End-to-end Triplet Loss based Emotion Embedding System for Speech Emotion Recognition

Puneet Kumar[†], Sidharth Jain[†], Balasubramanian Raman[†], Partha Pratim Roy[†] and Masakazu Iwamura[‡]

[†]Indian Institute of Technology, Roorkee, India; [‡]Osaka Prefecture University, Sakai, Japan

Abstract

In this paper, an end-to-end neural embedding system based on triplet loss and residual learning has been proposed for speech emotion recognition. The proposed system learns the embeddings from the emotional information of the speech utterances. The learned embeddings are used to recognize the emotions portrayed by given speech samples of various lengths. The proposed system implements Residual Neural Network architecture. It is trained using softmax pre-training and triplet loss function. The weights between the fully connected and embedding layers of the trained network are used to calculate the embedding values. The embedding representations of various emotions are mapped onto a hyperplane, and the angles among them are computed using the cosine similarity. These angles are utilized to classify a new speech sample into its appropriate emotion class. The proposed system has demonstrated 91.67% and 64.44% accuracy while recognizing emotions for RAVDESS and IEMOCAP dataset, respectively.

1. Introduction

- The proposed system is an end-to-end approach based on Residual Neural Network (ResNet). The model is trained using softmax pre-training and triplet loss function.
- The embedding values are calculated from the weights between the fully connected and embedding layers of the trained network. The embeddings are mapped onto a hyperplane, and angles among them are calculated and used to identify various emotions.

2. Contribution

The contributions of the paper are as follows.

- A deep neural end-to-end SER system based on triplet loss and residual learning has been proposed. The proposed system is capable of learning emotion-related information from a labeled emotional speech dataset in the form of embeddings.
- The embeddings learned by the proposed system are used to classify the speech samples of various lengths into appropriate emotion classes. Using the embeddings, the proposed system can estimate the emotions in unseen speech utterances.

3. Proposed Methodology

The proposed end-to-end system learns the embedding representations from emotional speech and uses them for speech emotion recognition. Various phases of the system are described below and represented in Fig. 1.

Phase I: Initialization and Pre-processing - The embeddings are initialized and data is pre-processed.

Phase II: Embedding Training - A fully connected layer projects the utterance-level representations as embeddings. Emotion characteristics of the speech are learned by training embedding vectors for each emotion. The training process first carries out the softmax-pretraining and then performs embedding training with Triplet Loss.

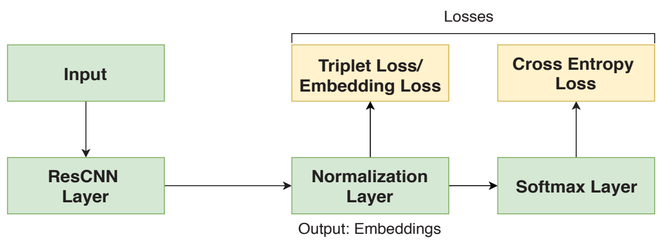


Fig. 1: Proposed methodology.

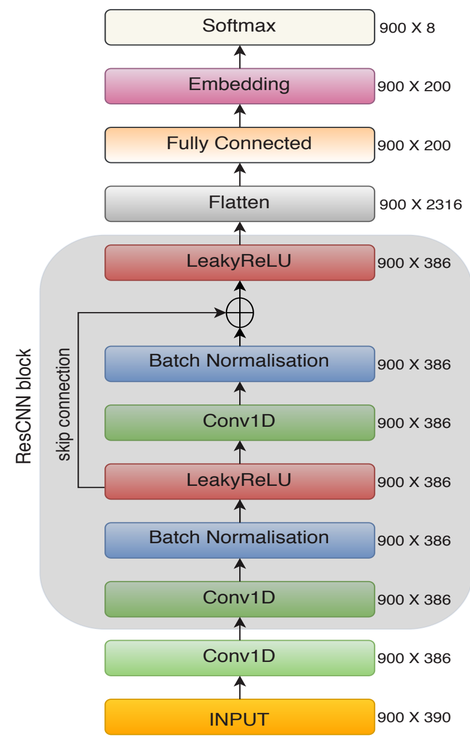


Fig. 2: Proposed architecture.

