



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 Wu³, Jia Jia^{1,2}, Helen Meng^{1,3}

¹Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
²Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing, China
³Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong



1. Introduction

Motivation

- Identify emotions directly from raw data (spectrograms), getting rid of feature engineering
- Extract discriminative features with larger inter-class variance and smaller intra-class variance to improve performance



Challenge

- How to design suitable model processing variable length spectrograms
- How to propose an appropriate method to extract discriminative features

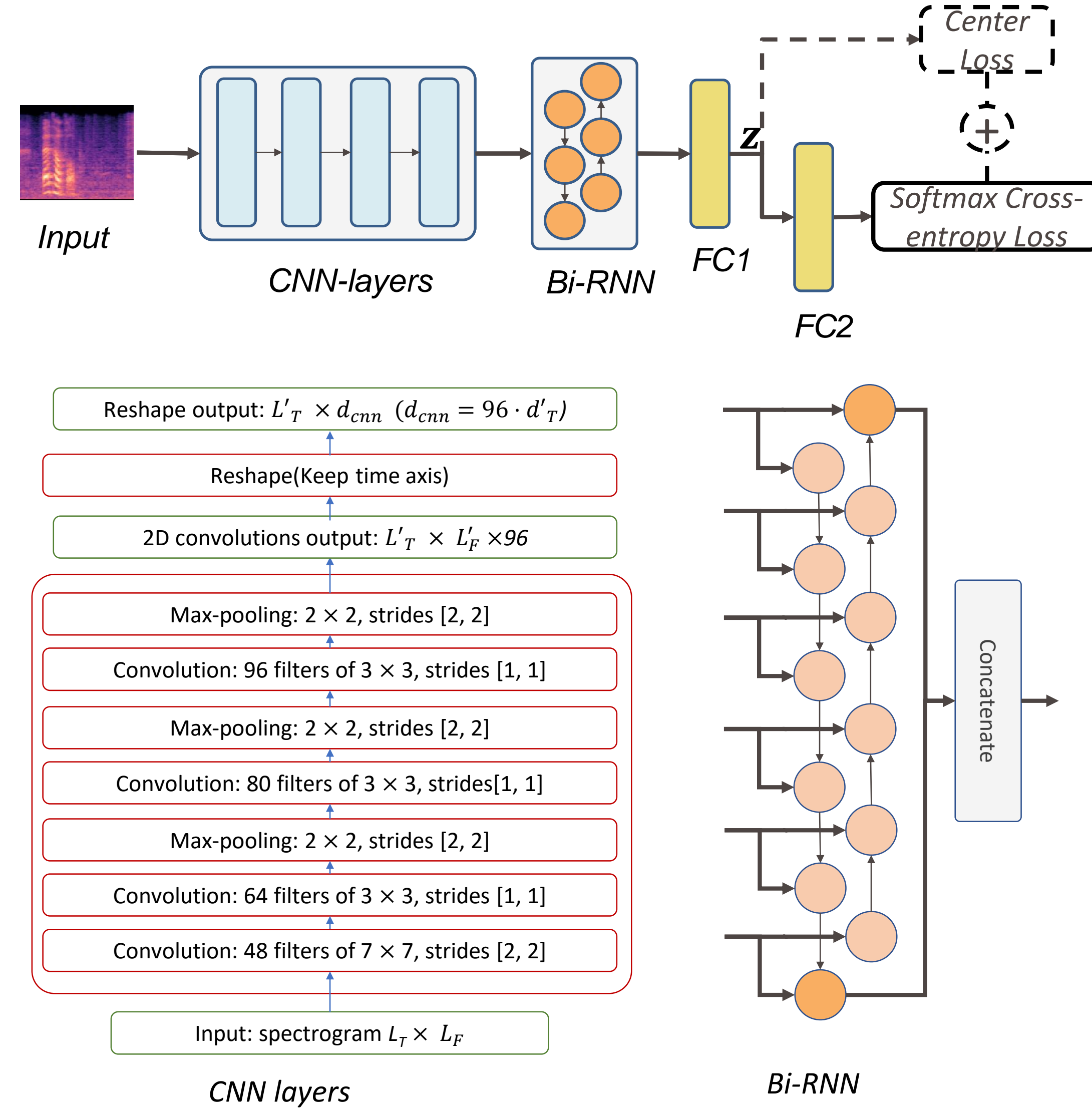
Contribution

- Apply CNN + Bi-RNN to extract features directly from spectrograms
- Introduce center loss together with softmax cross-entropy loss in SER task to learn discriminative features
 - ✓ Separable inter-class features
 - ✓ More compact intra-class features

2. Proposed Method

Model Architecture

- Input:** variable length spectrograms
- CNN layers:** extract spatial information from input, outputs a variable length sequence
- Bi-RNN:** compresses the variable length sequence down to a fixed-length vector, by concatenating the last output of forward RNN and backward RNN
- FC1:** outputs $z \in R^d$ as the learned feature, from which center loss is calculated
- FC2:** outputs posterior class probabilities, from which softmax cross-entropy loss is computed
- Softmax Cross-entropy Loss:** enables the network to learn separable features
- Center Loss:** pulls the features belonging to the same emotion category to their center



