Emotion Recognition through Speech Using Convolutional Neural Networks

Abstract:

In recent years, the intersection of technology and emotion recognition has gained significant attention. One fascinating area of research is the utilization of Convolutional Neural Networks (CNNs) to recognize emotions from speech signals. In this project, we explore the implementation of a CNN model for emotion recognition using the TESS Toronto emotional speech dataset.

Data Preprocessing:

The TESS dataset consists of audio recordings categorized into seven emotions: angry, disgust, fear, happy, neutral, surprise, and sad. The initial steps involve loading the data, extracting features from the audio files, and augmenting the dataset to enhance model robustness.

Librosa, a Python library for audio analysis, is employed to extract features such as zero-crossing rate, chroma_stft, MFCC, root mean square value, and mel spectrogram.

Augmentation techniques include introducing noise, time stretching, shifting, and pitch variation to diversify the dataset.

Data Exploration:

After preprocessing, the dataset comprises 5600 samples, each associated with a specific emotion label. The distribution of emotions is balanced, with 400 samples per emotion category.

Feature Scaling and Model Preparation:

To prepare the data for training, it is split into training and testing sets. Standard scaling is applied using sklearn's StandardScaler to normalize the features. The input shape of the data is then adjusted to fit the CNN model's requirements.
CNN Model Architecture:
The CNN model is designed with four convolutional layers, each followed by max-pooling to capture hierarchical features. Dropout layers are included to prevent overfitting, and dense layers provide the final classification. The model is compiled using the categorical crossentropy loss function and the Adam optimizer.

Training and Evaluation:

The model is trained over 50 epochs, with a ReduceLROnPlateau callback to adjust the learning rate dynamically. Training and testing accuracy and loss are monitored throughout the process.

Results:

The model achieves impressive results on the testing set, with an accuracy of approximately 97.71%. This demonstrates the effectiveness of using CNNs for emotion recognition in speech.

Conclusion:

In conclusion, this article provides insights into the application of Convolutional Neural Networks for emotion recognition in speech. The TESS dataset, preprocessing techniques, and model architecture contribute to building a robust system. Emotion recognition through speech holds promise in various applications, from human-computer interaction to mental health monitoring. Further research and refinement of models could lead to even more accurate and reliable emotion recognition systems.