Journal of Southeast University (English Edition) Vol. 35, No. 2, pp. 160 - 167 June 2019 ISSN 1003—7985

Transfer learning with deep sparse auto-encoder for speech emotion recognition

Liang Zhenlin[1] Liang Ruiyu[1,2] Tang Manting[3] Xie Yue[1] Zhao Li[1] Wang Shijia[1]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China)
([3]School of Computer Engineering, Jinling Institute of Technology, Nanjing 211169, China)

Abstract: In order to improve the efficiency of speech emotion recognition across corpora, a speech emotion transfer learning method based on the deep sparse auto-encoder is proposed. The algorithm first reconstructs a small amount of data in the target domain by training the deep sparse auto-encoder, so that the encoder can learn the low-dimensional structural representation of the target domain data. Then, the source domain data and the target domain data are coded by the trained deep sparse auto-encoder to obtain the reconstruction data of the low-dimensional structural representation close to the target domain. Finally, a part of the reconstructed tagged target domain data is mixed with the reconstructed source domain data to jointly train the classifier. This part of the target domain data is used to guide the source domain data. Experiments on the CASIA, SoutheastLab corpus show that the model recognition rate after a small amount of data transferred reached 89.2% and 72.4% on the DNN. Compared to the training results of the complete original corpus, it only decreased by 2% in the CASIA corpus, and only 3.4% in the SoutheastLab corpus. Experiments show that the algorithm can achieve the effect of labeling all data in the extreme case that the data set has only a small amount of data tagged.

Key words: sparse auto-encoder; transfer learning; speech emotion recognition

DOI: 10.3969/j.issn.1003-7985.2019.02.003

In the speech emotion recognition task, the classifier needs to determine the emotion category included in the speech according to the acoustic features and parameters of the speech. For speech emotion classifiers, their performance is usually positively correlated with the amount of data in the corpus; that is, the more data samples used for training, the higher the recognition rate will be obtained when testing. In addition, when the distribution of training data and test data is quite different, the performance of the classifier will decline significantly. Therefore, there are many similar problems in constructing the speech emotion recognition system. Firstly, in reality, it is very difficult to have a large number of annotated corpora due to the high cost of manual annotation. Secondly, due to the difference in data distribution between corpora, it is impossible to directly apply an existing speech emotion classifier to another scenario. Therefore, most of the actual work of deploying the speech emotion recognition system is to design an appropriate and effective feature representation to achieve the target classification performance. There are several speech emotion corpora at present, but they are usually quite different in spoken language, emotion types, performance or spontaneity, and the annotating scheme, such as classification or dimension[1]. In addition, when manually annotating an emotional corpus, as there are no prescribed standards but some subjective and artificial judgments, there are certain differences between different corpora. In order to reduce manual annotation, narrow the differences between corpora annotated in different ways, and accelerate the construction of the speech emotion system in practical application, speech emotion recognition requires a method to reuse existing corpora.

In recent years, transfer learning has the ability to transfer useful information from one or more source tasks to related target tasks[2], which has gradually attracted the attention of researchers. Transfer learning can improve the learning effect, especially when only a small amount of data is available in the target domain[3]. Transfer learning can also apply to speech emotion recognition tasks. For a new corpus with different data features or data distribution, it may only have a small amount of annotated data and a large amount of unannotated data. In this case, it is impossible to directly apply the model trained by other tagged corpora to this new corpus. Deng et al.[4] proposed a feature transfer scheme based on the sparse auto-encoder, in which an independent sparse encoder is established based on each emotion category to learn the features of each emotion category and perform feature transfer. Latif et al.[5] used deep belief networks (DBN) for knowledge transfer across corpora. Zong et al.[6] pro-

Received 2018-12-15, **Revised** 2019-04-10.

Biographies: Liang Zhenlin (1995—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61871213, 61673108, 61571106), Six Talent Peaks Project in Jiangsu Province (No. 2016-DZXX-023).

Citation: Liang Zhenlin, Liang Ruiyu, Tang Manting, et al. Transfer learning with deep sparse auto-encoder for speech emotion recognition [J]. Journal of Southeast University (English Edition), 2019, 35(2): 160 - 167. DOI: 10.3969/j.issn.1003-7985.2019.02.003.

posed a domain-adapted least squares regression algorithm to solve the feature space mapping problem of transfer. Song et al. [7] used the norm to search for potential feature subspace to minimize the difference between the target domain and the source domain data in a feature subspace.

