Abstract

Speech Emotion Recognition (SER) has been an active area of research ever since the need for smooth and natural Human-Computer Interaction (HCI) came into play. This thesis aims to develop an SER system based on an amalgamation of Tensor Factorization and Neural Network-based learning to mitigate several issues while using contemporary deep learning architectures. This, in turn, is helpful towards recognizing the mental health issues such as depression, anxiety, etc., from speech signals as it is shown in the literature that mental health and emotions are highly correlated. As such, this thesis tries to provide techniques to incorporate emotional information to assess mental health conditions from speech signals, thereby helping the psychologists assign a depression score to patients based on their experience and machine-generated score, thereby mitigating any human bias which might creep in human-only situations.

In the first work, several tensor-based architectures are explored for the task of Speech Emotion Recognition. A tensor Attention Layer is proposed, which helps to focus on class-specific regions of the speech mel-spectrograms and provides emotion-focused inputs to the Tensor Factorized Neural Network. A 3D AG-TFNN is proposed to leverage multi-dimensional information from 3D mel-spectrogram tensors, incorporating delta and double-delta information along the third mode of the input tensor. A parallel AG-TFNN is proposed to leverage complementary information from mel-spectrograms and modulation spectrograms and fused using a 3D tensor. Experimental evaluation on the state-of-the-art datasets such as Emo-DB and IEMOCAP demonstrates the effectiveness of the proposed approaches over the baseline CNN+LSTM architecture, with the added advantage of less computational complexity and a simpler architecture.

The second work delves into the domain of multi-cultural SER by focusing on the two aspects - universality and cultural specificity of emotions. Thus, we propose two methods, one incorporating cultural specificity and another demonstrating the universal nature of emotions across cultures. In the first method, we propose a novel technique to make a multi-cultural SER by incorporating impactful factors such as speaker and language as markers of cultural distinctiveness. We develop a language and a speaker model to get language and speaker embeddings, and a multi-modal fusion architecture is proposed to fuse the information along with emotional cues. Moreover, in the second method, a triplet-loss-based multi-cultural SER is proposed, which tries to normalize speaker and cultural variabilities and focuses on learning emotions, irrespective of culture. Experiments conducted on a collection of five language emotion datasets show the proposed technique's robustness

in predicting emotions in a leave-one-language-out setting. The system's design allows for incorporating a new language and speaker without needing to retrain the whole system again.

In the third work, we proposed a tensor-based architecture for the task of Multiple Instance Learning when a bag of utterances for a speaker is available, and inferences about the speaker label must be drawn using the bag of utterances. The conventional MIL architectures, such as the baseline CNN-MIL system, suffer from the inherent drawbacks of not considering relative and shared information across the utterances in a bag. These techniques rely on inferring labels for individual utterances and averaging or max-pooling the labels to infer the speaker-level labels. The tensor-based architectures solve this problem by considering the utterances as the third mode in addition to the time and frequency modes in speech spectrograms. As such, TFNNs, by their rich mathematical framework, try to capture the shared information across the utterances of a bag by tensor factorization where the input tensor is projected over three subspaces - time subspace, frequency subspace, and utterance subspace. This helps to utilize the shared information and generate a single speaker/bag level probability for the specified task. To this end, we proposed two tensor MIL architectures - 3D TFNN and 3D TFNN+Attention. Comparison with the state-of-the-art proves that both the proposed techniques effectively capture depression-related information across bags of utterances. Moreover, additional analysis on the optimal number of utterances per bag is also presented to shed light on the model performance when using varying bag sizes.

In the last work, we propose emotion information fusion using Tensor-based fusion approaches for depression classification. Since emotions in speech are highly affected due to an individual's underlying mental health issues, it becomes highly relevant when it comes to automatic assessment of clinical depression using speech. Two fusion approach is explored - Inner-Product based fusion and Elementwise Weighting. The emotion embedding tensors to be fused are generated using pre-trained TFNNs on six English SER datasets since the Depression dataset is also English. Moreover, a multimodal approach is also explored using audio and text modalities. BERT-based embeddings are explored for text transcripts and fused with audio embeddings learned from mel-spectrogram representations. Two fusion approaches are explored - Late feature fusion and Score Fusion.

The major contributions of the current thesis are as follows:

- A tensor attention layer to provide emotion focused tensor inputs to TFNN.
- Parallel AG-TFNN to leverage complementary information from two

speech representations.

- A Multi-lingual Emotion classification system that incorporates language and speaker embeddings along with emotion embeddings to adapt to new culture and speakers.
- A triplet-loss based multi-lingual architecture which tries to normalize language and speaker variabilities, thereby providing a more general solution to cross-cultural adaptivity of SER systems.
- A Tensor factorization based Multiple Instance Learning (MIL) architecture along with utterance level attention and statistics pooling for Depression classification from speech signals.
- Emotion information fusion to aid depression diagnosis using two tensor-fusion approaches - weighted fusion and inner-product-based fusion. Multi-task and Multi-modal architectures are also explored to exploit text-based sentiment embeddings to aid depression diagnosis.

Keywords: Speech Emotion Recognition, Deep Learning, Tensor Factorization, Mental Health, Depression Diagnosis, Multi-cultural, Fusion, Multi-modal, Multi-task