

Final Report

Team Name: NTU_b04705026_TeamName

Introduction & Motivation

這份專題的目標是使用機器學習訓練一個模型，給定一段音訊，辨識出他是真實世界中的什麼聲音。在過看三個題目在kaggle上leaderboard的分數後，覺得這個題目比較有潛力可以得到較高的分數，因此挑選這個題目。

Data Preprocessing/Feature Engineering

1. 資料處理:

a. 處理音訊長度不同^[1]

資料集一共有9400筆資料，共41個class，因為每筆音訊的長度並不相同，需要透過處理，讓fit進model的資料都相同長度。
對於不同model，有不一樣的目標長度

i. 原始音訊長度 < 目標長度

使用Padding的方式在音訊的後面補上constant的0。

ii. 原始音訊長度 > 目標長度

隨機決定符合目標長度要求的起始時間。在model training時，每個epoch都會重新裁切，可以彌補長音訊被裁切掉的損失與Bias。

2. 特徵抽取:

a. 取樣率 = 16000

b. 特徵一：Raw Audio

i. 使用原始音檔

c. 特徵二：MFCC

i. 使用梅爾頻率倒譜(Mel-Frequency Cepstrum)對每個音檔抽取特徵，同時再取其一次微分和二次微分做為第二和第三個channel。

ii. 實作：使用librosa套件中的librosa.feature.mfcc()函式^[2]

使用librosa套件中的librosa.feature.delta()函式^[2]

d. 特徵二：Log Mel Spectrogram

- i. 取出Mel Spectrogram後每個數值再取nature log^[2]
- ii. 實作：使用librosa套件中的librosa.feature.melspectrogram()函式
使用librosa套件中的librosa.feature.delta()函式^[2]

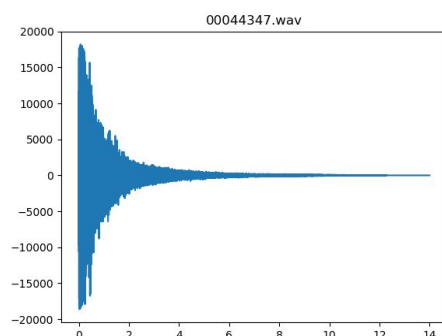
e. 資料增強(data augmentation)

抽取完特徵後我們也有嘗試data augmentation，使用keras 的 ImageDataGenerator，對資料進行平移(shift)、旋轉(rotate :30度)、縮放(zoom :0.8~1.2)、翻轉(flip)，以及推移(shear:0.2)，希望可以增加模型的。

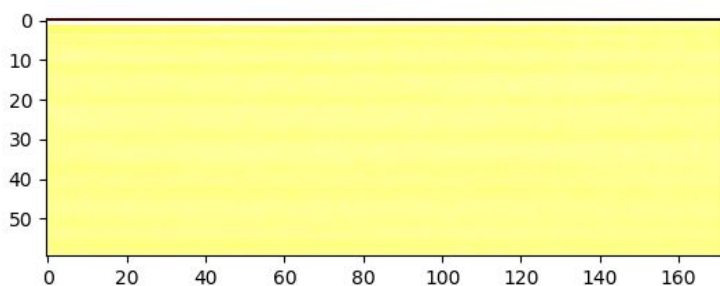
f. 資料標準化(data normalization)