In this paper, the deep sparse auto-encoder is used to learn the transfer of features so as to complete the establishment of a speech emotion classification model across the corpus. This method has two stages: In the first stage, a deep sparse self-encoder is trained to learn the potential low-dimensional feature representation of the data of the target domain. Therefore, the representation can reconstruct the data of the target domain. In the second stage, the representation method is applied to the source domain data. Therefore, it can be mapped from the low-dimensional features to the data of the target domain, and then the mapped data is used to train the classifier.

Compared with the previous method of feature transfer using an auto-encoder, the independent emotion auto-encoder is established for each emotion category, training and performing feature transformation independently. This paper establishes a unified auto-encoder for feature transfer for all emotion categories, so that the knowledge between different categories can also be transferred and assisted in feature transformation. Then, effective cross-category feature learning and feature transformation are established, so that the model can be better trained in the cross-corpus task, and be more generalized.

In addition, in the cross-corpus task of speech emotion recognition, the previous auto-encoders basically used a single-layer to ensure the efficiency of the auto-encoder training and reduce the complexity of the transmission. In this paper, a multi-layer deep sparse auto-encoder is used to construct a more complex feature transfer mapping space by using the powerful learning ability of a deep neural network. Moreover, by applying strong sparse constraints to the auto-encoder, the difficulty of training the deep auto-encoder and the degree of over-fitting are reduced. Experiments on CASIA and SoutheastLab corpora verify the effectiveness of the proposed method.

1 Basic Principles

This section introduces the basic principles of transfer learning. On this basis, this paper analyzes several situations that may occur in transfer learning for speech emotion recognition, and describes the corresponding specific cases.

First, for transfer learning, there are two basic elements: task and domain, where the symbol of the domain is D and the symbol of the task is T. For a domain D, it consists of two parts: feature space χ and marginal probability distribution $P(\boldsymbol{X})$, $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\} \in \chi$. Taking speech emotion recognition as an example, the feature space may be a MFCC coefficient of a certain dimension or a low-order descriptor of a certain order. $\boldsymbol{X}$ is the MFCC coefficient or low-order descriptor of a sample in the corresponding corpus. For two domains, as long as there is one difference between the feature space and the marginal probability distribution; the two domains are different.

For a given domain $D = \{\chi, P(\boldsymbol{X})\}$, task T also consists of two parts: the label space γ and the target prediction function $f(\cdot)$, i.e., $T = \{\gamma, f(\cdot)\}$. Also, taking the task of speech emotion recognition as an example, the label space γ is the type of emotion. For a classification task that needs to recognize four emotions, namely, "happy", "sad", "angry" and "indignant", γ is the label space constituted by the above four emotions. However, it is a discriminant function for the recognition of the above four emotions, which cannot be directly observed from the data, and can only be learned from it. From the perspective of probability, generally, it is also regarded as a conditional probability distribution.

For transfer learning, at least the existence of both source and target is required; that is, the transfer from the source to the target needs to be completed. For a given source domain D_S and source task T_S, as well as the target domain D_T and target task T_T; transfer learning is designed to take advantage of the knowledge of D_S and T_T to improve the performance of the target prediction function $f(\cdot)$ in T_T, where $D_S \neq D_T, T_S \neq T_T$.

For the speech emotion recognition task, according to the above definition, if the source domain D_S and the target domain D_T are different, then the feature space χ and the marginal probability distribution $P(\boldsymbol{X})$ of the two domains have at least one difference. If the feature space χ is different and the marginal probability distribution $P(\boldsymbol{X})$ is the same, it indicates that the task of the transfer learning is to transfer the knowledge based on the classification model with different features on the same corpora. For example, for the same corpora, the existence source domain model is characterized by MFCC, and attempts to transfer the knowledge of the source domain to a classification model characterized by low-level descriptors. If the feature space χ is the same and the marginal probability distribution $P(\boldsymbol{X})$ is different, it indicates that the task of the transfer learning is to transfer the knowledge based on the same feature classification model on different corpora. For example, for a corpus, the existence of the source domain model is characterized by MFCC, and it tries to transfer the knowledge of the source domain to the same MFCC but based on a classification model on another differently distributed corpus. If the feature space χ and the marginal probability distribution $P(\boldsymbol{X})$ are both different, it indicates that the transfer learning attempts to transfer knowledge of different feature-based classification models on different corpora.

Similarly, if the source task T_S and the target T_T are

different, it means that there is at least one difference between the label space γ and the conditional probability distribution $P(Y \mid \boldsymbol{X})$. If the label space γ is different, it indicates that the source task is different from the target task in attempting to complete the classification label. In speech emotion recognition, it can be the emotion type corresponding to the source task and the target task. In the case of conditional probability distribution $P(Y \mid \boldsymbol{X})$, it may be that the label data distribution of the source task and the target task is different. Similarly, taking speech emotion recognition as an example, it may be that the proportion of an emotion in the source task is much larger than the corresponding proportion in the target task.

