Unsupervised Cross-Corpus Speech Emotion Recognition Using Domain-Adaptive Subspace Learning

Na Liu [1,3], Yuan Zong [2], Baofeng Zhang [1], Li Liu [3],Jie Chen [3], Guoying Zhao [3], Junchao Zhu [1]

[1]*Key laboratory of computer vision and systems, Ministry of education, Tianjin University of Technology, China*

[2] *Research Center for Learning Science, Southeast University, China*

[3] *Center for Machine Vision and Signal Analysis, University of Oulu, Finland*

Introduction

- ***Motivation***: In this paper, we investigate unsupervised cross-corpus speech emotion recognition (SER). The training (source) and testing (target) speech signals come from two different corpora which may have different feature distributions and therefore lots of existing SER methods would not work.

- ***Solution***: (1) Construct a label space based on the label information provided by the source speech corpora to serve as the predefined common subspace for Domain-Adaptive Subspace Learning (DoSL). (2) Learn a projection matrix which transforms the source and target speech signals from the original feature space to a common subspace.

Method

Our DoSL aims at learning a projection matrix U to project the source speech feature matrix Xs from the original feature space to such common subspace spanned by the columns of Ls, which can be formulated as the following optimization problem:

$$\min_{\mathbf{U}} \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F \qquad (1)$$

Enforce the projected source and target speech features share the similar distributions

$$\min_{\mathbf{U}} \left\| \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbf{U}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t} \mathbf{U}^T\mathbf{x}_i^t \right\|^2 \qquad (2)$$

Minimizing the combination of the above objective functions in Eqs. (1) and (2)

$$\min_{\mathbf{U}} \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F + \lambda_1 \left\| \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbf{U}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t} \mathbf{U}^T\mathbf{x}_i^t \right\|^2 + \lambda_2\|\mathbf{U}^T\|_{2,1} \qquad (3)$$

Optimization

DoSL model is solved by using inexact augmented Lagrange multiplier (IALM) method. More specifically, by introducing a auxiliary variable Q which satisfies U = Q, we convert the optimization problem of DoSL to a constrained one which can be expressed as:

$$\min_{\mathbf{U}} \|\mathbf{L}^s - \mathbf{Q}^T\mathbf{X}^s\|_F^2 + \lambda_1 \left\| \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbf{Q}^T\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t} \mathbf{Q}^T\mathbf{x}_i^t \right\|^2 + \lambda_2\|\mathbf{U}^T\|_{2,1} \quad \text{s.t. } \mathbf{U} = \mathbf{Q} \qquad (4)$$

Subsequently, the Lagrange function of Eq. (4) can be obtained as follows:

$$\mathrm{L}(\mathbf{U},\mathbf{Q},\mathbf{T},\mu) = \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F + \lambda_1 \left\| \mathbf{Q}^T\bar{\mathbf{X}}_i^{st} \right\|^2 + \lambda_2\|\mathbf{U}^T\|_{2,1} + tr[\mathbf{T}^T(\mathbf{U}-\mathbf{Q})] + \frac{\mu}{2}\|\mathbf{U}-\mathbf{Q}\|_F^2 \qquad (5)$$

where $\bar{\mathbf{X}}_i^{st} = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_i^s - \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbf{x}_i^t$,T is the Lagrange multiplier, and $\mu > 0$ is the regularization parameter.

Iteratively minimize the Lagrange function of Eq. (4) until convergence:

1. Update Q

$$\min_{\mathbf{Q}} \|\mathbf{L}^s - \mathbf{Q}^T\mathbf{X}^s\|_F^2 + \lambda_1 \left\| \mathbf{Q}^T\bar{\mathbf{X}}_i^{st} \right\|^2 + \lambda_2\|\boldsymbol{U}^T\|_{2,1} + tr[\boldsymbol{T}^T(\mathbf{U}-\mathbf{Q})] + \frac{\mu}{2}\|\boldsymbol{U}-\boldsymbol{Q}\|_F^2$$

2. Update U:

$$\min_{\mathbf{U}} \frac{\lambda_2}{\mu}\|\mathbf{U}^T\|_{2,1} + \frac{1}{2}\left\| \mathbf{U}^T - (\mathbf{Q}^T - \frac{\mathbf{T}^T}{\mu}) \right\|_F^2$$

3. Update T and μ

$$\mathbf{T} = \mathbf{T} + \mu(\mathbf{U}-\mathbf{Q}), \mu = \max(\mu_{max}, \rho\mu),$$

4. Check convergence:$\|U - Q\|_\infty < \epsilon$

$$\text{emotion\_labels} = \operatorname*{argmin}_k\{[U_*^T X^t](k,:)\}$$

where$[U_*^T X^t](k,:j)$ means the k^{th} element of the j^{th} column (target speech signal) of the projected matrix $U_*^T X^t$

Results and discussion

Results of the cross-corpus SER experiments in terms of UAR and WAR, where the common emotion states (5 classes) are Angry, Disgust, Fear, Happy and Sad.

#	Source Corpus	Source Corpus	SVM		KMM		KLIEP		uLSIF		DALSR		DoSL	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
1	EmoDB	eNTERFACE	30.06_3	30.08_3	23.08_5	23.14_5	21.79_6	21.82_6	25.75_4	25.75_4	36.36_2	36.40_2	**37.49_1**	**37.51_1**
2	eNTERFACE	EmoDB	27.83_6	24.27_6	40.18_4	44.69_3	28.58_5	27.01_5	40.42_3	42.27_4	**44.41_1**	**52.27_1**	44.25_2	52.00_2
3	EmoDB	AFEW4.0	26.07_4	25.99_4	**30.39_1**	29.78_3	25.47_6	25.57_6	25.75_5	25.93_5	27.51_3	30.19_2	29.10_2	**31.00_1**
4	AFEW4.0	EmoDB	29.87_5	35.02_5	38.17_2	46.81_3	27.41_6	31.37_6	36.25_4	44.38_4	37.33_3	47.80_2	**39.66_1**	**50.00_1**
5	eNTERFACE	AFEW4.0	20.80_5	18.39_6	23.79_3	25.72_3	18.66_6	18.60_5	22.61_4	21.21_4	24.67_2	**27.70_1**	**24.83_1**	26.20_2
6	AFEW4.0	eNTERFACE	18.68_4	18.72_4	19.75_3	19.75_3	17.48_6	17.47_6	18.10_5	18.11_5	**21.93_1**	**21.96_1**	21.64_2	21.66_2

- ◆ It is convincing that the limited label information provided by a small number of samples in source database will lead to low recognition rate.
- ◆ The data imbalance between source and target databases is an important factor which will affect the cross-cropus speech emotion recognition tasks.

Acknowledgements

This research was supported by the Natural Science Foundation of China under Grants 61172185 and 61602345, the Application Foundation and Advanced Technology Research Project of Tianjin, the Academy of Finland, Tekes Fidipro program and Infotech Oulu.