DSP FINAL

r08944022 蔡仲閔

vtsai01@cmlab.csie.ntu.edu.tw

()

Table of Contents

Preprocessing

在處理音訊分類時，對於聲音的預處理一直都是相當重要的一環，這邊從音訊產生頻譜後，利用2D CNN的模型來偵測頻域特性並進行分類，此段將探討過程中遭遇的問題及發現。

Magnitude and Phase Spectrum

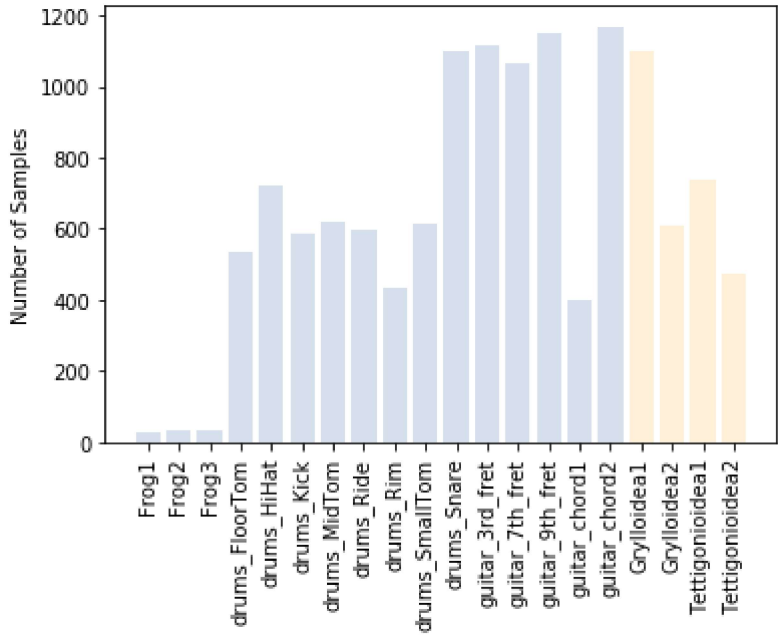
一開始選用 `scipy.signal.spectrogram` 來產生頻譜，而我們知道人類對於phase幾乎是無法辨別的，因此很自然的採用的是magitude，也得到不錯的結果。然而，這次的題目與一般音訊分類不同之處在於，**train及test的音訊是從同一段音源中sample出來**，因此即便人類無法辨別，仍有可能對我們的分類有幫助。

Mel Scale

如同上述我們利用頻譜來分類，然而仍有不足之處。 因為人耳對於音頻的感官，我們在這邊採用**Mel Scale**再對我們頻域值進行scaling，讓它更貼近人耳的聽感。 而這邊因為 `scipy.signal.spectrogram` 沒有內建這個功能，在這裡改用 `librosa.feature.melspectrogram` 來進行轉換。

Length of Sound

一開始為了節省訓練時間，我們將所有音訊切成等長去處理，但是這個方法畢竟可能遺漏了許多有用的訓練資訊，因此我們也嘗試了不同的方法。



上圖為training及validation set中各類別的數量，其中藍色為長度11025的sample，黃色則為22050。我們可以看到其實並不存在同類別不同長度的情況，雖然不正常，但是長度也許也會是有用的資訊。

Normalize

在將音訊轉換到頻譜之前，我們還需要將音訊進行normalize，一開始將音訊標準化到-0.5~0.5之間，但後來發現這樣的轉換會使的後面的fft失真，因此調整至0~0.5之間。

Model Description

訓練的部分我們使用 `torchvision` 提供的resnet (https://pytorch.org/docs/stable/torchvision/models.html) 來進行，我們知道Resnet(residual network)是強大的深度學習模型，它透過residual block之間的連結，並結合上一層的輸入作為下一層的輸入，來成功訓練足夠深的模型。 其中我也嘗試了resnet18及resnet50等不同深度的模型，其餘的細節將在下一章Experiments提到。
在完成訓練後，再使用ensemble的方法，將不同類型的訓練結果平均，得到最終提交的結果。

Experiments

在進行各種實驗後，固定了以下的設定，而我們將比較兩組結果，並使用 GTX 1070 進行訓練。

- normalize:

```
norm = lambda x : (data / (max+1e-6)) * 0.5
```

- padding: Pad to 22050
- optimizer: Adam

- learning rate: 0.01
- loss: CrossEntropyLoss
- batch size: 512

Generate Phase Spectrum

第一組我們選擇產生phase的spectrum，前面提到，一般來說人耳對於phase的敏銳度是相當低的，然而這次的dataset split的方式較為特別，也讓phase能發揮作用。
這裡採用的是 `scipy.signal.spectrogram` 來產生頻譜，其中參數 `fs=1.0, window=('tukey', 0.25), mode=phase` 。

Generate Magnitude Spectrum

第二組則是去產生Magnitude的spectrum，並使用Mel Scalse來使它更貼近人耳的感官，而這也是我們認為最能發揮功效的部分。然而實作上遇到一些問題，欲使用 `librosa.feature.melspectrogram` 來產生頻譜，卻發現它的運算時間相當長，幾乎是 `scipy.signal.spectrogram` 兩倍以上的時間，而 Data Generator 處理不及的情形下，也無法有效使用GPU訓練，妥垮整個運算速度。 為了提升訓練速度，我預先對音訊檔進行轉換，並另存成 `torch.tensor` 檔案，在訓練時只需要讀取預存的tensor即可。使用參數 `sr=22050, n_fft=2048` 。