2 Algorithm Design and Implementation

An auto-encoder is a neural network composed of several hidden layers that can set the target value equal to the input, and is used to find a common data representation from the input[8]. The auto-encoder can be divided into two parts. One part is the encoder and the other part is the decoder, as shown in Fig. 1. The former few hidden layers of the auto-encoder constitute the encoder, which is represented by a light blue graphic in Fig. 1. The output of the last layer can be regarded as the encoded feature, represented by a dark blue graphic. Then, several hidden layers constitute the decoder. The decoder represented by the light green graphic is responsible for decoding the coding features, obtaining the same output as the input, and completing the refactoring. In the process of input reconstruction, the auto-encoder learns the distribution of data through the repeated encoding-decoding process, and can compress and encode the data to obtain a more compact feature representation[9].

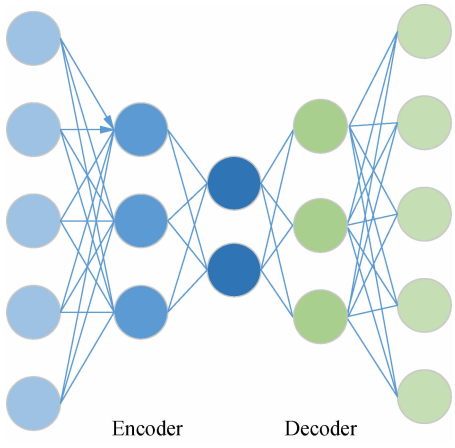

Fig. 1 Auto-encoder structure

For an auto-encoder, assuming that the input is $\boldsymbol{x}_i$; the output is $\boldsymbol{y}_i$; and the parameter is $\boldsymbol{\theta}$; then during its training, the goal of parameter optimization is

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \| \boldsymbol{x}_i - \boldsymbol{y}_i \|^2 \tag{1}$$

By limiting the expected activation of the hidden unit to sparse, that is, adding a regularization term[10], the deviation of the expected activation degree of the hidden unit from the target sparsity is penalized. Thus, Eq. (1) becomes the following loss function:

$$L = \sum_{i=1}^{N} \| \boldsymbol{x}_i - \boldsymbol{y}_i \|^2 + \beta \sum_{j=1}^{m} \mathrm{sp}(\boldsymbol{\rho} \| \hat{\boldsymbol{\rho}}_j) \tag{2}$$

where $\mathrm{sp}(\boldsymbol{\rho} \| \hat{\boldsymbol{\rho}}_j)$ is the sparsity penalty term, which is calculated as follows:

$$\mathrm{sp}(\boldsymbol{\rho} \| \hat{\boldsymbol{\rho}}_j) = \boldsymbol{\rho} \log \frac{\boldsymbol{\rho}}{\hat{\boldsymbol{\rho}}_j} + (1 - \boldsymbol{\rho}) \log \frac{1 - \boldsymbol{\rho}}{1 - \hat{\boldsymbol{\rho}}_j} \tag{3}$$

where $\hat{\boldsymbol{\rho}}_j$ is the average activation degree of all neurons in the hidden layer; $\boldsymbol{\rho}$ is the sparse degree; β is the penalty coefficient; and m is the number of all neurons. By adding sparse constraints[11], most neurons in the auto-encoder network are in an inhibitory state, while only a few neurons are active, which reduces the redundancy of the network and increases the robustness of the model[12].

In this paper, a five-layer deep sparse auto-encoder is constructed to learn the potential feature representation of the speech emotion corpus. Assuming that the source domain corpus is D_{S}, and the target domain corpus is D_{T}, the goal of transfer learning is to use the knowledge of the source domain corpus to establish a suitable speech emotion recognition model for D_{T} based on the use of a small amount of data in the target domain corpus D_{T}. In the algorithm proposed in this paper, the sparse auto-encoder is trained by a small amount of D_{T} data firstly. Therefore, the deep sparse auto-encoder can learn the data distribution and feature representation of D_{T} better, and save the trained sparse auto-encoder model. After that, the data of D_{S} is inputted into the deep sparse auto-encoder, and the coding features are taken out as the input features of the recognition model. Meanwhile, the data of D_{T} is inputted into the sparse auto-encoder and the coding features are extracted. The encoded feature extraction D_{T} is divided into a training set and test set, and the training set is merged with the feature extraction D_{S} to obtain a final training set. Finally, the classifier model is trained by the training set data, and after training, the model is verified by the test set data. The details of the specific algorithm are as follows:

1) Define the inputs as source domain corpus D_{S} and target domain corpus D_{T}. The initialization reconstruction training data D_{R} is an empty set, i.e., $D_{\mathrm{R}} = \varnothing$.

