AI CUP 2023 春季賽
多模態病理嗓音分類競賽報告

隊伍：TEAM_3135
隊員：張芷婷（隊長）
Private leaderboard：0.558209 / Rank 28

壹、環境

1. 作業系統：windows 10
2. 語言：python 3.9.43
3. 套件：
   - librosa: 0.10.0
   - pandas: 1.5.3
   - torch: 1.13
   - sklearn: 1.2.2
4. 訓練資料：皆使用官方提供之資料，並無額外資料集。
5. 訓練模型：皆為使用套件架構模型，無使用預訓練模型

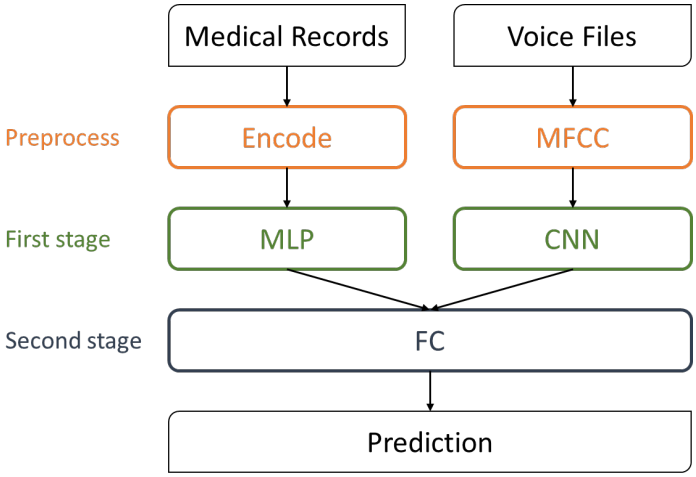
貳、演算方法與模型架構

在這次的比賽中，採用了**雙輸入模型**以整合生理資訊和聲音特徵。這種模型結構可以有效地結合兩種不同類型的輸入資料，並獲得更好的預測結果。

首先，使用 Multilayer perceptron（MLP）模型來訓練生理資訊。MLP 是一種基於神經網絡的監督式模型，由多個全連接層組成，適合處理複雜的非線性關係，具有高度靈活性、適應性等優點。在訓練過程中，使用了三層線性層（nn.Linear），並在每層**使用 CELU（Continuously differentiable exponential linear units）作為激活函數**以增強模型的非線性建模能力。

其次，使用 Convolutional neural network（CNN）模型來訓練聲音特徵。本次比賽會先將使用的聲音資料轉換為 **MFCC**（詳細請見資料處理），因此使用適合處理具有網狀結構的資料的 CNN。在訓練過程中，使用了三層卷積層（nn.Conv2d），並在每層使用 **CELU 作為激活函數**，再接上批次正則化層（nn.BatchNorm2d）與最大池化層（nn.MaxPool2d）。考慮到**聲音數據具有時間相關性**，因此在 CNN 中**使用了非正方形的濾波器**（filter）。這些非正方形濾波器能夠更能捕捉聲音資料中的時間和頻率特徵，提高模型對聲音數據的建模能力。

最後，將這兩個模型串接起來，並輸入至 Fully-Connected（FC）層中，並獲得最終的預測結果。FC 層的每個神經元都與前一層的所有神經元相連接，形成全連接的結構，適用於神經網絡的最後幾層。在訓練過程中，在第一層的全連接層使用 **celu 作為激活函數**再接上批次正則化層，第二層則使用 Softmax 作為激活函數並設置大小為 5，以得到各分類的機率。



參、創新性

1. 結合 MLP 及 CNN 成為**雙輸入模型**：由於生理資訊及聲音特徵為不同類型的資料，使用相同的模型可能無法充分利用這兩類資料的特性，因此在兩類資料使用不同的模型。在生理資訊使用 MLP 模型，這種模型可以處理複雜的非線性關係；而聲音部分則先轉換成圖像後使用 CNN 進行訓練，這樣可以更好地捕捉聲音資料中的時間和頻率特徵；最終再使用 FC 層作為最終輸出。
2. 使用 **CELU** 作為激活函數：CELU 為一種連續可微、自適應且具有非線性建模能力的激活函數，能夠提供平滑的梯度、適應不同尺度和分佈的數據、減少梯度消失的風險。與常見的 ReLU 相比，CELU 具有更強的非線性性質，可以更好地擬合複雜的非線性模式，提供更豐富的表達能力。
3. 在 CNN 中使用**非正方形濾波器**：由於**聲音數據具有時間相關性**，為了更好地捕捉聲音資料中的時間和頻率特徵，我們在 CNN 模型中採用了非正方形的濾波器。相較於傳統的正方形濾波器，這些非正方形濾波器更貼合聲音資料的特點，使模型能夠更準確地辨識和分析聲音的時間和頻率特徵。
4. 調整**損失函數權重**：由於此次比賽的資料集存在**分類比例不平衡**的情況，因此採取調整權重的策略來處理，以提高模型的表現。為了讓模型更加關注少量類別以提高對其的識別能力，**給予數量較少類別較高的權重**。
5. 使用**偽標籤**（pesudo labelling）：為了**增加訓練資料的多樣性及提高模型對未標記資料的泛化能力**，使用了偽標籤的方式。偽標籤將未標記的測試資料集的預測結果作為標籤，並將其與訓練資料集一起用於模型的訓練。
6. 使用**集成方式**：比賽中使用了不同參數訓練模型，為了**結合不同模型的預測結果**，使用了集成方式，以提高整體預測的準確性和穩定性。透過集成多個模型，可以彌補單一模型的缺點，並提供更穩定、更準確的預測結果。此外，為了讓模型更關注於少量類別，若有票數相同者則以少量類別為最終結果。

