INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Speech Emotion Recognition Framework for Better Discrimination of Confusions

Jiawang Liu, Haoxiang Wang

School of Computer Science and Engineering, South China University of Technology, China
202021044688@mail.scut.edu.cn, hxwang@scut.edu.cn

Abstract

Speech emotion recognition (SER) plays an important role in human-machine interaction (HMI). Various methods have been proposed for the SER task. However, a common problem in most of the previous studies is some specific emotions are grossly misclassified. In this paper, we propose a novel SER framework aiming at discriminating the confusions by utilizing triplet loss and data augmentation to enforce a CNN-LSTM model to emphasize more on these emotions which are hard to be correctly classified. Ablation experiments demonstrate the effectiveness of the proposed framework. On Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, our framework can achieve 79.52% of Weighted Accuracy (WA) and 78.30% of Unweighted Accuracy (UA). Compared to the other state-of-the-art models, our framework obtains more than 3.34% and 1.94% improvement on WA and UA respectively.
Index Terms: Speech Emotion Recognition, Triplet Loss, Data Augmentation

1. Introduction

Emotion intelligence is an essential part of human intelligence and plays a crucial role in human-computer interaction. As an important medium for human emotions, speech not only conveys non-linguistic information but also expresses emotions. Recognizing the emotions expressed in a speech signal is challenging work. The subjectivity in perceiving and defining an emotion causes confusion among some specific emotions. Besides, the deficiency of labeled emotion datasets also limits the development in SER.

Various methods have been proposed for the SER task. Traditional methods are based on acoustic feature extraction from speech utterances, such as pitch, jitter, Mel-Frequency Cepstral Coefficients (MFCCs), etc., and then these features are fed into classifiers such as Gaussian Mixture Model (GMM) [1], Hidden Markov Model (HMM) [2], Support Vector Machine (SVM) [3]. With the improvement of computing power, deep learning methods become mainstream, providing the excellent ability for high-level discriminative features capturing. In [4], RNN is used to learn the segment-level short-term acoustic features. Later on, LSTM is used as the mainstream RNN model for time-series feature modeling [5]. Meanwhile, as a classical deep learning method, CNN has been widely used in image processing. A real-time CNN based model is proposed to detect emotion from speech [6]. Moreover, a combination of CNN and RNN also becomes popular for SER, where the CNN is designed for frequency information extraction and RNN is used for learning long-term dependencies [7].

Recent SER works investigate architectures like attention, Transformer, transfer learning, and graph neural network (GNN). In [8], the authors propose self-attention for seeking the salient periods of emotion in speech, and gender classification is used as an auxiliary task for multitask learning. In [9], the authors investigate the effectiveness of the self-attention mechanism of Transformer. In another work, the authors propose to adopt a deep graph method for SER, which treats speech signal as a line or a cycle graph [10].

However, a common problem in most of the previous studies is some specific emotions are grossly misclassified. From the confusion matrices presented in [7, 11, 12] on IEMOCAP dataset, we can see there pervasively exists a common problem that most utterances with the happy label are misclassified as neutral. We speculate that this may be due to a similar activation level of neutral and happy, and the subtle difference cannot be captured by the models. Another cause of the confusion can be the imbalance in data samples. Generally, in an imbalanced dataset, the predicted labels prefer to majority classes. For example, in some previous research on the IEMOCAP dataset, happy emotional utterances are easily mistaken for neutral due to the bias in data samples [9, 11].

In this paper, we propose an SER framework for better discrimination of confusion by integration of triplet loss and data augmentation with a CNN-LSTM network. The contributions of this article are summarized as follows:

1. By using a joint loss of triplet loss and cross-entropy loss with the CNN-LSTM model in our framework, more discriminative emotional embeddings can be obtained. Extensive experiments have shown that the joint use of triplet loss and cross-entropy loss can reduce misclassification compared to the previous methods.
2. To overcome the imbalance of training dataset and improve the effectiveness of triplet loss, we use a data augmentation method in our framework to generate artificial MFCCs features for the minority emotional classes. Several comparative experiments are conducted to demonstrate the effectiveness of data augmentation.
3. Experimental results indicate that the proposed framework can reach 77.85% of WA and 76.78% of UA on the IEMOCAP dataset and can be further improved by data augmentation to 79.52% of WA and 78.30% of UA, which is state-of-the-art.