2) Divide D_{T} into training set $D_{\text{T-Train}}$ and test set $D_{\text{T-Test}}$, and part of data $D_{\text{T-Train-SA}}$ from the training set $D_{\text{T-Train}}$ will be used for training the sparse auto-encoder, and the ratio of $D_{\text{T-Train-SA}}$ to D_{T} is α.

3) Construct a deep sparse auto-encoder SA and use the data $D_{\text{T-Train-SA}}$ divided in step 2 for training.

4) Input the data $D_{\text{T-Train}}$ and $D_{\text{T-Test}}$ of the target domain corpus D_{T} into the SA to obtain the reconstructed data

$D_{\text{RT-Train}}$ and $D_{\text{RT-Test}}$.

5) Input the data of the source domain corpus D_{S} into the SA to obtain the reconstructed data D_{RS}.

6) Select part of the data with the proportion of η from $D_{\text{RT-Train-SA}}$, and merge with D_{RS} to form the final reconstructed training data D_{R}.

7) Construct classifier C, use D_{R} for training, $D_{\text{RT-Test}}$ for testing, and obtain the recognition accuracy of the cross-corpus task for the training model.

3 Experimental Results

3.1 Corpus and features

This paper uses two corpora: CASIA corpus[13] and SoutheastLab corpus. The CASIA Chinese emotional corpus is recorded by the Institute of Automation (Chinese Academy of Sciences). It consists of four professional speakers, six emotions (angry, happy, fear, sad, surprise, and neutral) with a total of 9 600 different pronunciations. Among them, 300 sentences are of the same text; that is to say, the same text is read with different emotions. These corpora can be used to compare and analyze the acoustic and rhythmic expression in different emotional states. The other 100 sentences are in different texts, which can literally tell what kind of emotion they belong to, making it easier for the recorder to express emotions more accurately. The SoutheastLab corpus is recorded by the Center for Signal Processing and Applied Research of Southeast University and includes a total of 6 237 samples of six emotions (angry, anxious, fearful, sad, tired and neutral). As only four emotions (anger, joy, sadness and neutral) are the same in the two corpora, only the data of these four emotions are selected from the two corpora for the experiment.

The 988-dimensional feature vector is obtained by using the OpenSmile tool[14] to extract the features of the two corpora. Since the value range of the feature vector is large, among -10^3 to 10^6, it is not conducive to the training of the neural network[15]. Therefore, the feature is z-score normalization, that is

$$\boldsymbol{x}_i = \frac{\boldsymbol{x}_i - \mu}{\sigma} \tag{4}$$

where μ and σ are the mean and variance of the data set, respectively. It should be noted that since the distribution of the two corpora is different, the z-score normalization of the two corpora is carried out independently.

3.2 Benchmark classification performance

This paper uses a variety of classifiers to classify the SoutheastLab data set and CASIA data set to obtain benchmark classification performance firstly. The results are shown in Fig. 2. The experiment uses a 5-fold cross-validation to evaluate the model. Among them, the classifiers used in the experiment are:

1) Deep belief net (DBN)[16-17], which has three layers of hidden layers with 1 024 neurons in each layer, is trained by greedy layer-by-layer training.

2) Deep sparse auto-encoder (deep SAE), which is combined with SVM to extract features using deep sparse auto-encoder, uses SVM for classification. The network structure of deep SAE is shown in Tab. 1.

3) Deep SAE is combined with DNN. Feature extraction is completed by using the deep sparse auto-encoder, and then classification is completed by DNN. The network structure of deep SAE is the same as that of deep SAE in 2). DNN is a five-layer neural network with 768 neurons in each layer, which adopts the activation function of Leaky ReLU[18], and is finally classified by Softmax.

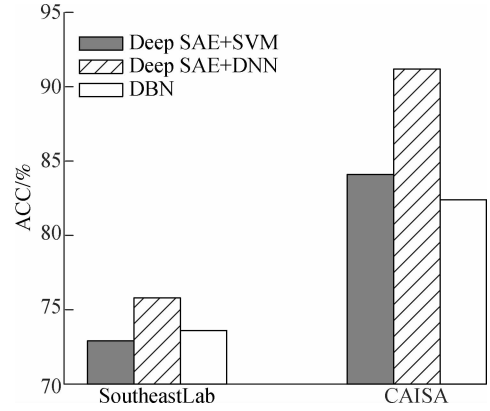

Fig. 2 Comparison of benchmark performance of DBN and Deep SAE methods on two corpora