同時也對每筆資料進行標準化，減去平均再除以標準差。

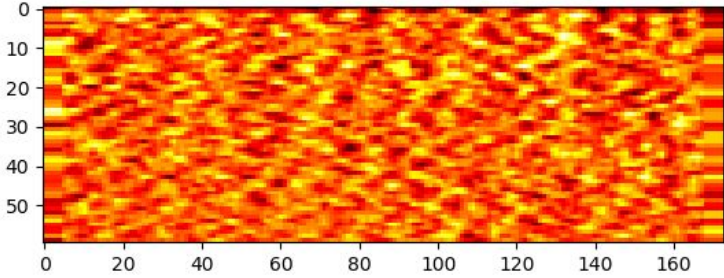
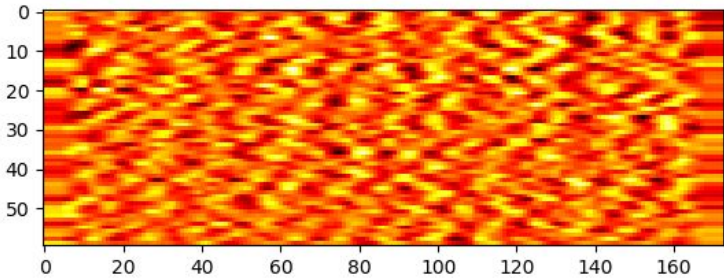
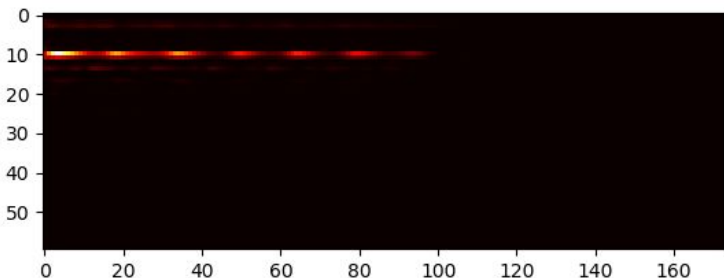
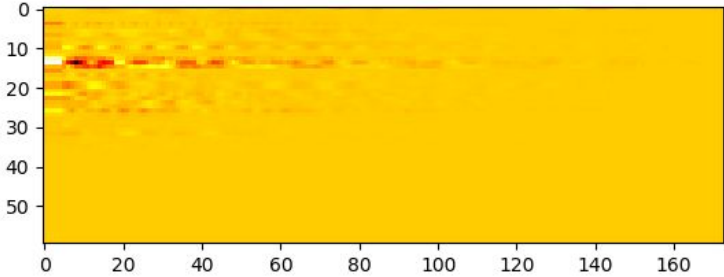
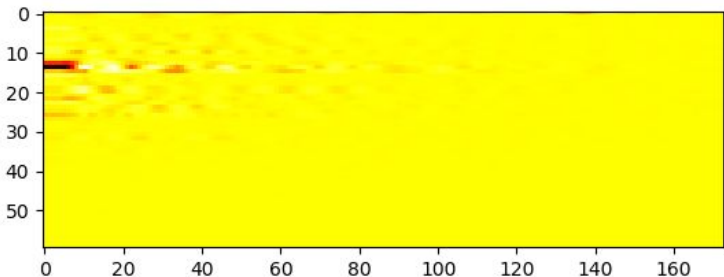
g. 資料特徵轉換範例