2. Proposed Framework

In this paper, we introduce triplet loss and data augmentation into an SER framework which is based on the CNN-LSTM model. Figure 1 shows an overview of the proposed framework. In the preprocessing phase, we extract MFCCs from speech segments, and data augmentation technique is applied on MFCCs to generate artificial samples for the minority emotional classes. Augmented MFCCs are then fed into the CNN-LSTM model, in which CNN is used for spatial feature extraction and a bi-directional LSTM is used to capture emotion salient part in time

http://dx.doi.org/10.21437/Interspeech.2021-718

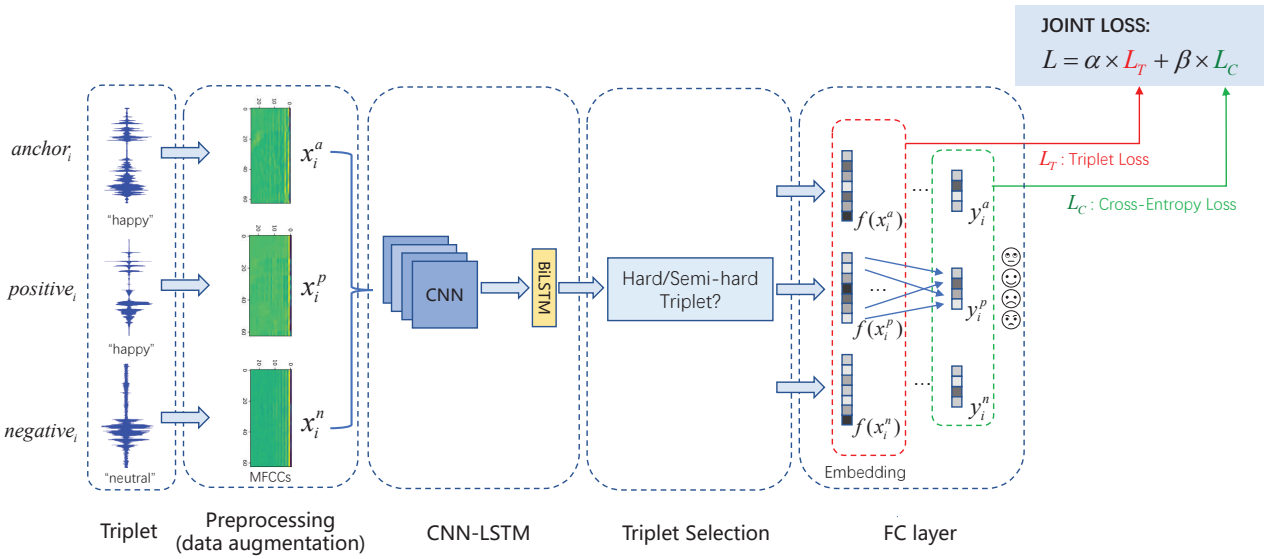

Figure 1: *Illustration of the proposed framework for SER, including four parts: preprocessing, CNN-LSTM network, triplet selection and FC layer.*

series. In the framework, Triplet selection aims to select hard and semi-hard triplets for loss calculation. Selecting only hard and semi-hard triplets can encourage the model to focus on indistinguishable samples and reduce computational complexity. Finally, a joint loss of triplet loss and cross-entropy loss is used for backpropagation.

2.1. CNN-LSTM

CNN-LSTM model is widely used in SER task for its excellent performance of high-level features extraction. In our proposed framework, CNN is used for spatial feature extraction and bi-directional LSTM is used for temporal feature extraction. In our experiment, four stacked convolution layers are used and batch normalization is performed after each convolution layer. The filter size of the first three convolution layers is 3x3 with a padding of 1 and we perform a convolution over frequency on the last convolution layer. The sequence features produced from the convolution layers are then fed into the bi-directional LSTM network with 32 cells per direction. The details about every convolution layer are shown in Table 1.

