Fusion of Deep Learning Features with Mixture of Brain Emotional Learning for Audio-Visual Emotion Recognition

125015103 Samatha A
125015120 Srividhya S
125015111 Shahana S

Multimodal emotion recognition, particularly in the domains of speech and facial expressions, has garnered significant interest in various applications such as e-learning , human-computer interaction, health monitoring , mobile computing and gaming .Multimodal emotion recognition is one of the main challenging tasks due to the multimodality characteristic of human emotional expression.

In response to these challenges, this paper introduces an innovative approach—a fusion model combining deep learning features with a Mixture of Brain Emotional Learning (MoBEL) model inspired by the brain's limbic system. The proposed model addresses the spatial-temporal correlations in video content, recognizing the intricate patterns of emotion expressed through both audio and visual cues.

Our methodology involves two primary stages: first, leveraging advanced deep learning techniques, including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), to extract highly abstract features from audio and visual data. Second, the introduction of the MoBEL model, which concurrently learns joint audio-visual features by simulating the intricate emotional processing observed in the brain's limbic system.

For visual modality representation, a 3D-CNN model captures spatial-temporal features of visual expressions, while Mel-spectrograms of speech signals undergo processing through a CNN-RNN architecture for spatial-temporal feature extraction in the auditory modality. The proposed high-level feature fusion with the MoBEL network aims to exploit the correlation between visual and auditory modalities, thereby enhancing the accuracy of emotion recognition that detects 7 basic emotions with around 90% accuracy.

Our model training was conducted on the Zenodo RAVDESS dataset. Through this work, we contribute to the ongoing discourse on the challenges of multimodal emotion recognition and emphasize the critical importance of effective feature extraction and integration, especially when considering spatial-temporal correlations.

References:

[1] Farhoudi, Z., & Setayeshi, S., (2021). Fusion of deep learning features with mixture of brain emotional Learning for audio-visual emotion recognition. Speech Communication, 127, 92-103.

A. Emily

Dr. Emily Jenifer A (AP III / SoC)