Center Loss

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

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

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

Softmax Cross-entropy Loss

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

Joint Loss

$$L = L_s + \lambda L_c$$

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

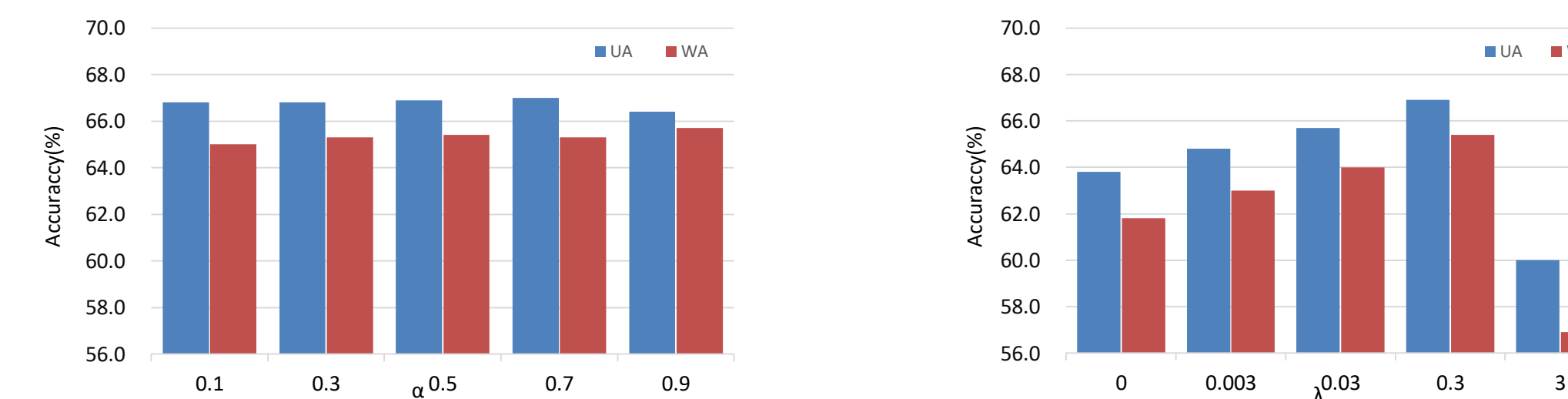
3. Experiments and Results

Experimental Setup

- Dataset:** IEMOCAP
 - ✓ 4 emotion categories: neutral, angry, happy and sad (happy and excited merged as happy)
 - ✓ 5 subsets
 - Randomly divided the total 5531 utterances, but keeping the distribution portion of emotion categories
 - 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: 40ms window length and 10ms shift
 - ✓ Sample rate: 16KHz
 - ✓ DTF length: 1024
 - ✓ The number of Mel bands: 128
- Evaluation metrics**
 - ✓ **Unweighted Accuracy (UA):** the mean value of the recall for each class
 - ✓ **Weighed Accuracy (WA):** the number of correctly classified samples divided by the total amount of samples

Experiments

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

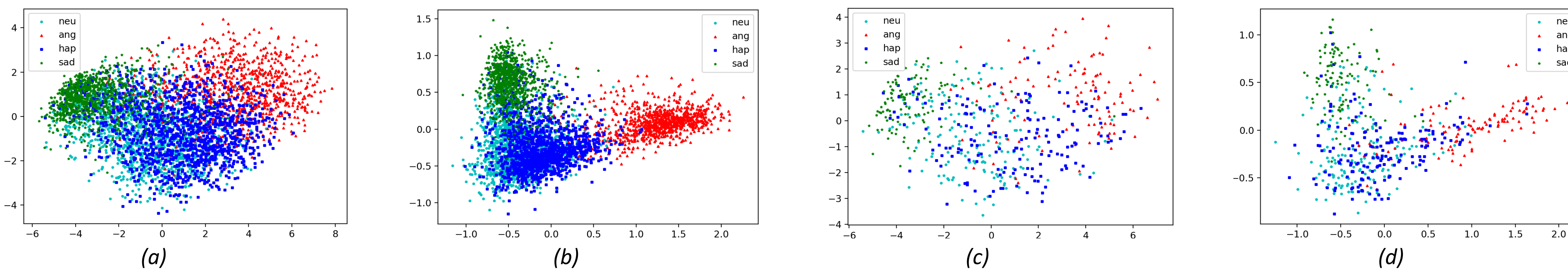


- Experiments with different λ on Mel and STFT

- ✓ The UA and WA on setting 1 ~ setting 4 (%)
- ✓ Confusion matrix on setting1|setting2|setting3|setting4 (%)

	λ, α	Input	UA	WA														
Setting1	$\lambda=0$	Mel	63.80	61.83														
Setting2	$\lambda=0.3, \alpha=0.5$	Mel	66.86	65.40														
Setting3	$\lambda=0$	STFT	60.97	58.93														
Setting4	$\lambda=0.3, \alpha=0.5$	STFT	65.13	62.96														

- 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



4. Conclusion

Conclusion

- Introducing center loss with proper λ could effectively improve the SER performance on both STFT spectrogram and Mel-spectrogram input
- Mel-spectrogram input, reducing the dimension based on human hearing characteristics, outperforms STFT spectrogram input
- The 2-D PCA embedding illustrates the discriminative power of 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)