Table 1: *Details about convolution layers. Four stacked convolution layers are used in our experiment.*

layer	kernel_size	in_channels	out_channels	padding
Conv1	(3,3)	1	8	1
Conv2	(3,3)	8	16	1
Conv3	(3,3)	16	32	1
Conv4	(26,1)	32	32	0

2.2. Joint Loss

The triplet loss minimizes the distance between the anchor and positive, both of which have the same emotion label, and maximizes the distance between the anchor and negative, which have different labels. The purpose of training is to make an anchor utterance x_i^a more similar to all other positive utterances x_i^p than it is to any negative utterance x_i^n. The final embedding space is expected to follow:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + m < \|f(x_i^a) - f(x_i^n)\|_2^2 \quad (1)$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \Gamma$$

where m denotes margin which is enforced between positive and negative. Γ is the set of all possible triplets and has cardinality N. $(f(x_i^a), f(x_i^p), f(x_i^n))$ is defined as a triplet. Triplet loss L_T is given by:

$$L_T = \frac{1}{N} \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + m \right] \quad (2)$$

Triplet loss and cross-entropy loss are simultaneously used for model training with different weights α and $(1-\alpha)$, where cross-entropy loss includes emotion label information. The final loss L is calculated by:

$$L = \alpha \times L_T + (1-\alpha) \times L_C \quad (3)$$

where L_C presents the cross-entropy loss as defined in:

$$L_C = -\sum_{k=1}^{K} \hat{y}_k \log(y_k) \quad (4)$$

$\hat{y} \in \{0,1\}^K$ is a one-of-K label vector and y denotes the probability output of the softmax layer.

2.3. Triplet Selection

Even a small dataset can generate a large number of triplet samples. But most triplets will not contribute to training, and rather cause slower convergence. Therefore, the triplet selection strategy is important to model performance. We expect to acquire productive triplets for discriminating emotion classes with less time consuming.

In our experiment, we start by randomly selecting anchor, positive, and negative samples to form triplets, with a total of 50,000 triplets. In the training process of a mini-batch, semi-hard triplets that follow:

$$\begin{aligned} \|f(x_i^a) - f(x_i^p)\|_2^2 &< \|f(x_i^a) - f(x_i^n)\|_2^2 \\ &< \|f(x_i^a) - f(x_i^p)\|_2^2 + m \end{aligned} \tag{5}$$

and hard triplets that follow:

$$\|f(x_i^a) - f(x_i^n)\|_2^2 < \|f(x_i^a) - f(x_i^p)\|_2^2 \tag{6}$$

are selected for computing the joint loss. The selection of semi-hard and hard triplets makes the training focus on indistinguishable samples and reduces computational complexity.

2.4. Data Augmentation

As illustrated in Figure 2, in the IEMOCAP dataset, the distribution of samples is imbalanced. Besides, the imbalance of triplet distribution is also observed in the triplet selection. The triplets which consist of neutral sample are far more than the triplets consisting of angry sample. Based on the hypothesis that the imbalance of triplet distribution might be caused by the imbalance in training data samples, a data augmentation method named SpecAugment [13] is used to enrich the minority emotion classes. SpecAugment is a simple but effective data augmentation method for automatic speech recognition (ASR) proposed by Google Brain. By applying time warping, frequency masking, and time masking on the log mel spectrogram, the SpecAugment helps the network learn robust features for the transformation of time direction, and fractional loss in frequency domain and time domain. Because of the similarity between ASR and SER task, we use SpecAugment for MFCCs generation. The augmented numbers of all classes are equal to the most numerous category in the original samples.

3. Experiment

3.1. Dataset

To evaluate the proposed framework, we conduct our experiments on the IEMOCAP dataset [14]. It comprises 12 hours of data collected over five sessions. Each session is composed of two speakers (female and male) in scripted and improvised scenarios. The dataset contains 10039 utterances annotated by at least three expert evaluators with one of the following classes: happy, angry, neutral, sad, surprise, excited, fear, frustrated, disgust and others. To be consistent with previous studies [9, 15], our experiment considers four emotions: happy, angry, sad, and neutral selected only from the improvised utterances with excitement replacing happy class.