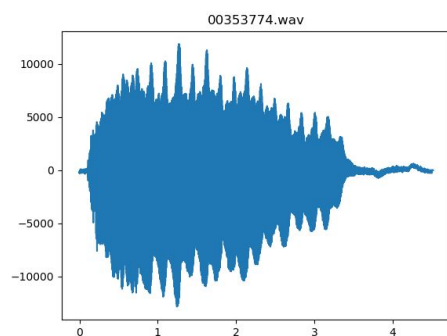


Data:00044347.wav,Hi-hat,manually_verified=0

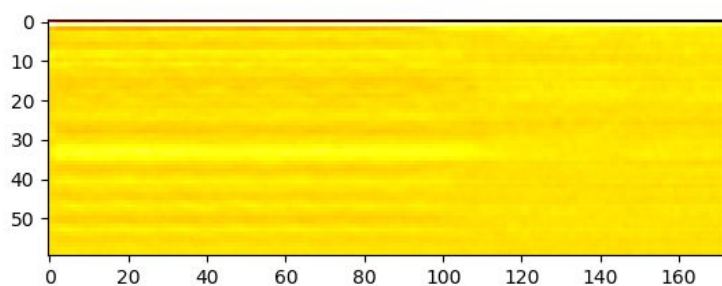


mfcc

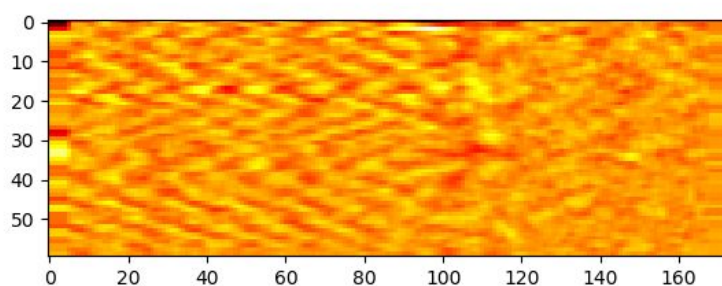
 <p>A heatmap visualization of the mfcc_delta feature. The x-axis represents time from 0 to 160, and the y-axis represents frequency from 0 to 50. The plot shows a dense, noisy pattern of red and orange colors, indicating high energy across the entire frequency spectrum over time.</p>	mfcc_delta
 <p>A heatmap visualization of the mfcc_delta_delta feature. The x-axis represents time from 0 to 160, and the y-axis represents frequency from 0 to 50. The plot shows a dense, noisy pattern of red and orange colors, indicating high energy across the entire frequency spectrum over time.</p>	mfcc_delta_delta
 <p>A heatmap visualization of the log melspectrogram feature. The x-axis represents time from 0 to 160, and the y-axis represents frequency from 0 to 50. The plot shows a dark background with a prominent horizontal band of red and orange colors near the top (low frequencies), indicating high energy in the lower frequency range over time.</p>	log melspectrogram
 <p>A heatmap visualization of the log melspectrogram_delta feature. The x-axis represents time from 0 to 160, and the y-axis represents frequency from 0 to 50. The plot shows a dark background with a prominent horizontal band of red and orange colors near the top (low frequencies), indicating high energy in the lower frequency range over time.</p>	log melspectrogram_delta
 <p>A heatmap visualization of the log melspectrogram_delta_delta feature. The x-axis represents time from 0 to 160, and the y-axis represents frequency from 0 to 50. The plot shows a dark background with a prominent horizontal band of red and orange colors near the top (low frequencies), indicating high energy in the lower frequency range over time.</p>	log melspectrogram_delta_delta



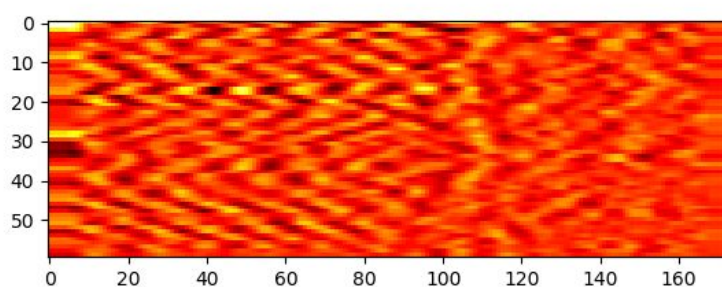
Data:00353774.wav,Cello,,manually_verified=1



