Yes/No Classifier Performance Report for HW2

Authors: Suyash Singh, Harsh Vasoya, Junaid Shah
Student IDs: s307798, s308347, s319928
Politecnico di Torino

1 Introduction

The task at hand was to significantly improve a binary classifier within the constraints of IoT devices, which demand high accuracy and speed with limited resources. This report delves into the methodology, results, and implications of our advanced Yes/No classifier.

2 Methodology for Hyper-parameter Discovery

In optimizing hyper-parameters, our strategy was specifically designed for the dichotomy of 'yes' and 'no' audio classification. A targeted grid search honed in on parameters with the most impact on accuracy and TFLite model size, prioritizing real-world applicability and performance. The selection process was guided by validation set results, focusing on meeting critical project benchmarks for accuracy and computational efficiency.

3 Pre-processing Hyper-parameters

We invested heavily in pre-processing, with a focus on feature engineering to ensure superior input quality. Our fine-tuning aimed to capture the distinct elements of 'yes' and 'no' within the audio signals.

Parameter	Value
Sampling Rate	16000 Hz
Frame Length	0.016 s
Frame Step	0.012 s
Number of Mel Bins	10
Lower Frequency	20 Hz
Upper Frequency	4000 Hz
Number of Coefficients	30

Table 1: Pre-processing hyper-parameters

4 Training Hyper-parameters

We meticulously calibrated the learning rate and batch size, introducing sparsity to avoid overfitting. This precise fine-tuning ensured the model's learning efficacy.

Parameter	Value
Batch Size	30
Initial Learning Rate	0.005
End learning Rate	1.e-5
Epochs	20
Final Sparsity	0.70

Table 2: Training hyper-parameters

5 Model Architecture and Optimizations

We engineered a model optimized for IoT constraints, enhancing audio signal classification via strategic modifications to convolutional layers and sparsity induction through pruning. Leveraging TensorFlow's capabilities, we applied an adaptive learning rate for efficient training convergence and post-training quantization to minimize the model size, successfully achieving a synergy between compactness and accuracy.

6 Results

Our model demonstrated exceptional performance, surpassing all expectations. It achieved an accuracy of **99%**, with a model size of only **16.509 KB**, and reduced latency by **59.83%**, setting new standards for IoT classifiers.

Metric	Requirement	Achieved
Accuracy (%)	> 98.9	99
TFLite Size (KB)	≤ 25	16.509
Total Latency Savings (%)	> 35	59.83

Table 3: Model performance metrics

7 Discussion

This project sets a new bar in IoT machine learning with a TensorFlow model that efficiently decodes 'yes/no' prompts, outperforming benchmarks in accuracy and latency. Employing MFCCs and model pruning, strikes a balance essential for resource-constrained environments, paving the way for advanced, responsive IoT applications and future technological strides.