Learning Discriminative Features from Spectrograms using Center Loss for Speech Emotion Recognition

Dongyang Dai[1,2], *Zhiyong Wu*[1,2,3], *Runnan Li*[1,2], *Xixin Wu*[3], *Jia Jia*[1,2], *Helen Meng*[1,3]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

[2] Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing, China

[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

1. Introduction

Motivation

- Identify the emotional state from speech

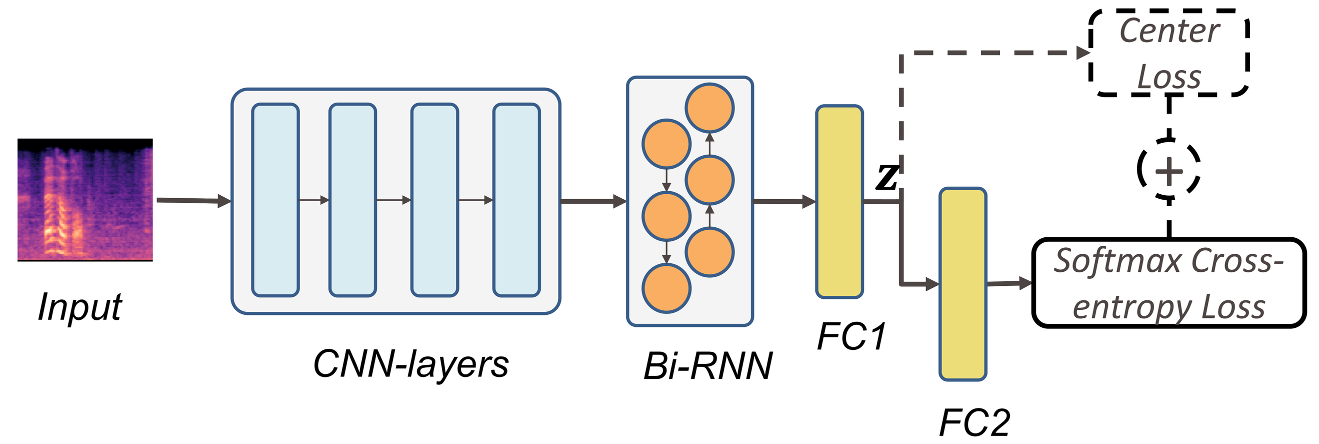
Challenge

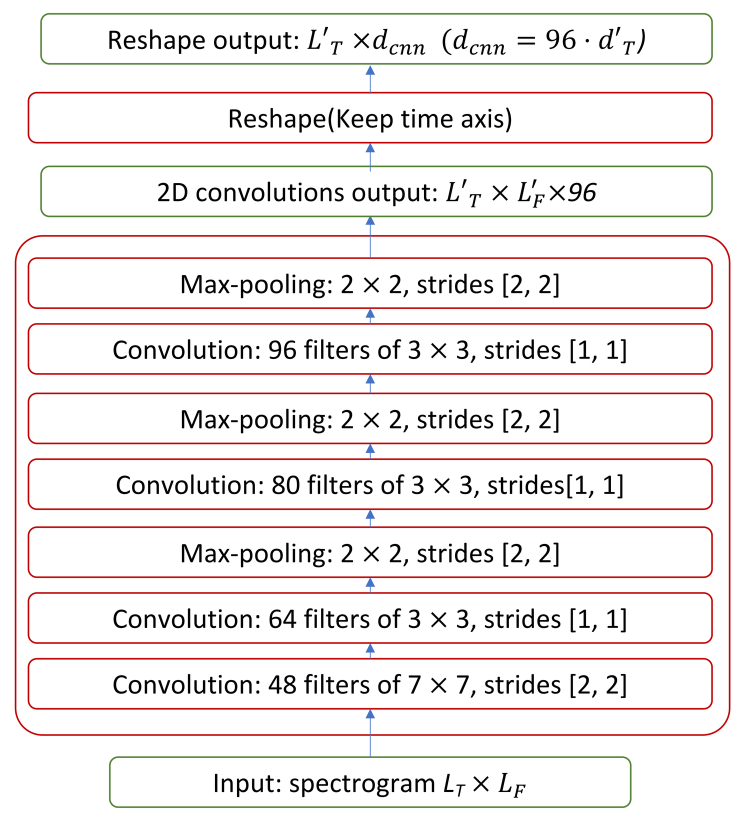
- Emotions are naturally ambiguous
- How to extract features containing enough emotional information