肆、資料處理

這次比賽中，資料處理分成四個部分：為了處適應資料集中類別間分布不均而使用的分層採樣（Stratified sampling）、處理聲音資訊的 Mel-frequency cepstral coefficients（MFCC）、將生理資料做編碼、及增加樣本的偽標籤（Pseudo-Labelling）。

首先，由於此次比賽的資料集中在類別間分布不均，因此使用 sklearn 中的 StratifiedKFold 方法來進行資料切分。這個方法可以確保在切分資料集時，**每個類別的資料比例都能保持一致**，並提高模型的穩定性和泛化能力。

接著，使用官方所提供的 **MFCC** 來處理聲音資訊。MFCC 是一種常用的聲音特徵表示方法，它將聲音信號轉換為一系列特徵向量，這些特徵向量反映了聲音信號在頻率和時間上的變化。另外，在這次比賽中**主要調整了其中的 n_mfcc 參數**。

再來是為生理資料進行處理。首先須將清除含有缺失值的資料，再來類別型資料編碼，及將數值資料做標準化。除了使用官方提供的程式，另外使用了 pd.get_dummies 方法將類別變數進行 one-hot 編碼，將其轉換為二元的特徵表示。這樣可以確保模型能夠更好地理解和利用這些類別特徵。

最後，為了增加樣本，**使用了偽標籤**來改善模型的表現。偽標籤是指將未標記的測試資料集的預測結果作為標籤，並將其與訓練資料集一起用於模型的訓練。這樣可以增加訓練資料的多樣性，提高模型對未標記資料的泛化能力。

伍、訓練方式

1. 載入資料集，並進行前處理（如資料處理所述）。在處理 MFCC 時，使用五種 n_mfcc 參數（13、17、21、30、50）。
2. 將處理後的資料建成 Dataset 並用 Dataloader 讀取之，以讓資料分成小批次進行訓練並進行預測，並使用 Cross entropy 計算損失以更新參數。訓練過程中，以 Unweighted average recall 評估模型，並儲存最高分模型以做後續使用。
3. 將測試資料集的預測結果儲存後，把這些預測結果作為額外的訓練資料集，再併入原始的訓練資料中以訓練模型，即使用偽標籤的方式。重複以上步驟 15 次。
4. 利用集成方式將五個不同 n_mfcc 參數的模型預測結果結合起來。為了讓模型更注重於少量類別，若有票數相同者，則以在資料集中數量最少的分類作為最後結果。

透過以上步驟，我們能夠綜合利用不同 n_mfcc 參數值的模型，並透過偽標籤和集成技術提升模型的預測能力和穩定性，以獲得更好的結果。

參數：

- Batch size: 32
- Epoch: 150
- Optimizer: SGD
  SGD 為一種常用的優化器，具有計算效率高、隨機性、適用於大型資料集等優點，並可以提高模型的泛化能力。
  - Learning rate: 1e-2

- Weight decay: 1e-4
- Loss function: Cross Entropy
  - Weight: 為資料集中**各類別的倒數形成的張量**。
    由於此次比賽的資料集分類比例不平衡，透過調整權重的方式，即增加數據量較少類別的權重，可以幫助模型更好地處理少數類別，並改善模型表現。

陸、分析與結論

模型架構：

在官方所提供的範例程式中，所使用的模型皆為 DNN 模型。然而在訓練過程中，Public 分數的上限只能達到約 0.44，因此嘗試其他模型。

在聲音資料的部分，由於聲音資料與頻率、時間等相關，因此改以圖像的方式擷取聲音資料的特徵，並以 CNN 模型進行訓練。而在生理資訊部分，則使用 MLP 模型進行訓練。

最終結合了以上兩種模型，達到 Public 分數約 0.60 的結果。因此顯示不同模型對於不同類型的資料可能具有不同的優勢，而更具資料特性選擇不同模型對於提高預測準確度有顯著的改善。

方式	**皆為 DNN**	**結合 MLP 與 CNN**
最高 Public 分數	0.44	0.60

聲音特徵：

在將聲音轉成圖形特徵的過程中嘗試了兩種方式：Short-Time Fourier Transform（STFT）及 Mel Frequency Cepstral Coefficients（MFCC）。

STFT 是一種將聲音訊號轉換為時域與頻域表示的方法，透過計算不同時間窗口的頻譜，可以捕捉到聲音訊號在不同時間和頻率上的變化。然而，在這次比賽中，STFT 並沒有達到更好的效果。

方式	**STFT**	**MFCC**
最高 Public 分數	0.59	0.60

最終則以 MFCC 得到了較好的結果，但由於我並無處理聲音資料的經驗，因此無法解釋其中的原因。未來若有機會可以再嘗試不同聲音處理方式，如頻譜圖、線性預測編碼等方式。

結論：

- 在模型方面捨棄原本的 DNN，結合 MLP 及 CNN 得到了更好的結果。
- 將聲音資訊轉為圖像的 MFCC 特徵，可以獲得更高的分數。

柒、程式碼

Github 連結：[nicochang18/AICUP_2023_Spring_acoustics]

捌、使用的外部資源與參考文獻

[1] **Multi-layer Perceptron.** Mitra, S., & Pal, S. K. (1995). Fuzzy multi-layer perceptron, inferencing and rule generation. IEEE Transactions on Neural Networks, 6(1), 51-63.

[2] **Continuously differentiable exponential linear units.** Barron, J. T. (2017). Continuously differentiable exponential linear units. arXiv preprint arXiv:1704.07483.
[3] **Mel Frequency Cepstral Coefficients.** Logan, B. (2000, October). Mel frequency cepstral coefficients for music modeling. In Ismir (Vol. 270, No. 1, p. 11).
[4] **Short-Time Fourier Transform.** Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 32(2), 236-243.