**Deep-Learning-for-Autonomous-Driving**

**Lab4 : Pruning & Quantization**

ID : 310605007

Member : 鄭晴立

Department : 機器人學程

目錄

[表目錄 2](#_Toc121222282)

[找不到圖表目錄。 2](#_Toc121222283)

[圖目錄 2](#_Toc121222284)

[1. Task 1 : Training 3](#_Toc121222285)

[1.1. Result 3](#_Toc121222286)

[1.2. Dirty data 3](#_Toc121222287)

[1.3. Data augmentation 3](#_Toc121222288)

[2. Task 2 : Fine-grained Pruning 4](#_Toc121222289)

[3. Task 3 : Coarse-grained Pruning 4](#_Toc121222290)

[4. Task 4 : Post Training Static Quantization 4](#_Toc121222291)

# 表目錄

# 找不到圖表目錄。

# 圖目錄

**找不到圖表目錄。**

# Task 1 : Training

在本次的實驗中，我在Training過程做了兩個調整，一個是Data augmentation，一個是刪除壞掉的音訊資料，最好的model可以到**91.47%，**以下先展示本次的training結果。。

## Result

圖 一 是model的Confusion Matrix，可以發現二類、第六類及第七類容易混淆，其分別是up、on和off，其發音非常類似；而第零類及第八類，分別為yes和stop，因為發與其他有巨大差別，所以分辨率特別高。

從

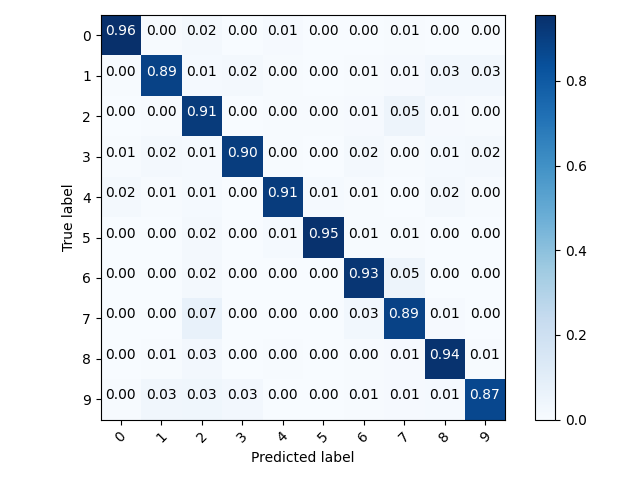


圖 一 最佳model的Confusion Matrix

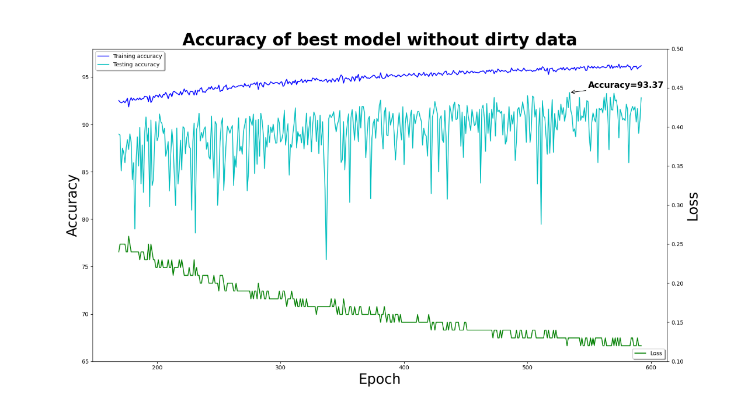


圖 二 訓練過程取縣

## Data augmentation

經過觀察，大多數的聲音檔噪音皆很大，有語調高低的不同、速度

的不同。原本我想利用torchaudio函式中sox\_effects.apply\_effects\_tensor去做處理

於是我利用librosa函式做處理

一張含有 文字 的圖片

自動產生的描述

## Dirty data

經過觀察，在Data裡面，有非常多音訊資料是有問題的，是人耳無法分辨的，包括幾種情況：

* 噪音太大，以至於完全聽不到正確發音
* 聽得懂在念什麼，但完全分類錯誤
* 聽不懂在念什麼，其發音完全無法分辨為該類別

我本來想利用波型去濾除錯誤檔案

，於是我利用第一次train出來的91%的model去找Confusion Matrix

，將每一次判斷錯的training data和testing data列出來，一個一個去聽，找出錯誤檔案

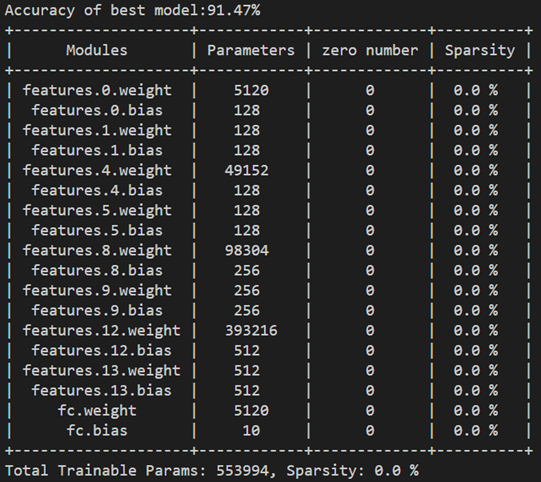
|  |  |  |
| --- | --- | --- |
|  | 用dirty data去訓練 | 用clean data去訓練 |
| 在dirty data evaluation |  |  |
| 在clean data evaluation | - |  |

