reduce the size of the model of a factor of x3/x4 with a not too high increasing of the loss;
    - Structured pruning with an **alpha** value of **0.25**, this reduced a lot the model size but had some bad consequences on the loss;
    - Magnitude-Based Pruning with PolynomialDecay and an initial **sparsity** of **0.3** and a final sparsity of **0.9**, this allowed to reduce a lot the compressed size of the model with a small impact on the loss.

The results are:

Compressed tflite model size: 1.93kB, MAE: temp: 0.425 humi: 1.721.

The solution for the version **B** is:

- The model architecture is a **CNN** with 1 Conv-1D layer with 64 filters, 1 Dense layer with 64 neurons and a final Dense layer with 12 neurons;
- 3 types of optimizations are used: Weight-Only Post-Training Quantization, Structured Pruning (with **alpha=0.07**), Magnitude-Based Pruning with a final **sparsity** of **0.7.**

The results are:

Compressed tflite model size: 1.40kB, MAE: temp: 0.427, humi: 1.883.

Ex2) Keyword Spotting

The solution for the version **A** and **C** is:

- The model architecture is a **DSCNN** with 3 Conv-2D layers with 256 filters each and 1 final Dense layer with 8 neurons, the possible labels;
- The dataset is generated with **mfcc=True**, **frame_length=480** and **frame_stride=320.** These reduced both the inference time and the preprocessing time;
- 2 types of optimizations are used: Weight-Only Post-Training Quantization, Structured Pruning (with **alpha=0.282**), no Magnitude-Based Pruning is applied;
- Even if Magnitude-Based Pruning is not applied the model is compressed.

The results are:

Compressed tflite model size: 23.22kB, Accuracy: 0.907, Inference Latency: 1.10ms, Total Latency: 38.10ms.

The solution for the version **B** is:

- The model architecture is a **DSCNN** with 3 Conv-2D layers with 256 filters each and 1 final Dense layer with 8 neurons, the possible labels;
- The dataset is generated with **mfcc=True**, **frame_length=320** and **frame_stride=240.** These reduced a lot the inference time while keeping the accuracy high enough;
- 2 types of optimizations are used: Weight-Only Post-Training Quantization, Structured Pruning (with **alpha=0.285**), no Magnitude-Based Pruning is applied since this increased the both the inference time;
- Even if Magnitude-Based Pruning is not applied the model is compressed.
- Compressed.

The results are:

Compressed tflite model size: 23.25kB, Accuracy: 0.906, Inference Latency: 1.49ms.