3.2. Experimental setup

We employ a 5-fold cross-validation technique in our evaluations. 80% of the dataset are selected randomly as the training dataset and the remaining 20% as the test dataset. In the train set, an utterance is segmented into segments with an overlap of one second between each segment, which with a length of 2 seconds. In the test set, the length of overlap is set to 1.6 seconds, which is consistent with [15]. To extract MFCCs, windowing is used on the raw wave with a small window of length 40ms and a small step size of length 10ms. In the test process, we get the final prediction by averaging the prediction of each segment of a source utterance. We implement the proposed model by using the PyTorch deep learning framework. Adam optimizer is chosen with the initial learning rate of 1e-4 and weight decay of 1e-6. We conduct the experiments on a GTX 1080Ti GPU, with the batch size set as 32. WA and UA are calculated for evaluation. To reduce the error, we average the accuracy using five different random seeds for each parameter setting.

The experiment can be divided into two parts. In the first part, comparison experiments are performed on the IEMOCAP dataset to evaluate the effectiveness of triplet loss and data augmentation. Different weights are associated with triplet loss and entropy-loss to explore the optimized joint loss. In the second part, we evaluate our experimental results compared with other state-of-the-art works.

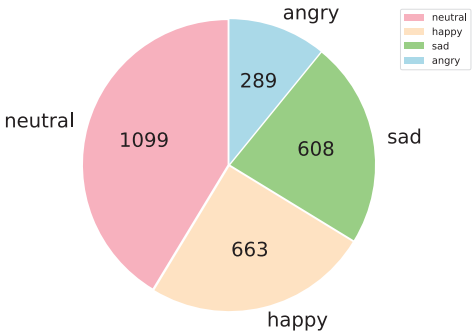

Figure 2: *Distribution of the four emotions used in experiment*

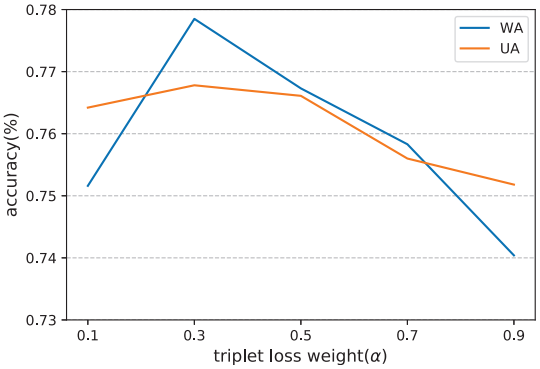

Figure 3: *The impact of different weights between triplet loss and cross-entropy loss.* α *is the triplet loss weight and* $(1-\alpha)$ *is the cross-entropy loss weight*

3.3. Impact of triplet loss

Figure 3 shows the WA and UA evaluated on the different weights between triplet loss and cross-entropy loss. When the triplet loss weight is set to 0.3, we obtain the highest WA and competitive UA.

Table 2: *Impact of triplet loss*

Methods	WA(%)	UA(%)
CNN-LSTM(no triplet)	77.38	75.84
CNN-LSTM(with triplet)	77.85	76.78