Tab. 1 Deep sparse auto-encoder parameters

Component	No.	Layer	Neurons	Activation
Encoder	1	Fully connected layer	988	ReLU
	2	Fully connected layer	768	ReLU
	3	Fully connected layer	512	ReLU
Decoder	1	Fully connected layer	768	ReLU
	2	Fully connected layer	988	

As can be seen from Fig. 2, on both the SoutheastLab and CASIA corpora, the combination method of deep SAE and DNN achieve the best classification performance, reaching 91.2% recognition accuracy on the CASIA corpus and 75.8% accuracy on the SoutheastLab corpus. In the SoutheastLab corpus, the combination of deep SAE and SVM reached 72.9%, slightly lower than that of DBN(73.6%). However, in the CASIA corpus, it is 1.7% higher than DBN. Through the analysis of the above experimental results, it can be found that the recognition rate of unsupervised feature extraction using deep SAE is generally higher than that of the DBN method.

3.3 Speech emotion transfer

First, this paper experiments based on the case that η is 0, this is to say that there is no target domain data $D_{\text{RT-Train-SA}}$ but only source domain data D_{RS} in the recon-

structed training data D_R for training SVM. In the experiment, SVM is adopted as the classifier, and the proportion of data $D_{RT\text{-}Test}$ used for testing is fixed at 50%. The experimental results are shown in Fig. 3. Among them, the solid line shows the result of the SoutheastLab corpus as the target domain and the CASIA corpus as the source domain in the transfer experiment. It can be found that when the SoutheastLab data set is the target domain and the CASIA data set is the source domain, for the SVM classifier, which is trained by the reconstructed data obtained from the transformation of CASIA, its recognition accuracy on the target domain test data is only slightly higher than 25%. However, with the increase in proportion α of data $D_{RT\text{-}Test}$ used to train the deep sparse auto-encoder, when sparse auto-encoders are trained by using more target domain data, the recognition rate of the classifier starts to increase, which indicates that the sparse auto-encoder uses the target domain data to learn its potential feature representations. With the increase of data, the feature representation learned can also be more effective and reliable. Even though the data distribution of the two corpora is quite different during knowledge transfer, the sparse auto-encoder with more robust coding ability can map the source domain data to a more accurate joint distribution space.

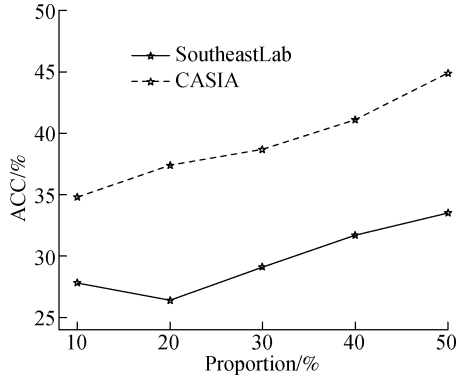

Fig. 3 SVM recognition rate of reconstructed data training when $\eta = 0$

The dotted line in Fig. 3 shows the results of the transfer state experiment based on the CASIA corpus as the target domain and the SoutheastLab corpus as the source domain. It can be seen that when using CASIA as the target domain and using SoutheastLab as the source domain data for data reconstruction, the classification performance on the CASIA corpus is better than the experimental results described in the previous paragraph, which is also consistent with the conclusion in Fig. 2. At the same time, with the increase in proportion α of $D_{RT\text{-}Test}$ used in CASIA to train the deep sparse auto-encoder, the recognition rate of the classifier starts to increase, indicating that the deep sparse auto-encoder does transfer the knowledge of the SoutheastLab corpus. Therefore, the reconstructed data used to train the classifier contains this part of the knowledge, and the classifier can use this part of knowledge to classify the CASIA corpus.

It can be seen from the above experiments that the performance of the trained SVM classifier is poor when there is no target domain data $D_{RT\text{-}Train\text{-}SA}$ in the reconstructed training data D_R, and the classification accuracy rate of the SoutheastLab corpus is only about 30%, and the classification accuracy rate of the CASIA corpus is about 44%, which is also unsatisfactory. The reason is that when the knowledge transfer is performed on the SoutheastLab and CASIA corpora with large differences in data distribution, no matter which is the target domain and which is the source domain; the target domain data D_{RT} is never added to the reconstructed training data D_R, and there is only the source domain data D_{RS}. Therefore, the classifier in the process of learning lacks effective guidance, and makes its conditional probability distribution completely based on the source domain data distribution, which cannot be well applied to the classification of the target domain discrimination. Secondly, this paper experiments on the reconstruction training data of different η values for training the classifier. In order to ensure the fairness of the experimental comparison, the experiment uses a fixed α value. The data $D_{RT\text{-}Test}$ used for testing has a fixed proportion of 50% of D_{RT}, and uses the DNN and SVM models as classifiers for comparison. Fig. 4 shows the classifier performance under different η when the target domain corpus is CASIA. It can be seen from the figure that as the target data $D_{RT\text{-}Train\text{-}SA}$ is gradually added to

