AI CUP 2023 春季賽
多模態病理嗓音分類競賽報告

隊伍：TEAM_3134
隊員：李昶億（隊長）
Private leaderboard：0.607568 / Rank 6

壹、環境

作業系統: Windows 11
語言: Python 3.9.12
套件:

```
numpy==1.24.3
pandas==1.5.3
maplotlib==3.7.1
seaborn==0.12.2
torch==1.12.1
torchversion==0.13.1
librosa==0.10.0.post2
sklearn==1.2.2
```

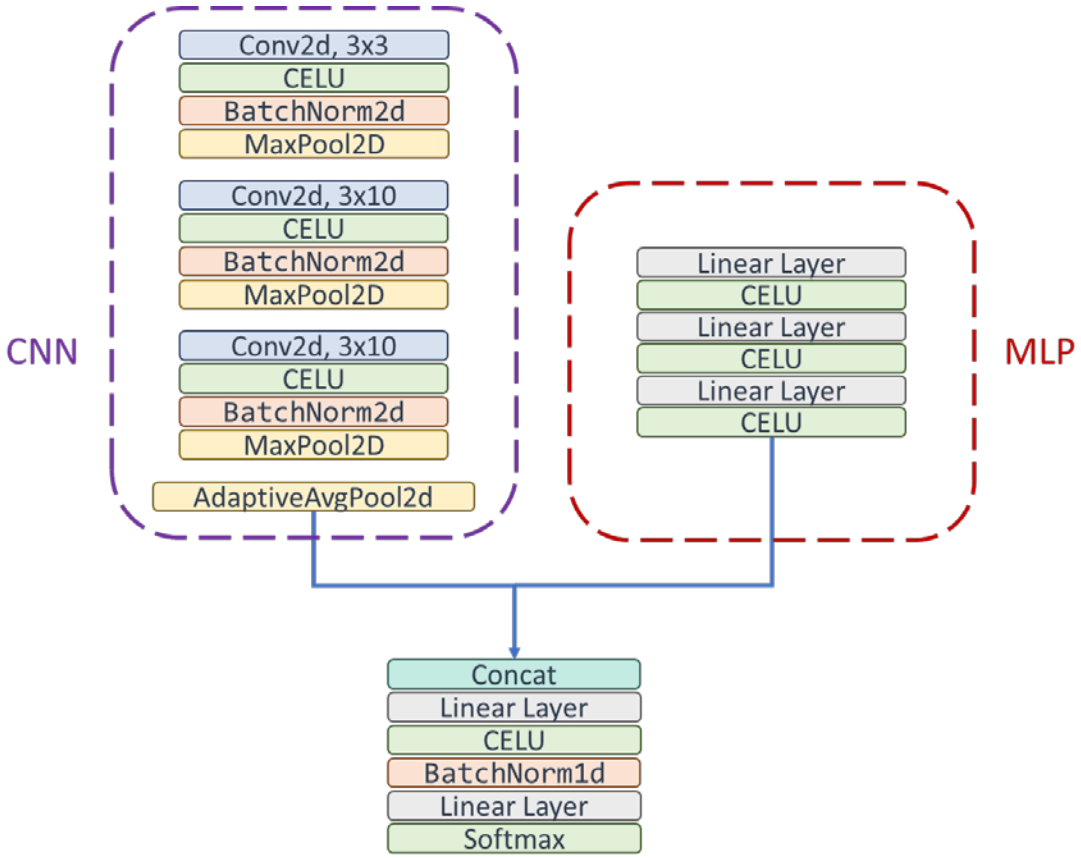
預訓練模型的部分，我們並無使用別人發布過的模型權重，是透過自己自訂義一個獨特的網絡結構來進行訓練，並達到分類的目的；此外，資料集的部分也沒有使用額外的資料集，所有的訓練或驗證資料皆是由官方所提供的。

貳、演算方法與模型架構

我們使用繼承nn.Module來建立一個模型，架構包含卷積神經網路（CNN）和多層感知器（MLP）兩個部分，如圖一。CNN的部分用來訓練MFCC聲音資訊，而MLP的部分用來訓練生理資訊，並在最後使用全連接層（FC）作為分類器。