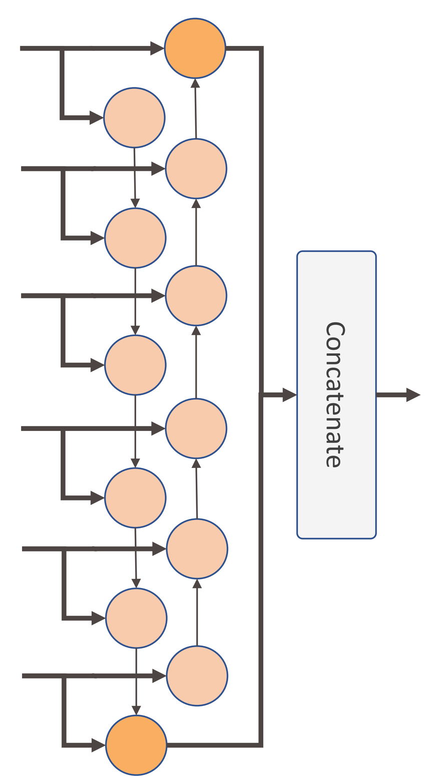
Contribution

- Introduce center loss together with softmax corss-entropy loss in SER task to learn discriminative features
- Extract features and identify emotions directly from spectrograms

2. Proposed Method







Model Architecture

- Input: variable length spectrograms (STFT or Mel-spectrograms)
- CNN layers: extract spatial information
- Bi-RNN: compresses the variable length sequence down to a fixed-length vector
- FC1: output $z \in R^d$ as the learned feature and calculate center loss according to z
- FC2: outputs posterior class probabilities, used to calculate softmax cross-entropy loss
- Softmax Cross-enctropy Loss: enables the network to learn separable features
- Center Loss: pulls the features belonging to the same emotion category to their center

Softmax Cross-entropy Loss

$$L_s = -\frac{1}{\sum_{i=1}^{m} \omega_{y_i}} \sum_{i=1}^{m} \omega_{y_i} log(\frac{e^{W_{y_i}^{T} z_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T z_i + b_j}})$$

- ω_j: in inverse proportion to the sample number of the j-th class in training set

Center Loss

$$L_c = \frac{1}{\sum_{i=1}^{m} \omega_{y_i}} \sum_{i=1}^{m} \omega_{y_i} ||z_i - c_{y_i}||^2$$

$$c_j^{t+1} = \begin{cases} (1-\alpha)c_j^t + \alpha \dot{c}_j^t & \sum_{i=1}^{m} \delta(y_i = j) > 0 \\ c_j^t & \sum_{i=1}^{m} \delta(y_i = j) = 0 \end{cases}$$

$$\dot{c}_j = \frac{\sum_{i=1}^{m} \delta(y_i = j) z_i}{\sum_{i=1}^{m} \delta(y_i = j)}$$

- L_c: center loss
- c_j: the global class center of features corresponding to the j-th emotion, updated per mini-batch iteration
- $\dot{c}_j$: the j-th class center of features from a mini-batch
- α: controls the update rate of c_j

Joint Loss

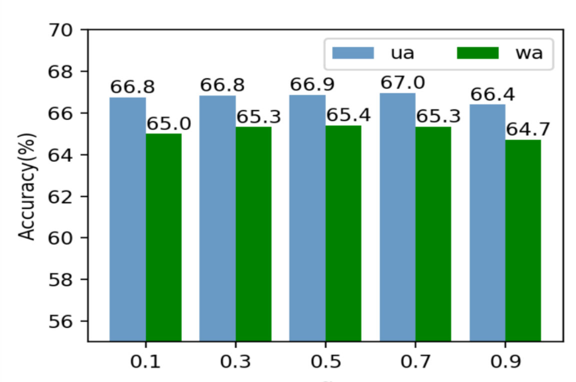
$$L = L_s + \lambda L_c$$

- λ: trades off center loss against softmax cross-entropy loss.