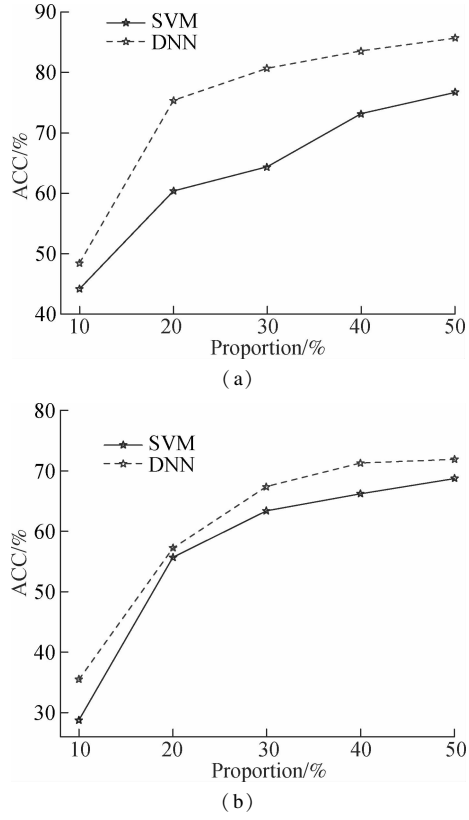

Fig. 4 Comparison of classifier performance under different η. (a) The target domain corpus is CASIA; (b) The target domain corpus is SoutheastLab

the reconstruction training data D_R, the performance of the model is gradually improved. For the speech emotion recognition model with DNN as the classifier, η increases from 0% to 50%, and the recognition accuracy of the model increases by about 41%, which is close to the performance of the benchmark classifier. In addition, when η is 10%, which means that only a small amount of target domain data is added, the classification accuracy of the DNN model increases from 48.2% to 75.1%, with an increase of nearly 30%. For the speech emotion recognition model with SVM as the classifier, η increases from 0% to 10%, and the recognition accuracy of the model increases by about 22%. When η increases to 50%, the recognition accuracy of the model increases by about 45%, and the improvement is nearly doubled, indicating that adding a certain proportion of the target domain data $D_{\text{RT-Train-SA}}$ in the reconstruction training data D_R can help the classifier to establish a conditional probability distribution that is suitable for identifying the target domain data.

Fig. 4 also shows the performance comparison of SVM and DNN classifiers under different η when the target domain corpus is CASIA and SoutheastLab. Compared with the previous experiment, it can be seen that when the target domain corpus is SoutheastLab, the classification accuracy of SVM and DNN is relatively close, and the curve trend is also similar. Similarly, as the target domain data is gradually added to the reconstructed training data, the classification accuracy of SVM and DNN is gradually increased. Moreover, when η increases from 0% to 10%, the accuracy of the SVM model increases by about 29%, and the DNN model increases by about 23%. When η is 50%, the SVM achieves a recognition accuracy of 69.2%, and the DNN is 72.4%, which is also close to the performance of the benchmark classifier in Experiment 1. Compared with the benchmark classifier which used all the target data to train, our model has only 2% difference in CASIA, as well as 3.4% in the SoutheastLab corpus. The above analysis shows that in the process of training the classifier with the reconstruction training data, the target domain data can play a role in guiding the model learning. Using only a small amount of the labeled target domain data can enable the model to achieve better performance.

3.4 Cross-linguistic speech emotion transfer

A five-layer deep sparse auto-encoder is constructed and used to learn the potential feature representation of the speech emotion corpus. The parameter settings are shown in Tab. 1. As it is a sparse self-encoder, the encoder needs to delete the acquired information first (that is, the number of neurons is lower than that in the upper layer; therefore, the information is lost), and then the reconstruction can extract the domain invariant features of the target domain data. The sparse auto-encoder needs to reduce the dimension of the input 988 dimension. If the single layer neural network is directly reduced to 512 or 256 dimensions, it will cause excessive information loss, so a dense layer of 768 neurons is introduced between the reduced-dimensional networks. Therefore, the selection of neurons is 988, 768, 512, 768 and 988. Please note that when we reconstruct the low-dimensional structural representation, we use the output of the layer containing 512 neurons.