Result

Experiments	**training loss**	**validation loss**	**validation accuracy**	**Public Set Score**
Phase Spectrum	0.298	0.462	0.960	**0.953**
Magnitude Spectrum	**0.091**	**0.334**	**0.982**	0.946
Ensemble	-	-	-	**0.978**

從上表的結果可以看到，Magnitude Spectrum 在各項數據的表現上幾乎都比Phase Spectrum 來的優秀，也符合我們一開始的預期，然而，Phase Spectrum 在 validation set 上也來到96%的準確率，似乎也有其出色之處，而最驚人的莫過於Public Set上的結果，Phase Spectrum 甚至超過了 Magnitude Spectrum。 因此與其選擇一種方法，最適當的方式應該是綜合各家所長，從下面兩張圖(fig1, fig2)可以看到 Magnitude Spectrum 唯獨在 guitar_3rd_fret 與 guitar_chard2 容易混淆，而 Phase Spectrum 卻可以準確分辨，因此我們決定在這邊採用Ensemble的方式，透過兩個輸出的平均來找到最好的結果，而最後我們也在 public set 上拿到很好的成績，**0.97818**，截至12/26 11:00 a.m. 仍是排行榜上最佳的結果。

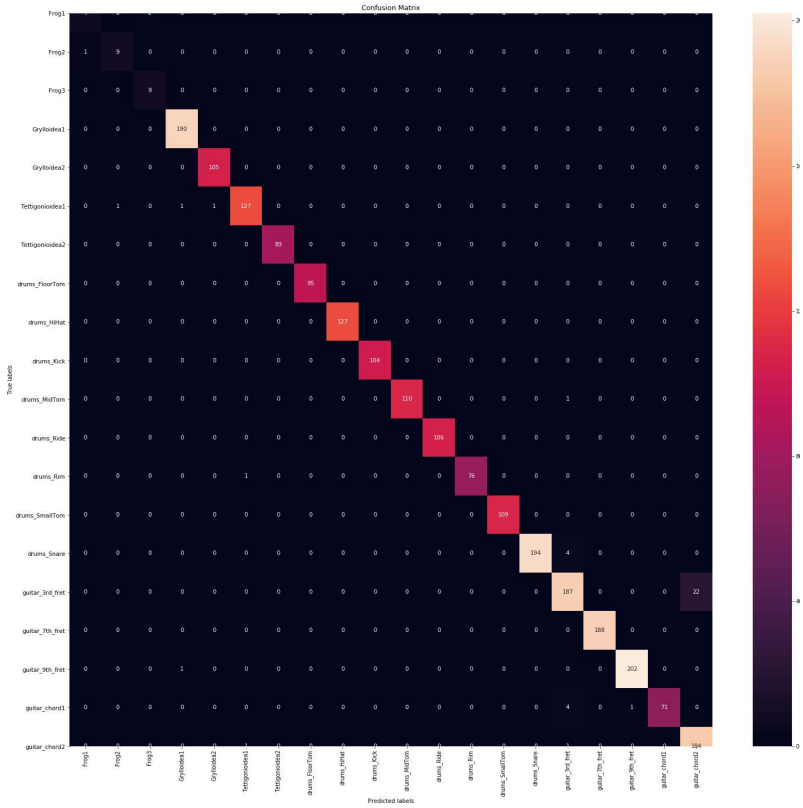


fig1. Confusion Matrix of Magnitude Spectrum Method

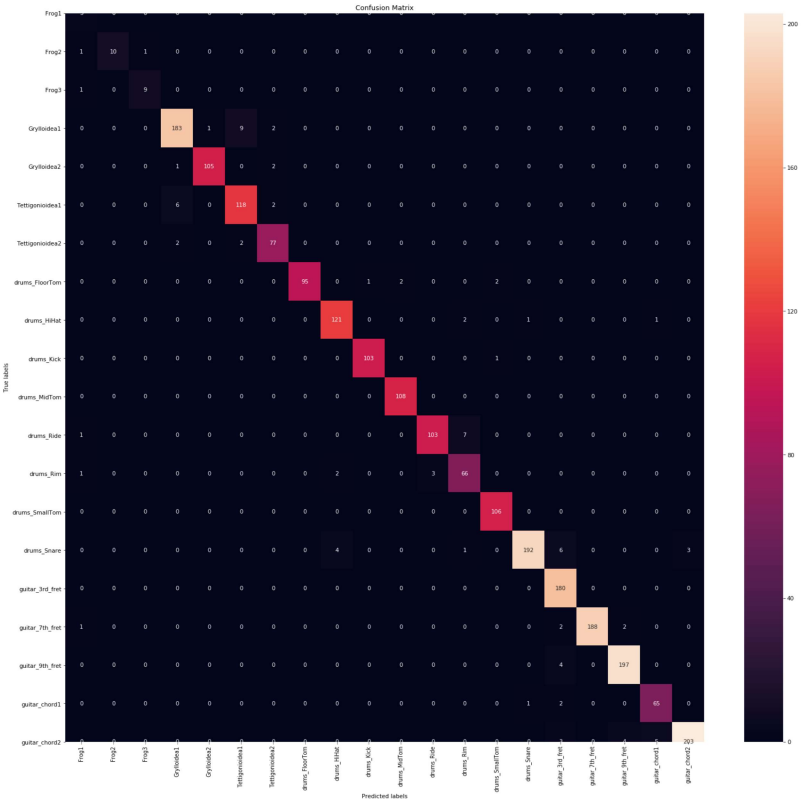


fig2. Confusion Matrix of Phase Spectrum Method