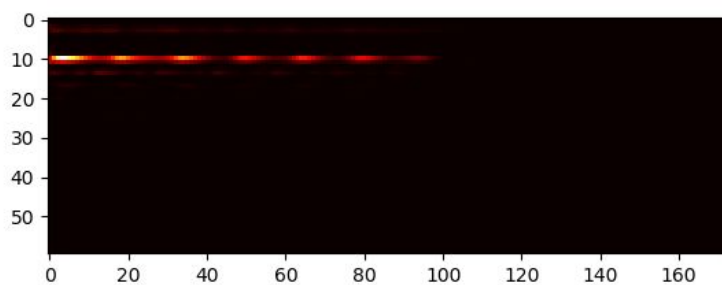
mfcc



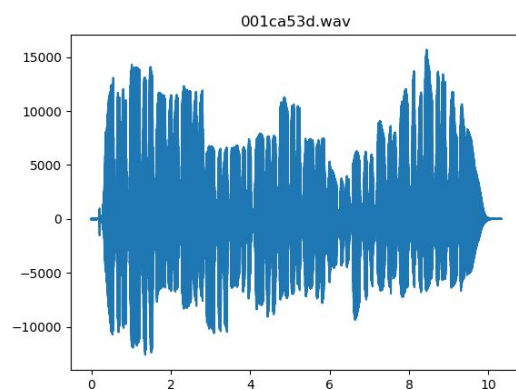
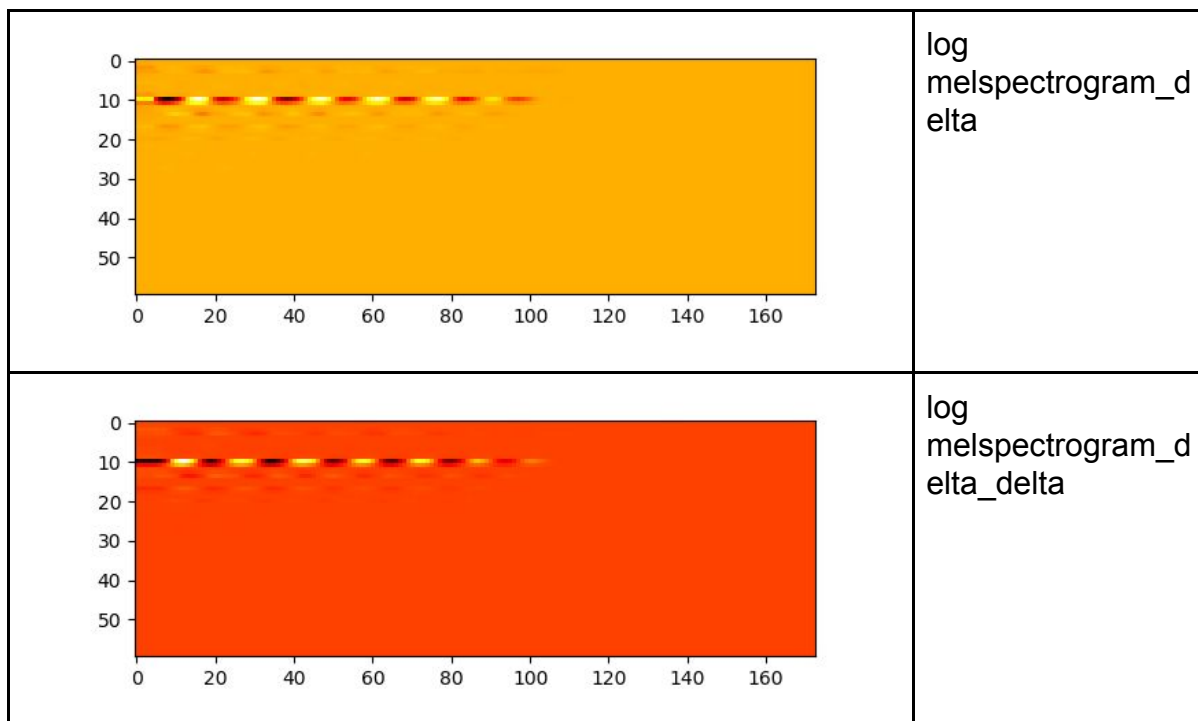
mfcc_delta



