深度學習應用於語音情緒辨識

國立中正大學 National Chung Cheng University
電機工程學系
專題生:陳睿宇、陳頡毅　指導教授:余松年

(一)摘要

本專題結合語音特徵擷取、Mixture of Experts（MoE）與卷積神經網路（CNN），設計語音情緒辨識系統。以 MFCC 與 Mel Spectrogram 為雙特徵輸入，透過專屬卷積分支與動態門控機制融合判斷結果，提升辨識準確率。系統以 RAVDESS 與 EmoDB 為資料來源，採用 Hold-out 驗證，並以 Accuracy 與 Cohen's Kappa 評估效能，證實架構具良好穩定性與泛化能力。

(二)實作方法

本研究的實作流程包含五個主要步驟：資料前處理、特徵擷取、模型設計、訓練與驗證，以及效能評估。系統以 Python 為主要開發語言，採用 TensorFlow 與 Keras 建構模型，並使用 Librosa 進行語音特徵轉換與視覺化。

首先，在資料前處理部分，選用公開語音情緒資料庫 RAVDESS 與 EmoDB 作為訓練與測試來源。兩者情緒類別涵蓋中性、快樂、悲傷、憤怒、恐懼、厭惡、驚訝與無聊等八種情緒，經統一標籤後以 LabelEncoder 進行數值化。為確保一致性，所有音訊長度經標準化並轉為相同取樣率。接著進行特徵擷取，以 MFCC 與 Mel Spectrogram 為主要特徵。每段語音轉換成 128×128 的特徵矩陣並正規化後輸入模型。MFCC 能反映語音能量分布與共振峰特性，Mel Spectrogram 則保留時間與頻率的能量變化，有助 CNN 擷取情緒模式。

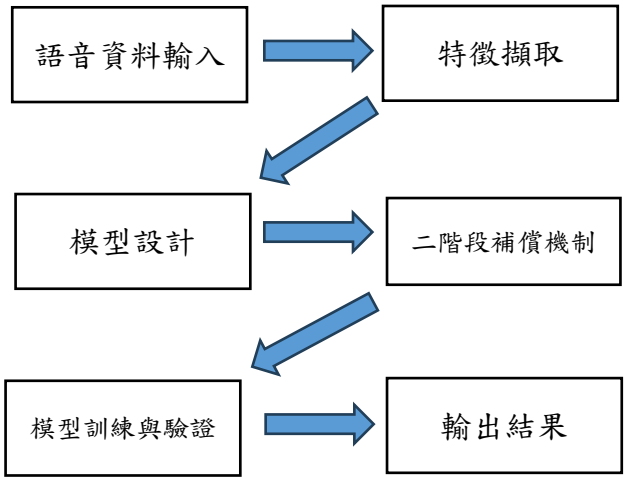
在模型設計部分，本研究採用 多專家學習架構。系統設有兩個 CNN 分支，分別輸入 MFCC 與 Mel 特徵，各自學習不同特徵空間的情緒資訊，並透過 Gating Network 自動加權融合兩者結果。各分支皆使用多層卷積與池化結構，並採 LeakyReLU 激活函數以避免神經元死亡問題。

此外，為改善情緒混淆，本研究設計二階段分類補償機制。第一階段將情緒分為 Positive、Negative、Neutral，第二階段再針對各群組細分類，如將正向群組區分為 happy 與 surprise，負向群組再區分為 sad、angry、fear、disgust 等，以提升辨識準確率。

模型以 Adam 優化器 訓練 30 個 epoch，並採用 Hold-out（8:2）驗證策略，以提升結果穩定性。最終以 Accuracy 與 Cohen's Kappa 為主要評估指標，並搭配 Precision、Recall、F1-score 與混淆矩陣分析各情緒辨識效果。

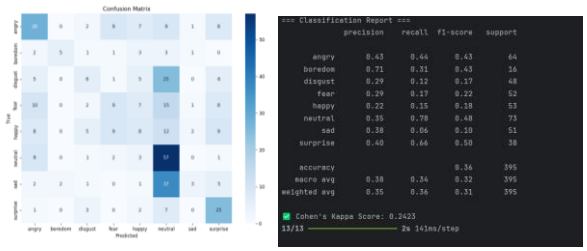
整體而言，透過 MoE 架構與二階段補償策略，本研究成功整合多特徵輸入並提升模型對多情緒類別的辨識效能。

(三)實驗流程圖

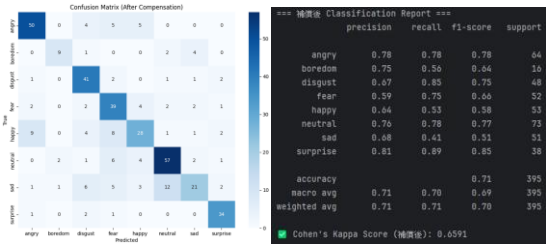


(四)實作成果

MOE



二階段補償



(五)結論

本研究成功設計一套結合 MoE 與 CNN 的語音情緒辨識系統。透過 MFCC 與 Mel Spectrogram 雙特徵輸入及動態門控機制，模型能自動調整各專家權重，有效提升情緒分類的準確率與穩定性。同時導入二階段分類補償機制，改善易混淆情緒之辨識問題。實驗結果顯示，本架構在公開資料庫上具良好表現，並經 Hold-out 驗證 證實其可靠性。未來將持續擴充資料集、優化 gating 設計並導入注意力機制，以提升跨語料辨識能力，推進語音情緒辨識於智慧助理與人機互動領域的實際應用。