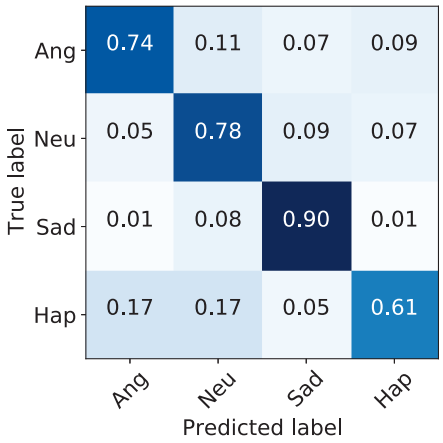

Figure 4: *The confusion matrix of training without triplet loss*

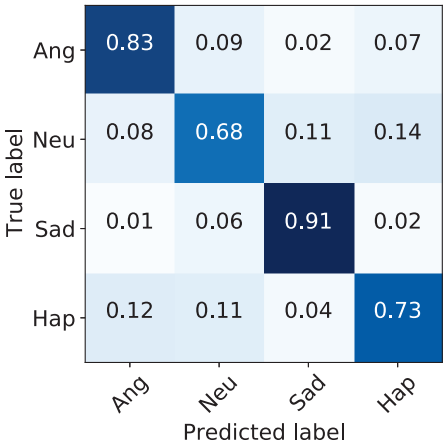

Figure 5: *The confusion matrix of training with triplet loss*

In Table 2, we present the comparison between training with triplet loss and without triplet loss. As we can observe, the model with triplet loss outperforms that without triplet loss and obtains an absolute improvement of 0.47% and 0.94% for WA and UA, respectively. It should be noted that the results of the first line are obtained directly on CNN-LSTM model without triplet generation and triplet selection process. Further analysis of the confusion matrix shown in Figure 4 and Figure 5 reveals that the utilization of triplet loss reduces confusion among some classes which are difficult to distinguish. As we can see, the recognition accuracies of emotions such as angry, sad, and happy have all been improved.

Table 3: *Performance comparison between training with original dataset and training with an augmented dataset*

Data Augmentation	WA(%)	UA(%)
CNN-LSTM(no augmentation)	77.85	76.78
CNN-LSTM(data augmentation)	79.52	78.30

We also investigate the effectiveness of the SpecAugment data augmentation method. A comparison study is conducted on the proposed framework with triplet loss used. Compared with the model without data augmentation, we can observe that the model trained with the enhanced dataset further improves WA and UA's performance by 1.67% and 1.52%, respectively, as shown in Table 3. It proves our hypothesis that the augmented and balanced dataset improves the effectiveness of triplet loss and the generalization of the model.

As the ablation experiment results suggest, triplet loss and data augmentation adopted in our proposed framework both contribute to the improvement of the accuracy by distinguishing easily confused emotion classes.

3.4. Comparison with Other Works

Table 4: *Comparison of the accuracy of previous research and our model.*

Methods	WA(%)	UA(%)
CNN-LSTM [16]	67.30	62.00
3-D ACRNN [7]	-	64.74
CNN-Attn [17]	71.75	68.06
Self-Attn [9]	70.17	70.85
ACNN [15]	76.18	76.36
Our proposed	**79.52**	**78.30**

Table 4 summarizes the performance of our framework and some recent state-of-the-art SER networks. On the IEMOCAP dataset, our proposed triplet loss method achieves state-of-the-art results of 79.52% on WA and 78.30% on UA. Compared with previous attention methods [15], we obtain improvement by about 3% on WA and 2% on UA, giving the credit to the excellent discrimination ability of triplet loss and data augmentation.

4. Conclusion

This paper proposes an SER framework using triplet loss and data augmentation for better discrimination of confusions in emotional classification. Experiments conducted on the IEMOCAP dataset show that the proposed framework achieves outstanding results superior to other state-of-the-art models. In future work, artificial data enhancement can be further explored. To compare with SpecAugment adopted in current experiments, we can try different GANs for data augmentation to generate more emotional utterances for model training.

5. Acknowledgements

This work was supported by Guangdong Basic and Applied Basic Research Foundation, 2021A1515011852; The Fundamental Research Funds for the Central Universities, x2js-D2190680.

6. References

[1] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. Interspeech*, 2006, pp. 809–812.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[3] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 864–867.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, 2014, pp. 223–227.

[5] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. Interspeech*, 2018, pp. 272–276.

[6] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, 2017, pp. 5115–5119.

[7] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, oct 2018.

[8] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, 2019, pp. 2803–2807.

[9] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2578–2582.

[10] A. Shirian and T. Guha, "Compact graph architecture for speech emotion recognition," 2020,*arXiv:2008.02063*.

[11] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech*, 2018, pp. 3683–3687.

[12] T. Fujioka, T. Homma, and K. Nagamatsu, "Meta-learning for speech emotion recognition considering ambiguity of emotion labels," in *Proc. Interspeech*, 2020, pp. 2332–2336.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, dec 2008.

[15] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. Annu. Comput. Commun. Workshop Conf.*, 2020, pp. 1058–1064.

[16] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, 2017, pp. 1089–1093.

[17] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. Interspeech*, 2018, pp. 3087–3091.