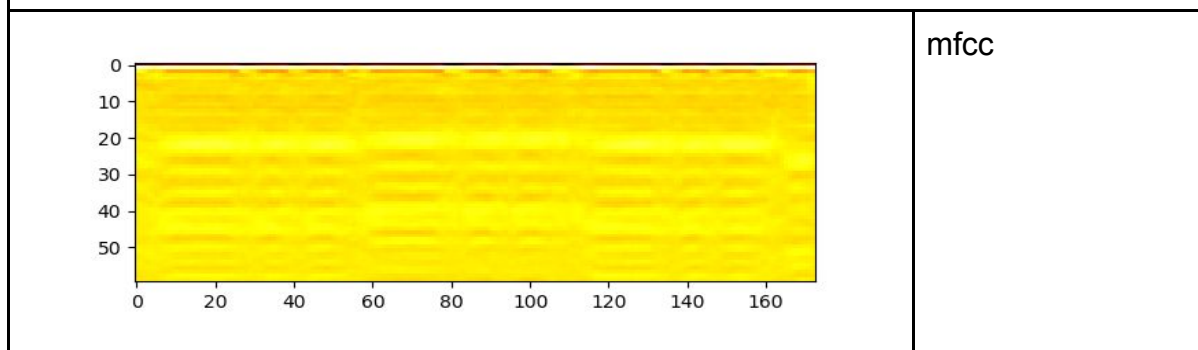
mfcc_delta_delta

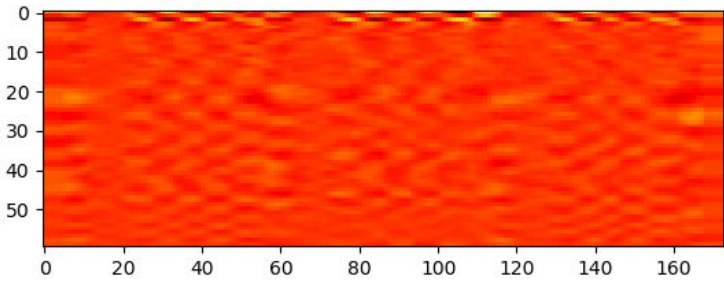
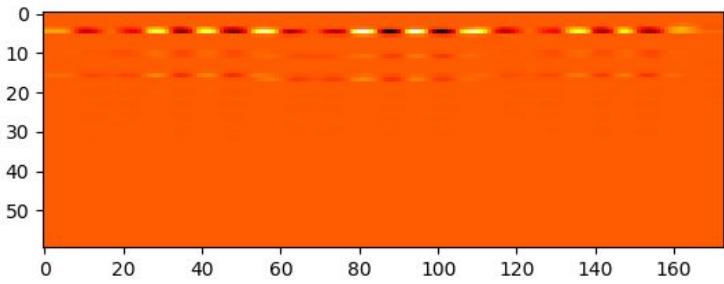
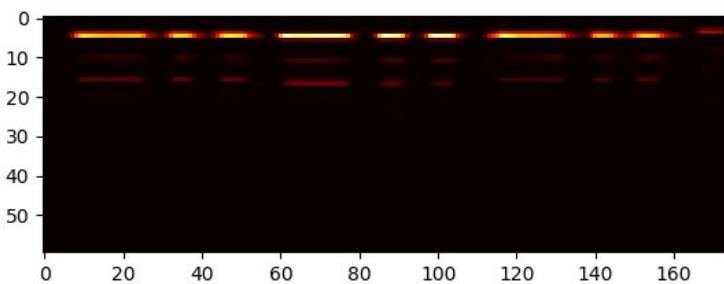
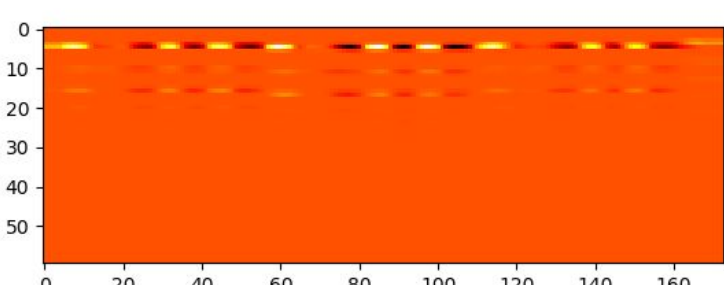
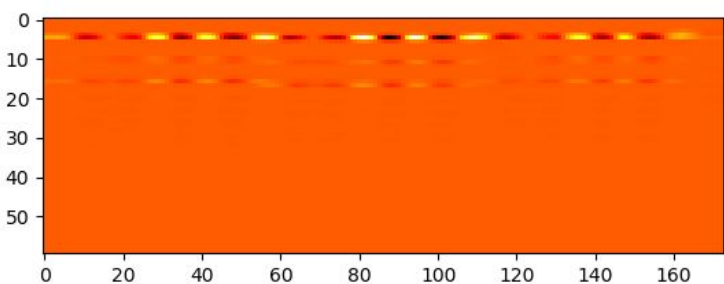


log
melspectrogram



Data:001ca53d.wav,Saxophone,manually_verified=1



 <p>A spectrogram showing MFCC delta values. The y-axis represents frequency from 0 to 50, and the x-axis represents time from 0 to 160. The plot is filled with a dense, noisy pattern of orange and red colors, indicating significant variation across the frequency spectrum over time.</p>	mfcc_delta
 <p>A spectrogram showing MFCC delta-delta values. The y-axis represents frequency from 0 to 50, and the x-axis represents time from 0 to 160. The plot shows a relatively uniform orange background with some faint horizontal lines of slightly higher intensity near the top, indicating lower variation compared to the delta plot.</p>	mfcc_delta_delta
 <p>A spectrogram showing log-mel spectrogram values. The y-axis represents frequency from 0 to 50, and the x-axis represents time from 0 to 160. The plot is predominantly black, with several distinct, bright horizontal lines of yellow and orange, representing specific frequency components that are active over time.</p>	log melspectrogram
 <p>A spectrogram showing the delta values of the log-mel spectrogram. The y-axis represents frequency from 0 to 50, and the x-axis represents time from 0 to 160. The plot is mostly orange with some faint horizontal lines of higher intensity near the top, similar to the mfcc_delta_delta plot.</p>	log melspectrogram_d elta
 <p>A spectrogram showing the delta-delta values of the log-mel spectrogram. The y-axis represents frequency from 0 to 50, and the x-axis represents time from 0 to 160. The plot is mostly orange with some faint horizontal lines of higher intensity near the top, similar to the mfcc_delta_delta plot.</p>	log melspectrogram_d elta_delta