This experiment adopts two corpora of different languages: the Chinese corpus CASIA and the English corpus Enterface. The CASIA corpus is composed of six emotions, namely, angry, happy, fear, sad, surprise, and neutral, with a total of 9 600 different pronunciation samples. Enterface contains six emotions: anger, disgust, fear, happy, sad, and surprise. There are 71 samples for each type of emotion, and 426 samples in total. The two corpora share the same five emotions, namely fear, happy, sad, surprise and anger. Therefore, the experiment uses the data of the five emotions in the two corpora.

The 988-dimensional feature vector is obtained by using the OpenSmile tool[10] to extract the features of the two corpora. As the value range of the feature vector is large, from -10^3 to 10^6, it is not conducive to the training of the neural network. Therefore, we do z-score normalization on the features. It should be noted that since the distribution of the two corpora is different, the z-score normalization of the two corpora is carried out independently.

The data amount of the two corpora varies greatly, the CASIA corpus has 9 600 samples while Enterface only has 426 samples. It is difficult for Enterface to transfer CASIA as the source domain data set. Therefore, the experiment uses CASIA as the source domain data set and Enterface as the target domain data set for feature transfer learning.

The experimental classifier uses SVM as the benchmark classifier with a penalty parameter C of 1.5. The Enterface corpus is divided into two equal parts. One is for training the deep sparse auto-encoder, it conducts feature transformation after the training of the sparse auto-encoder as the data of the training SVM, and the other is used to verify the performance of the classifier as test data after feature transformation.

As shown in Fig. 5, the brighter the color, the better the recognition effect. The ordinate on the upper graph represents the actual label category, and the abscissa represents the predicted label category. After the feature transformation of the deep sparse auto-encoder, the overall recognition accuracy obtained by using SVM as a classifier is about 57%. Among them, the sad category has the highest recognition rate of 65%, and the happy category has the lowest recognition accuracy rate of 48%. The happy category and the surprise category have the

highest degree of confusion, which is 21%.

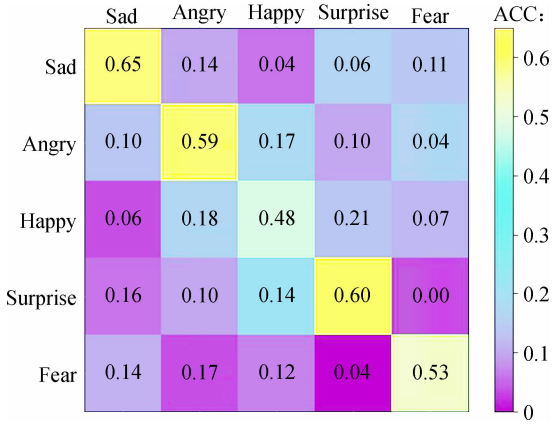

Fig. 5 Confusion matrix of the Enterface corpus

For the cross-lingual emotional transfer experiment, due to the huge difference between corpora and the small overlapping degree of potential feature subspace between them, it is difficult to search for potential feature subspace by using the deep sparse auto-encoder as the feature transformation network. When the encoding space corresponding to the deep sparse auto-encoder differs greatly from the potential feature subspace, the transfer feature obtained from the feature transformation is difficult to effectively transfer the source domain knowledge. Therefore, it can not help the target domain to establish a strong conditional probability distribution.

4 Conclusions

1) A unified feature transfer model for all categories is established by using the deep sparse auto-encoder. Therefore, in the process of transferring knowledge from the source domain corpus to the target domain corpus, the cross-category knowledge transfer is reinforced rather than the transfer knowledge or the transfer feature for a specific category.

2) Furthermore, by increasing the sparsity constraint, the deep sparse auto-encoder can resist over-fitting and learn a more accurate potential joint probability distribution.

3) The experiments demonstrate that the increase in proportion of target domain data has a positive effect on the accuracy. It means that the model performance could increase rapidly with a small amount of labeled target domain data, which has a practical application potential for many scenarios that need to transfer learning.

References

[1] Schuller B, Vlasenko B, Eyben F, et al. Cross-corpus acoustic emotion recognition: Variances and strategies [J]. *IEEE Transactions on Affective Computing*, 2010, **1**(2): 119 - 131. DOI:10.1109/t-affc.2010.8.

[2] Lim H, Kim M J, Kim H. Cross-acoustic transfer learning for sound event classification[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China, 2016: 16021470.

[3] Torrey L, Shavlik J. Transfer learning[M]//*Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2010: 242 - 264. DOI: 10.4018/978-1-60566-766-9.ch011.