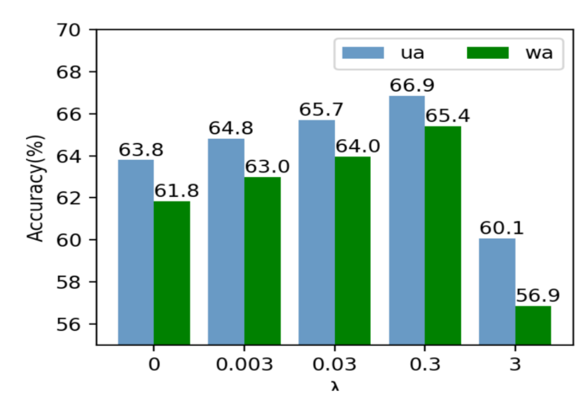
3. Experiments and Results

Experimental Setup

- Data
  - Dataset: IEMOCAP
  - Neutral, angry, happy, sad and excited (merges happy and excited as happy, 5531 utterances)
  - 5 subsets (keep the emotion distribution), 4 subsets for training, half of the last subset as development set and half as test set
- Settings of spectrograms
  - Model input: log scale STFT spectrogram or Mel-spectrogram
  - Hamming window
  - Window size: 40msec
  - Window Shift: 10msec
  - Sample rate: 16KHz
  - DTF length: 1024
  - The number of Mel bands: 128
- Metrics
  - The unweighted accuracy: UA, the mean value of the recall for each class
  - The weighed accuracy: WA, the number of correctly classified samples divided by the total amount of samples



(a)



(b)

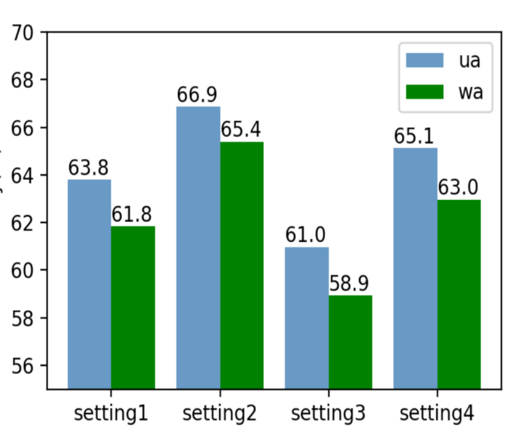


Table 1: confusion matrix on, (a) setting 1, (b) setting 2

	neu	ang	hap	sad
neu	0.575	0.095	0.164	0.166
ang	0.119	0.691	0.155	0.035
hap	0.211	0.162	0.511	0.115
sad	0.138	0.026	0.060	0.776

(a)

	neu	ang	hap	sad
neu	0.637	0.067	0.167	0.127
ang	0.108	0.705	0.167	0.020
hap	0.219	0.131	0.556	0.094
sad	0.128	0.025	0.070	0.777

(b)

Table 2: confusion matrix on, (a) setting 3, (b) setting 4

	neu	ang	hap	sad
neu	0.544	0.093	0.185	0.177
ang	0.127	0.681	0.167	0.025
hap	0.216	0.186	0.476	0.122
sad	0.161	0.039	0.062	0.737

(a)

	neu	ang	hap	sad
neu	0.573	0.073	0.196	0.157
ang	0.103	0.720	0.153	0.022
hap	0.205	0.161	0.518	0.114
sad	0.125	0.028	0.053	0.793