Model Description

1. 模型選擇 - Convolutional Neural Net^[1]

使用Convolution的想法是，聲音的特徵出現的位子並不重要，重點是有沒有出現（與圖像辨識一樣，物品有出現就好，並不在意他在哪裡出現），再來就是一段聲音如果sampling rate變成一半，其實我們還是可以辨別出來聲音的類型，（如同照片解析度下降，一樣可以辨識出物品種類）。

綜合以上聲音與圖片特性的相同，我們決定採用Convolutional Neural Network當作我們3個模型的主要架構。

2. 1D convolution model^[1]

a. 使用Feature

- i. 裁切後的音訊檔。
為了增加模型多樣性，使用沒有傅利葉轉換的Input。

b. 使用Layer

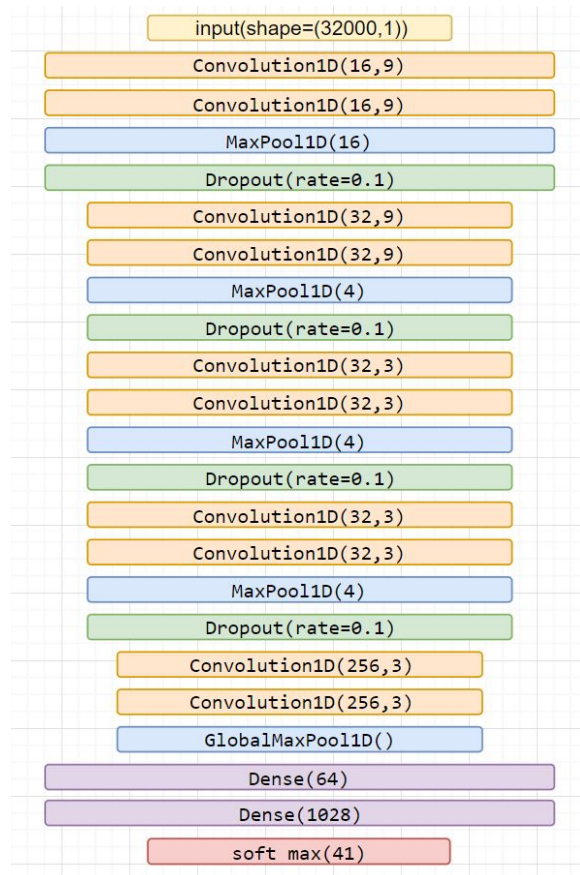
- i. ONE dimensional Convolutional Layers
- ii. Max Pooling Layer
- iii. Dropout Layer
避免Over-fitting
- iv. Fully Connected Layer
使用模型前段Extract的特徵，把後面Dense Layers當作分類器。
- v. Activation Function
使用 ReLU(Rectified Linear Unit)
最後一層Dense Layer使用Softmax代表機率分佈

c. 降低模型變異 - 10-Fold 算術平均

為了降低模型的變異(Variance)，我們使用10-Fold Cross Validation，輪流取不重複的10%資料，當成validation set，訓練出10個小模型。

最後的大模型用來inference，資料會forward pass經過10個小模型，最後經過“**算術平均**”每個小模型預測的分佈，取得最終大模型的預測結果。

d. 模型架構



3. 2D convolution model - MFCC_[1]

a. 使用Feature

- i. MFCC 梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients)