[4] Deng J, Zhang Z X, Marchi E, et al. Sparse autoencoder-based feature transfer learning for speech emotion recognition[C]// 2013 *Humaine Association Conference on Affective Computing and Intelligent Interaction*. Geneva, Switzerland, 2013: 511 - 516.

[5] Latif S, Rana R, Younis S, et al. Cross corpus speech emotion classification—An effective transfer learning technique[EB/OL]. (2018-01-22)[2018-11-20]. https://www.researchgate.net/publication/322634480_Cross_Corpus_Speech_Emotion_Classification_-_An_Effective_Transfer_Learning_Technique.

[6] Zong Y, Zheng W M, Zhang T, et al. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression[J]. *IEEE Signal Processing Letters*, 2016, **23**(5): 585 - 589. DOI:10.1109/lsp.2016.2537926.

[7] Song P, Zheng W M. Feature selection based transfer subspace learning for speech emotion recognition [J]. *IEEE Transactions on Affective Computing*, 2018: 1. DOI:10.1109/taffc.2018.2800046.

[8] Xu J, Xiang L, Liu Q S, et al. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images[J]. *IEEE Transactions on Medical Imaging*, 2016, **35**(1): 119 - 130. DOI: 10.1109/tmi.2015.2458702.

[9] Sarath C A P, Lauly S, Larochelle H, et al. An autoencoder approach to learning bilingual word representations [C]//*International Conference on Neural Information Processing Systems*. Kuching, Malaysia, 2014: 1853 - 1861.

[10] Goodfellow I J, Le Q V, Saxe A M, et al. Measuring invariances in deep networks[C]// *International Conference on Neural Information Processing Systems*. Bangkok, Thailand, 2009: 646 - 654.

[11] Mairal J, Bach F, Ponce J. Online learning for matrix factorization and sparse coding[J]. *Journal of Machine Learning Research*, 2009, **11**(1): 19 - 60.

[12] HintonG E. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, **313**(5786): 504 - 507. DOI:10.1126/science.1127647.

[13] Pan S F, Tao J H, Li Y. The CASIA audio emotion recognition method for audio/visual emotion challenge 2011 [C]// *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction*. Memphis, TN, USA, 2011: 388 - 395.

[14] Eyben F, Wöllmer M, Schuller B. openSMILE—The Munich versatile and fast open-source audio feature extractor[C]//*ACM International Conference on Multimedia*. Firenze, Italia, 2010: 1459 - 1462.

[15] Larochelle H, Bengio Y, Louradour J, et al. Exploring Strategies for training deep neural networks[J]. *Journal of Machine Learning Research*, 2009, **1**(10): 1 - 40.

[16] Bengio Y, Lamblin P, Dan P, et al. Greedy layer-wise

training of deep networks[J]. *Advances in Neural Information Processing Systems*, 2007, **19**(2007): 153-160.
[17] Hinton G E. Deep belief networks[J]. *Scholarpedia*, 2009, **4**(5): 5947. DOI:10.4249/scholarpedia.5947.
[18] Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network[EB/OL]. (2015-11-27)[2018-11-20]. http://de.arxiv.org/pdf/1505.00853.

一种基于深度稀疏自编码的语音情感迁移学习方法

梁镇麟[1] 梁瑞宇[1,2] 唐曼婷[3] 谢 跃[1] 赵 力[1] 王诗佳[1]

([1] 东南大学信息科学工程学院, 南京 210096)
([2] 南京工程学院通信工程学院, 南京 211167)
([3] 金陵科技学院计算机工程学院, 南京 211169)

摘要: 为了提高跨语料库的语音情感识别效率,提出了一种基于深度稀疏自编码的语音情感迁移学习方法. 算法首先通过训练深度稀疏自编码器来对目标域中的少量数据进行重建,使得编码器可以学习到目标域数据低维度的结构表征. 然后,将源域数据和目标域数据通过训练好的深度稀疏自编码器,得到靠近目标域低维度的结构表征的重建数据. 最后,利用部分重建的含标签的目标域数据与重建的源域数据混合后共同训练分类器,以便完成对源域数据的引导. 在 CASIA、SoutheastLab 语料库上的实验表明,通过少量数据迁移后的模型识别率在 DNN 上达到了 89.2% 和 72.4%. 和完整原始语料库训练的结果相比,在 CASIA 上仅下降了 2%,在 SoutheastLab 上仅下降了 3.4%. 实验说明,该算法能够在数据集只有少量数据有标签的极端情况下,达到逼近于所有数据都有标签的效果.

关键词: 稀疏自编码器;迁移学习;语音情感识别
中图分类号: TN912.3