清3词的逻辑关系明确 ① Motivation应该说明 (1.1) 为什么要基于spectrogram进行SER; (1.2) SER中discriminative features的重要性, 从而引出文章主题 "Learning Discriminative Features".
② Challenge应该强调 (2.1) 如何design model architecture 以接受 variable-length spectrogram 作为输入; (2.2) 强调如何设计model以得到更有区分性的features.
③ Contribution: (3.1) 通过CNN+Bi-LSTM来解决变长spectrogram的问题; (3.2) 通过引入center loss来学习区分性特征.

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

背景图取错啦

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

- Extract valid features from raw data for emotion recognition

Speech → Features → Emotion

强调spectrogram;
强调Discriminative Features

Challenge

- How to design a suitable model processing directly on raw data
- Emotions are naturally ambiguous

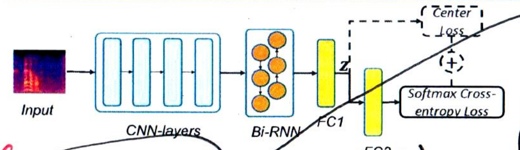
要突出具体的强调

Contribution

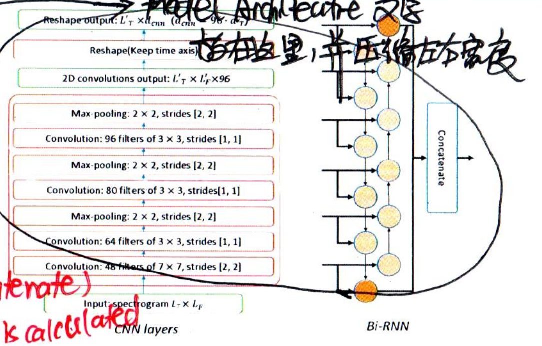
- Extract features and identify emotions directly from spectrograms
- Introduce center loss together with softmax corss-entropy loss in SER task to learn discriminative features
  - ✓ Separable inter-class features
  - ✓ More compact intra-class features

标记统一用"√"号, 并与其他地方字体大小一致

2. Proposed Method



Model Architecture 放这里, 并压缩左右宽度



放左侧

Model Architecture

- ~~Input: variable length spectrograms (STFT or Mel-spectrograms)~~ (这里不用说)
- CNN layers: extract spatial information
- Bi-RNN: compresses the variable length sequence down to a fixed-length vector (强调Bi-RNN的输出如何concatenate)
- FC1: output $z \in R^d$ as the learned feature ~~and calculate center loss according to z~~, from which center loss is calculated
- FC2: outputs posterior class probabilities, ~~used to calculate softmax cross-entropy loss~~, from which softmax ... loss is computed
- 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

① 要先介绍center loss, 所以放在第一

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

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

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

② cross entropy 大家都比较清楚, 放在第二, 不用详细介绍

- ~~L_c: center loss~~
- c_j: the global class center of features corresponding to the j-th emotion class, 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

③ 最后说明应当组合 joint loss.

$$L = L_s + \lambda L_c$$

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

3. Experiments and Results

Experimental Setup

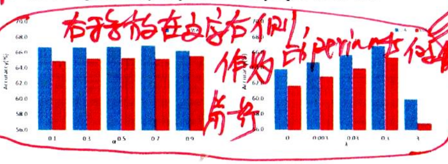
- Dataset: IEMOCAP ✓ 4 Emotion categories: N, A, H, S. (happy and excited are merged as happy)
  - ~~Dataset: IEMOCAP~~
  - ~~Neutral, angry, happy, sad and excited (merges happy and excited as happy, 5531 utterances)~~
  - ✓ 5 subsets ~~(keep the emotion distribution)~~, 4 subsets for training, half 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: 40ms window length and 10ms shift
  - ~~Window size: 40msec~~
  - ~~Window Shift: 10msec~~
  - Sample rate: 16KHz
  - DTF length: 1024
  - The number of Mel bands: 128 for Mel-spectrogram
- ~~Metrics~~ Evaluation Metrics
  - ~~The~~ Unweighted Accuracy (UA): the mean value of the recall for each class
  - ~~The~~ Weighted Accuracy (WA): the number of correctly classified samples divided by the total amount of samples

粗体强调突出

✓ 5 subsets
- randomly divided 5531 utterances, but keeping the distribution portion of emotion categories
- 4 subsets for traing, half ... and half ...

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



两个图放在左下方侧, 作为Experiments的第一行

- 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

	Setting1	Setting2	Setting3	Setting4
UA(%)	63.80	66.86	60.97	65.13
WA(%)	61.83	65.40	58.93	62.96

- Confusion matrix on setting1/setting2/setting3/setting4(%)

	neutral	angry	happy	sad
neutral	57.5/63.7/54.4/57.[illegible]	9.5/6.7/[illegible]/7.3	[illegible]/19.6	[illegible].7
angry	11.9/10.8/12.7/10.3	69.1/70.5/68.1/72.0	15.5/16.7/[illegible]/15.3	3.5/[illegible]
happy	21.1/21.9/21.6/20.5	16.2/13.1/18.6/16.1	51.1/55.6/[illegible]/51.8	11.5/9.4/12.2/[illegible]4
sad	13.8/12.8/16.1/12.5	2.6/2.5/3.9/2.8	6.0/7.0/6.2/5.3	77.6/77.7/[illegible]/79.3

精简 混淆 混淆 混淆 混淆

① 表格使用立体的形式

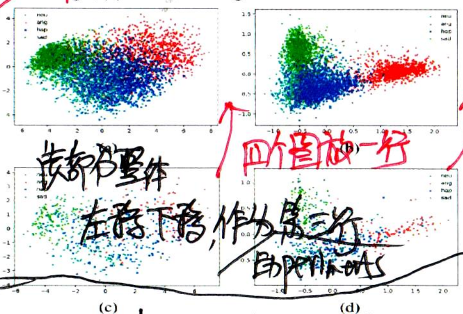
λ, α	input
1	
2	
3	
4	

② 同样设计三行
③ 保留四个部分, 但合为一个表

三个表格放在一行, 作为Experiments的第二行

- 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

放右侧



四个图放一行

去掉那个整体左移下移, 作为第三行 Experiments

4. Conclusion

Conclusion

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

缩减宽度

Experiments 部分分三行呈现结果