CNN的部分使用了3層的卷積層，分別是conv1、conv2、conv3；輸出通道數為16、32、64；卷積核大小依序為（3， 3）、（3， 10）、（3， 10），選擇使用非正方形卷積核是參考了這篇文章[1]：因為音訊與時間密切相關，因此增加卷積核的橫向尺寸可以更好地捕捉時間資訊；為了提升模型的非線性能力在每層輸出後添加了CELU（Continuously Differentiable Exponential Linear Unit）激活函數；為了加速模型的收斂與穩定性，每層卷積層後都加入了Batch Normalization。此外，我們還使用了最大池化對圖片進行向下採樣，以減少圖片尺寸並更好地提取關鍵特徵。這些操作有助於提升模型的效能和準確性。

MLP的部分，包含了三層線性層，分別是linear1、linear2、linear3，每個線性層的輸出維度分別為1024、256、128。每層全連結層的輸出後都添加了CELU來提高非線性能力。透過MLP，我們可以從生理資訊中提取重要特徵並進行降維。



圖一、CNN + MLP 模型

最後連接MLP提取的生理特徵與CNN所提取的音訊特徵，並輸入到全連接層分類器（FC）進行分類預測。FC由兩個線性層組成，為fc1、fc2。首先將先前的CNN和MLP部分得到的特徵連接起來，形成一個綜合的特徵向量。這個特徵向量通過線性層fc1會變成128維，接著通過線性層fc2，將輸入維度從128降低到5。最後，將fc2的輸出通過Softmax函數，並將輸出轉換為表示各類別概率的形式，再取最高機率的類別作為最後結果。

參、創新性

我們使用了幾種創新的方法

1. **使用雙input的模型結構同時對MFCC（Mel-frequency cepstral coefficients）聲音資訊與生理資訊做處理，並將它們結合起來以提高分類的性能。**傳統上，聲音資訊和生理資訊通常被獨立應用於模型中，並分別進行處理和分析。然而，在本次比賽中，我們的模型將這兩種不同類型的資訊同時引入到模型中，並在網絡中進行特徵結合。此舉可以揭示出聲音資訊和生理資訊之間的潛在關聯性，也能提高分類的準確度。
2. **在多個層中使用CELU（Continuously Differentiable Exponential Linear Unit）作為激活函數。**傳統上，在神經網絡中常用的激活函數包括ReLU（Rectified Linear Unit）、Sigmoid和Tanh等，CELU是一種較新的激活函數。首先，CELU是一種連續可微的激活函數，表示在訓練過程中可以計算其導數，這對於本次使用的SGD（隨機梯度下降）反向傳播非常重要。相比於一些非連續或不可微的激活函數（如ReLU），CELU的連續性可以更好地幫助梯度的傳

播，從而提高模型的收斂速度和穩定性。其次，CELU和ReLU相比，能夠在負值區間具有負的響應，表示CELU能夠更好地處理負值的輸入，並提供更豐富的模型表示能力、幫助模型更好地捕捉到數據中的細節。

3. 在處理本次比賽的資料集中，發現五類資料時數據量極度不平衡的情況，其中最多的類別和最少的類別之間存在約15倍以上的差距。這種數據不平衡的情況容易對模型的訓練及預測產生影響，因為模型會更傾向於預測數量較大的類別。為了解決這個問題，我們在定義交叉熵損失函數（CrossEntropyLoss）時改變權重（weight）這個參數。透過將每個類別的倒數作為權重，可以平衡各個類別之間的樣本數量差異。**將數量較少的類別賦予較高的權重，使其在計算損失時具有更大的影響力**，從而提升數量少之類別的預測準確性。這種使用權重來抹平數據量不平衡造成的問題在處理不平衡數據集時非常重要。透過這種方式，模型可以更「公平地」對待各個類別，如此才到找到每個類別的重要特徵，提高整體的預測性能。