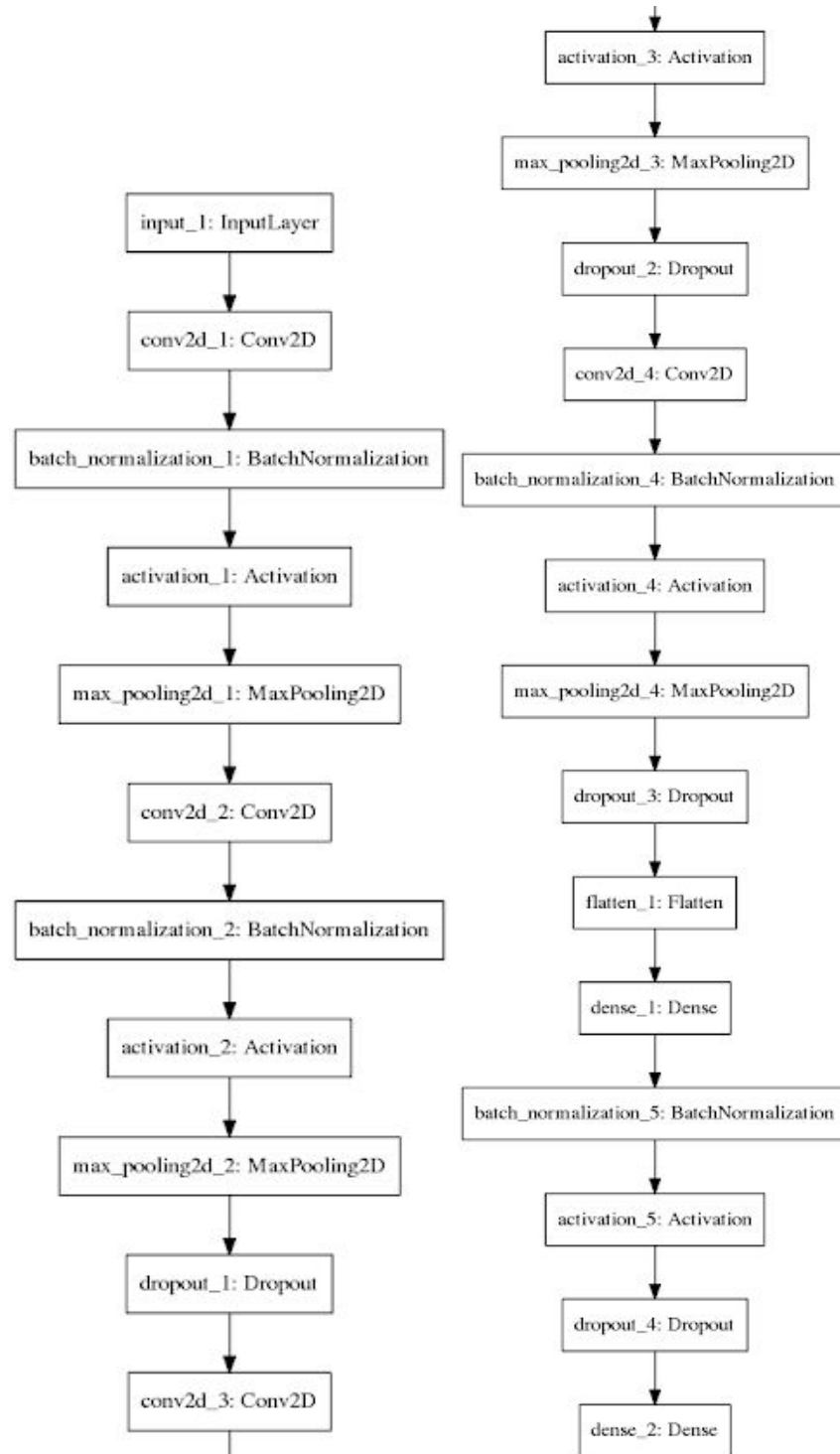
b. 使用Layer

- i. TWO dimensional Convolutional Layers
- ii. Batch Normalization Layer
可使用比較大的Batch size / 避免Internal covariate shift
- iii. Max Pooling Layer
- iv. Dropout Layer
避免Over-fitting
- v. Activation Function
使用 ReLU(Rectified Linear Unit)
最後一層Dense Layer使用Softmax代表機率分佈