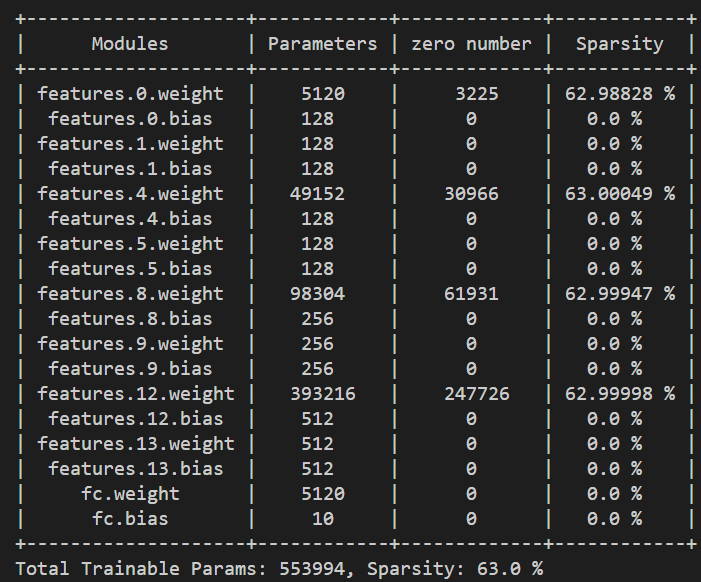
|  |  |  |
| --- | --- | --- |
|  | 用dirty data去訓練 | 用clean data去訓練 |
| 在dirty data evaluation |  |  |
| 在clean data evaluation | - |  |

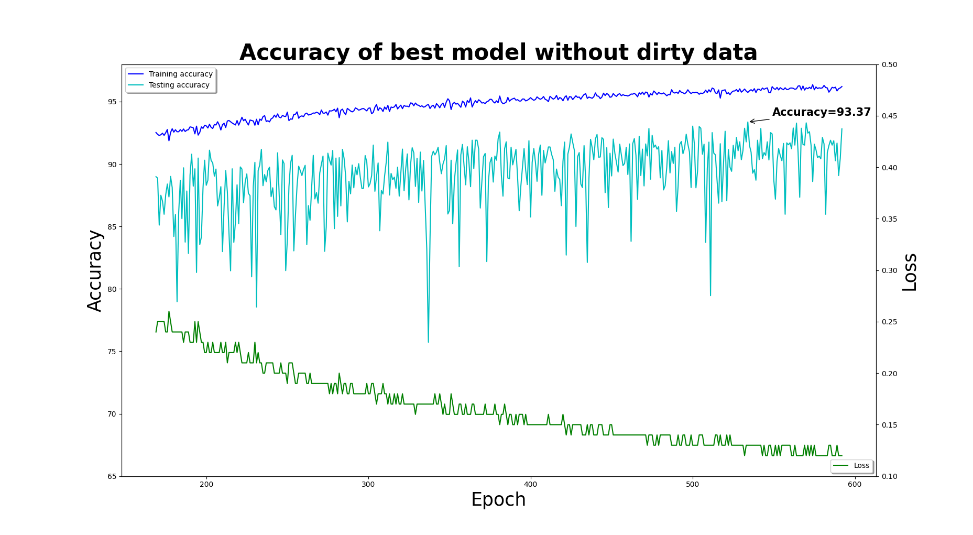
# Task 2 : Fine-grained Pruning

從下圖可以看到，Fine-grained之前Accuracy是前面所說的91.47%，sparsity為0

## Result







## Method

首先，將原本的model做deepcopy，避免更改到原本的model，並將需要Prining的layer取出來之後，將weight轉成numpy形式，再利用numpy的percentile函式取出絕對值最大的前幾百分位，將其作成一個mask，最後套上原本的weight就完成了。

一張含有 文字 的圖片

自動產生的描述

# Task 3 : Coarse-grained Pruning

## Result

## method

# Task 4 : Post Training Static Quantization

# Report

## What are the pros and cons of fine-grained pruning and coarse-grained pruning? Which one is more hardware-friendly and why? (5%)

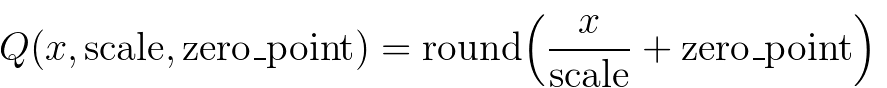
## Give your opinions on the benchmark table. (about speed and accuracy) (5%)

## Why do we fuse modules in quantization flow? (5%)

如同助教在影片所說，pytorch 所提供的quantization並不支援cuda，因此龐大的torch在cpu裡面會花很多時間去計算，因此fuse\_modules能將多層layer的model融合成一個模塊，以提升它的運算速度，但官方文件有說他也可以提升精確度，但我就不太確定是為什麼。

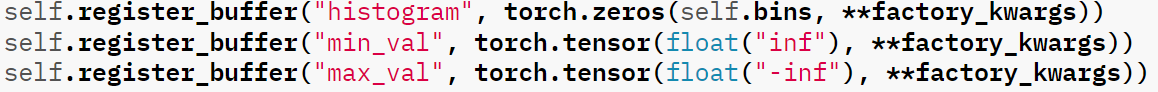
## In step 3 of post-training static quantization (PTSQ), we insert observers. What’s the use of these observers? (5%)

Quantization 轉化的方式為以下公式，需要找出scale及zero-point。從官方文件可以看到，會先利用add\_parameter去加入observer，加入之後會再去計算register中的變量histogram、min\_val、max\_val，透過這幾個參數去找出開頭提的cale及zero-point，並計算最終quantization的值



一張含有 文字 的圖片

自動產生的描述



## We use PTSQ in task 5. What’s the difference between PTSQ and post-training dynamic quantization? Which one is more hardware-friendly and why? (5%)

## Other (10%)