肆、資料處理

資料前處理分為聲音資訊與生理資訊兩部分：聲音資訊將轉為MFCC圖片，而生理資訊則是要經過編碼、填補缺失值等處理方式。

首先是聲音資訊：在本次比賽中使用librosa讀取聲音訊號並以官方提供的預設值給定n_fft和hop_length兩個參數，計算出MFCC特徵 [2] 。MFCC透過分幀、傅立葉變換等數學運算等多種處理方式，使其能夠捕捉聲音信號的特徵。接著將計算得到的MFCC特徵儲存為.npy文件，供後續訓練使用。

而對於生理資訊的處理，首先為了避免缺失值對模型訓練造成不良影響，我們將性別編碼為0和1，並針對有缺失值的PPD項則將其填充為0。此外，為了避免數值過大對模型訓練產生偏差，對年齡和Voice handicap index進行了正規化處理：分別將年齡及Voice handicap index除以與40，使這兩個特徵的範圍縮小，並能與其他特徵有相近的數值範圍。最後，我們使用one-hot encoding在類別型資料上，使其能夠更好的被模型利用。

結束前處理後，需將資料進行切分。由於這次比賽的各類數量太懸殊，我們使用StratifiedKFold來切分訓練與驗證資料，以確保不同類別間的比例。若使用隨機切分則可能導致少數族群全部被分到訓練或驗證的情況，讓模型訓練難度提升。

最後，我們將資料集封裝成PyTorch的Dataset物件，並使用DataLoader來讀取資料。訓練集的批次大小為32，並且將其隨機打亂，以提高訓練效果。

伍、訓練方式

在建立模型後我們將其放入GPU做訓練，以達到更快速的訓練速度，在訓練前先定義了損失函數與優化器。

損失函數定義為nn.CrossEntropyLoss，CrossEntropyLoss適用於多類別分類任務的損失函數。此外為了應對資料類別不平衡的問題，我們將權重（weight）調整為各類別資料量的倒數，這樣可以在訓練過程中平衡各個類別的重要性。

優化器定義為SGD（隨機梯度下降），它是一種常見的優化算法。我們將模型的參數放入優化器，讓優化器去更新模型的參數。學習率（lr）定義為0.01，而weight_decay參數則用於控制正規化項的權重，以防止模型過度擬合訓練數據。

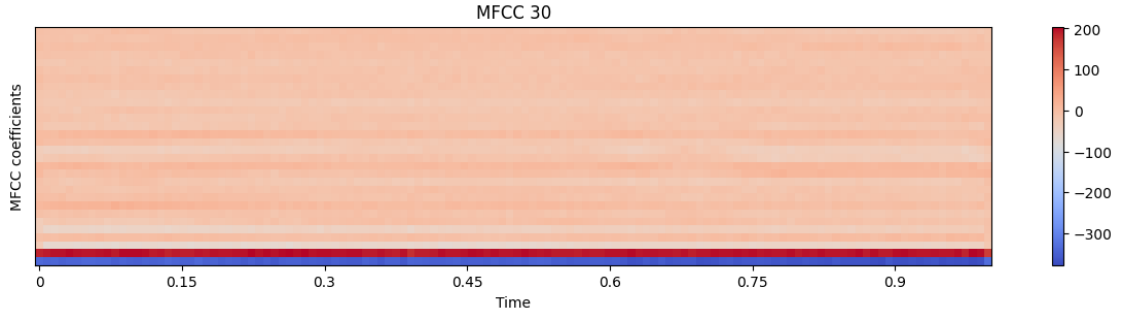
我們設定了150個epoch（訓練迭代次數），這足夠使模型收斂。在每個epoch中，模型對訓練資料進行一次前向傳播且計算輸出結果，然後將其與正確答案比對並計算損失值，接著根據損失值進行反向傳播，以便更新模型的參數。

在每個epoch結束後，我們使用UAR（Unweighted Average Recall）值以對模型進行評估。UAR是一種評估模型分類性能的指標，它計算所有類別的Recall分數並取平均值。UAR越高，表示模型更能抓出各個類別的資訊。在訓練過程中，我們也保留了最高UAR分數的模型，以供後續使用。

在比賽中，我們調整了五種n_mfcc參數，並得到五個模型。並使用集成（ensemble）方式結合每個模型的預測結果。我們將這五個模型對測試資料進行預測後得到的類別機率分佈結果進行相加，得到一個新的陣列，並取最高的機率最為最後的結果，這樣做可以提升預測準確度且更穩定和準確的預測結果。