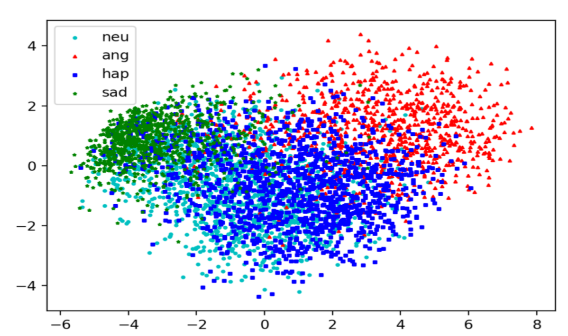
(b)

Experiments

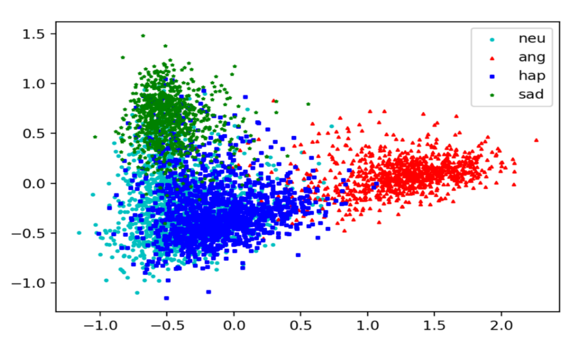
- The effect of hyperparameter α and λ on Mel-spectrogram
  - (a) fixing $\lambda = 0.3$, (b) fixing $\alpha = 0.5$
  - not sensitive to α
  - can be significantly improved with proper value of λ
- Experiments with different λ on Mel and STFT

	Setting1	Setting2	Setting3	Setting4
λ, α	λ=0	λ=0.3, α=0.5	λ=0	λ=0.3, α=0.5
Input	Mel	Mel	STFT	STFT

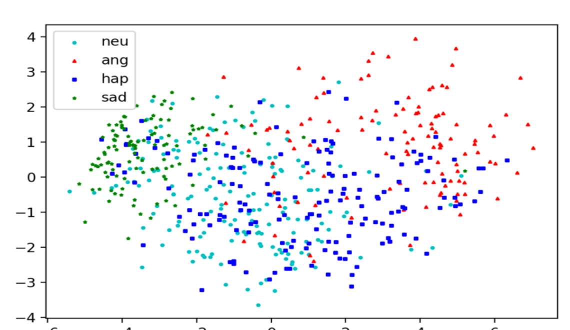
  - The UA and WA on setting 1 ~ setting 4
  - The Confusion matrix on setting 1 ~ setting 4
- PCA embedding of feature z
  - (a) training set on setting 1, (b) training set on setting 2, (c) test set on setting 1, (d) test set on setting 2



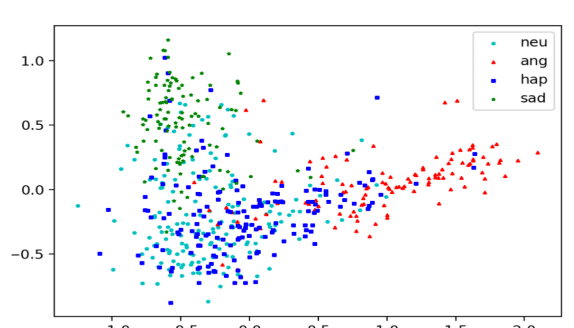
(a)



(b)



(c)



(d)

4. Conclusion

Conclusion

- Introducing center loss with proper λ could effectively improve the SER performance on bath STFT spectrogram and Mel-spectrogram input
- Mel-spectrogram input, reducing the dimension based on human hearing characteristics, over performs STFT spectrogram input
- The 2-D PCA embedding illustrated the discriminative power when using center loss, which enables the neural network to learn more effective features for SER

5. Acknowledgment

- This work is supported by National Natural Science Foundation of China (NSFC) (61433018, 61375027), joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N CUHK404/15) and National Social Science Foundation of China (13&ZD189)