c. 降低模型變異 - 10-Fold 算術平均

一樣使用10-Fold Cross Validation方法，做出十個小模型，最後經過”算術平均”每個小模型預測的分佈，取得最終大模型的預測結果。

d. 模型架構



4.2D convolution model - Log Mel Spectrogram

a.使用Feature

- i. Log Mel Spectrogram

b.Layer與模型架構 同上

c.降低模型變異 - 5-Fold 幾何平均

使用5-Fold Cross Validation方法，做出十個小模型。

每一個Fold的小模型會使用不同random seed去裁切，目的是希望能夠cover更多長音訊的片斷，增加模型的準確率。

最後經過”幾何平均”每個小模型預測的分佈，取得最終大模型的預測結果。

Experiment and Discussion

1.訓練細節

- a. Batch Size: 32
- b. Optimizer: Adam (Amsgrad)
- c. Initial Learning Rate: 1e-3
 - i. Adaptive Learning Rate: 降低LR如果6個Epochs Validation Loss沒有降低。
(這個方法通常讓模型Validation準確率增加了5%~10%)

2.單一模型表現

編號	Feature	一階微分	二階微分	CNN Structure	Avg. Val Accuracy	Public LB Score
(A)	Raw	x	x	1D	0.54	0.699
(B)	MFCC	x	x	2D	0.62	0.825
(C)	MFCC	v	v	2D	0.69	0.844
(D)	LogMel S.	v	v	2D	0.73	0.901

3. 單一模型討論

a. 1D vs 2D

- i. 採用原始音訊檔的模型相較於Spectrogram類型的模型，準確率相差一大截。
不過，(A) 模型與其他3個模型迥然不同，可以彌補其他模型的不足，會在下段討論1D模型權重不同的差異。

b. 一二階微分

- i. 由 (B) (C)可知採用一階微分(Delta) 以及 二階微分(Delta Delta)，會明顯增進模型表現。
單純MFCC 只會專注在音訊的“能量”，一二階微分可以補足音訊“動能”的資訊。
後續模型接採用一/二階微分當作第2/第3個Channel。

c. MFCC vs Log Mel Spectrogram

- i. 由準確率可知 Log Mel Spectrogram明顯勝出，其原因可能跟MFCC主要設計用來辨識人的語音，但是這個Task是對於很多像話的聲音進行分類，Mel Spectrogram相較於MFCC可以抓取到人聽不到的特徵，有額外的資訊，進而有較好的表現。[3]

4. Ensemble 模型表現

編號	RAW	MFCC_d_d	Log Mel Spectrogram_d_dd	權重方法	Public LB Score
(E)	0.5	0.5	x	算術	0.880
(F)	0.20	0.40	0.40	幾何	0.907
(G)	0.33	0.33	0.33	幾何	0.914

5. Ensemble 模型討論

- a. Raw & MFCC的模型兩者取機率分佈的平均，模型表現並沒有突破0.9，甚至比單一Log Mel Spectrogram還要差。
相較於模型(C)的0.844，只有微幅上升。我們認為應該減低Raw的權重

- b. 降低Raw的權重，以及增加Log Mel Spectrogram讓模型表現進步，我們認為應該是Log Mel Spectrogram貢獻了大部份模型的正確度，因為單純Log Mel Spectrogram就有0.901。因此可能跟Raw的權重不太有關。
- c. 於是我們使用三種模型權重均等，表現是我們所有模型中最好的。

Conclusion

以模型的表現來說，影響最大的就是特徵抽取的方式。如果不做特徵轉換用原始音檔，模型表現會非常差，不過同時也是因為模型的差異，最後把原始音檔當feature的模型加入ensemble當中，會彌補轉換成spectrogram兩個模型的不足。

MFCC與Log Mel Spectrogram的比較。因為前者會對Log Mel Spectrogram做一個Discrete Cosine Transform，比較專注於人類的聲音與聽覺，喪失一些聲音原有的特性。例如Bass這個聲音特徵頻率是低於人類聽覺的最低頻率(20Hz)，但是這裡的頻率又可以準確分類出這個項目。Log Mel Spectrogram並不會有這一層的資訊損失，所以對於辨別”多類型”的聲音優勢十分明顯。 [3]

Reference

- [1] <https://www.kaggle.com/fizzbuzz/beginner-s-guide-to-audio-data>
- [2] <https://librosa.github.io/librosa/index.html>
- [3] <https://www.quora.com/What-are-the-advantages-of-using-spectrogram-vs-MFCC-as-feature-extraction-for-speech-recognition-using-deep-neural-network>