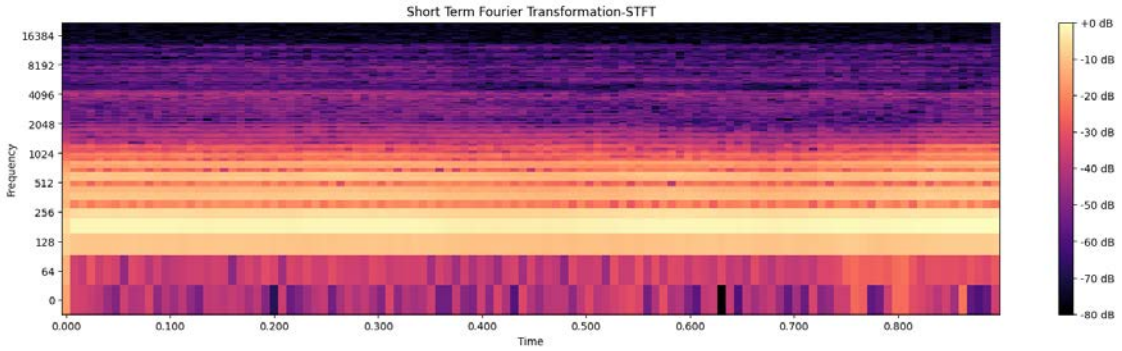
陸、分析與結論

在參加次比賽之聲音疾病分類之前，我並未有分析聲音資訊的經驗，透過參考官方提供的基準範例程式，我學習到了許多有用的知識。其中最重要的是如何將聲音轉換為MFCC（Mel-frequency cepstral coefficients）特徵資訊，如圖二。



圖二，MFCC （n_mfcc= 30）

在這之外也嘗試過使用STFT，即把聲音轉成聲譜圖，如圖三。但訓練效果並不理想，這可能是因為STFT未能捕捉到在本次比賽中與聲音疾病分類相關的關鍵特徵。



圖三、STFT聲譜圖

在結果方面，根據之前參與其他比賽的經驗，了解到集成方法可以有效地提升模型的準確度。因此在訓練模型的過程中，我們刻意地建立多種不同的模型進行訓練，以便進行集成。在這次的比賽中，我們選擇了五個具有相同結構但使用不同數量的MFCC圖像進行訓練的模型。這五個模型的MFCC數量分別為13、17、21、30和50，將這些模型的訓練結果進行集成，以獲得最終的集成預測結果。透通過這種方式，充分利用不同特徵數量的MFCC之特點和優勢，從而達到更高的整體準確度。

接著嘗試了添加偽標籤（Pseudo Label），他是一種半監督式學習的技術，就是將沒有答案的資料（public data & private data）放入模型做預測，把預測結果做為沒有答案的資料的標籤，並放入下一輪訓練中。然而，由於偽標籤本身可能存在一定的錯誤性，可能會導致額外偏差，且這次的分類任務相對較難（分五類而且各類的數量差甚大），所以最後並沒有獲得更好的結果。

總結來說，這次比賽我學到了如何處理聲音資料並成功應用於雙輸入模型，同時處理兩種不同類型的資料；在方法學上我學到了MFCC、STFT等聲音處理方法，這些方法對於聲音分析和特徵提取是非常重要的。在實戰方面也更加了解到集成方法的強大，讓我可以有60%的UAR，並獲得第六名的好成績，也體會到在分類任務過於困難時，偽標籤並不一定能有更好的結果，這點提醒了我在使用偽標籤時要謹慎評估其效果。

柒、程式碼

程式碼放置於Github，Github連結：https://github.com/JulianLee310514065/AICUP_audio_2023

捌、使用的外部資源與參考文獻

[1] Audio Event Detection (https://github.com/harmanpreet93/audio_classification)

[2] AI CUP 2023 春季賽【多模態病理嗓音分類競賽】巡迴課程(https://www.youtube.com/playlist?list=PLk_m5EiRQRF3j35iw-93Wh4cGa5fy41gu)