4. Implementation

- **Experimental Setup**: Model training has been performed on Nvidia RTX 2070 GPU with 2304 CUDA cores, 288 Tensor cores, and 8 GB Virtual RAM. Model testing has been carried out on Intel(R) Core(TM) i7-7700, 3.70 GHz, 16GB RAM CPU system with Ubuntu 18.04.
- **Dataset and Training Strategy**: SER experiments have been performed on *RAVDESS* and *IEMOCAP* datasets. Final implementation has been carried out using 70%-30% training-testing split and 10-fold cross-validation.

5. Ablation Study

The ablation study to choose the appropriate network architecture has been performed. Fully Connected (FC) network, Convolutional Neural Network (CNN), Residual Neural Network (ResNet), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) based RNN, and Gated Recurrent Unit (GRU) based RNN has been evaluated. Their details have been presented in Table 1. Here 'x' represents the total number of layers. The analysis is performed for x = 6 to 15. ResNet, as described in Fig. 9, performed best among these networks. It has been chosen as a suitable architecture for the proposed implementation.

Table 1: Summary of the ablation study

Architecture	Details	x	Accuracy
FC-x	Fully Connected network with x layers	7	47.22%
CNN-x	Convolutional Neural Network with x layers	8	58.33%
RNN-x	Simple x-layered Recurrent Neural Network	8	55.56%
LSTM-x	x-layered RNN with LSTM units	7	56.94%
GRU-x	x-layered RNN with GRU units	8	54.16%
ResNet-x	Residual Neural Network with x layers	11	**61.11%**

6. Results

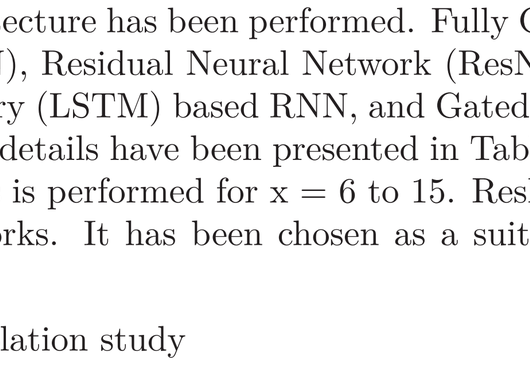
The results of the proposed system in terms of speech emotion classification and learned embeddings have been described as follows.

Table 2: Result comparison for RAVDESS

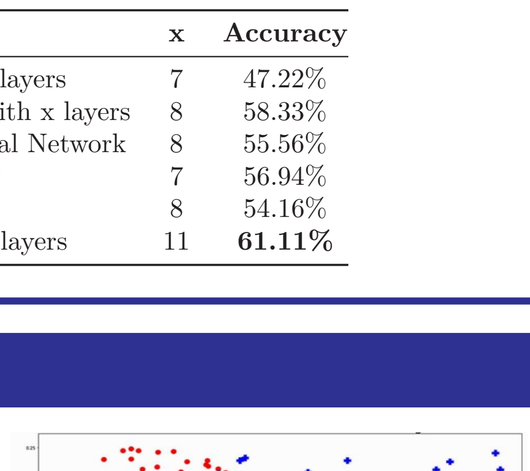
Method	Author	Accuracy
Proposed Method		**91.67%**
Convolutional Neural Network	M. G. Pinto	91.53%
Artificial Neural Network	K. Tomba et al.	89.16%
Multi Task Hierarichel SVM	B. Zhang et al.	83.15%
Bagged Ensemble of SVMs	A. Bhavan et al.	75.69%
Convolutional Neural Network	D. Issa et al.	71.61%

Table 3: Result Comparison for IEMOCAP

Method	Author	Accuracy
RNN + Attention	N. Majunder	64.50%
Proposed Method		**64.44%**
Memory Network	D. Hazarika et al.	63.50%
CNN + Mel Filterbanks	Z. Aldeneh and E. Provost	61.80%
Memory Network	S. Poria et al.	56.13%
CNN + LSTM	J. Zhao	52.14%



(a) RAVDESS dataset



(b) IEMOCAP dataset

Fig. 3: Visualization of Emotion Embeddings

7. Conclusion and Future Work

An end-to-end emotion embedding system has been proposed to learn the emotional patterns from speech in the form of an embedding matrix. The emotion embedding matrix thus prepared has been used for speech emotion recognition, and it demonstrated comparable recognition results to the state-of-the-art methods.

It is required to check the angles for each speech utterance with each emotion class. This process can be optimized to reduce computational requirements. It is also aimed to use the learned embeddings for other speech processing tasks such as emotional speech synthesis.