Linked source and target domain subspace feature transfer learning - exemplified by speech emotion recognition

Jun Deng, Zixing Zhang, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Deng, Jun, Zixing Zhang, and Björn Schuller. 2014. "Linked source and target domain subspace feature transfer learning - exemplified by speech emotion recognition." In *2014 22nd International Conference on Pattern Recognition, 24-28 August 2014, Stockholm, Sweden*, edited by Magnus Borga, Anders Heyden, Denis Laurendeau, and Michael Felsberg, 761–66. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICPR.2014.141.

Linked Source and Target Domain Subspace Feature Transfer Learning – Exemplified by Speech Emotion Recognition

Jun Deng[1], Zixing Zhang[1], Björn Schuller[2,1]
[1]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2]Machine Learning Group, Department of Computing, Imperial College London, London, U. K.
{jun.deng, zixing.zhang}@tum.de, bjoern.schuller@imperial.ac.uk

***Abstract*—The typical inherent mismatch between the test and training corpora and by that between 'target' and 'source' sets usually leads to significant performance downgrades. To cope with this, this study presents a feature transfer learning method using Denoising Autoencoders (DAEs) to build high-order subspaces of the source and target corpora, where features in the source domain are transferred to the target domain by an additional neural network. To exemplify effectiveness of our approach, we select the INTERSPEECH Emotion Challenge's FAU Aibo Emotion Corpus as target corpus and further two publicly available databases as source corpora for extensive and reproducible evaluation. The experimental results show that our method significantly improves over the baseline performance and outperforms today's state-of-the-art domain adaptation methods.**

***Index Terms*—feature transfer learning; denoising autoencoders; cross-corpus; domain adaptation; speech emotion recognition**

I. Introduction

This paper addresses the common situation in practice where training and test samples come from different corpora. In this ever-present case, many traditional machine learning methods may not live up to expectations as a common assumption is not met, namely the request that training and test data instances are drawn from the same feature space and the same distribution [1].

The influence of such differences in training and test databases can be alleviated by building a feature representation that incorporates domain knowledge into the data [2]. However, such feature engineering can be very application-specific and labour-intensive. Therefore, directly learning the underlying explanatory factors hidden in the corpora seems more promising, and more importantly able to expand the scope of applicability to novel target tasks.

Representation learning, i. e., learning transformations of the data that make it easier to extract useful information when building classifiers or other predictors, is recently gathering a lot of attention [3], [4]. The key idea of representation learning is to make use of deep architectures which leads to abstract representation. Generally, more abstract concepts are invariant to most local changes of the input. Following this spirit, representation learning can potentially be used to deal with the problem caused by the situation discussed above. For example, the previous works include [5] where part of the present authors proposed feature transfer learning based on a sparse autoencoder method for discovering knowledge in acoustic features from small labelled target data to improve performance of speech emotion recognition when applying the knowledge to source data [5]. Further, in [6] a stacked denoising autoencoder with sparse rectifiers is used for domain adaptation in large-scale sentiment analysis.

In this paper, we propose a feature transfer learning method by using a combination of denoising autoencoders (DAEs) and regression neural networks (NNs). In our approach, we first train exclusive DAEs for source and target data in an unsupervised way so as to build two subspaces. By training a DAE for input data, the subspace gets implicitly grounded by the input data modality, allowing us to give a high-order feature representation for each input instance. Besides, we transfer target data into the source subspace as well. Then, a regression NN is used to discover the difference between the resulting features for target data in the source subspace and the ones in the target subspace. We expect that the NN becomes a link which is able to compensate for the disparity between the source domain and the target domain to a certain degree. Therefore, we feed the resulting high-order representations for the source data into the NN to predict new high-order representations in the target subspace, in turn, leading to reducing the disparity between high-order features for source data. Afterwards, we use the new high-order features for source data in the target subspace as training set and the original subspace features for target data as test set to carry out normal supervised algorithms for classification.

As a real-life test-case to exemplify effectiveness of our proposed approach, we decided for the INTERSPEECH Emotion Challenge task aiming at recognition of negative versus neutral speech. In this field, the two disjoint databases could have different acoustic conditions and/or type of emotions such as acted, elicited, or naturalistic. It may even be that the spoken languages or the emotion annotation schemes are different. These differences are known to produce a detrimental effect on the real-world performance of acoustic emotion recognition systems, since in training they will not have prepared for data subsequently encountered in use.

The remainder of this paper is organised as follows. Sec-

tion II first discusses related work. We then present the proposed feature transfer learning method in Section III. In Section IV, next, we briefly introduce the selected real-world classification task for exemplification of effectiveness, including three chosen speech databases and acoustic features used, and demonstrate experiments on the three corpora. Finally, we draw a conclusion and point out promising future work in Section V.

II. Related Work

Transfer learning has been proposed to deal with the significant problem of how to reuse the knowledge learnt previously from 'other' data or features [1]. Among the various ways of transfer learning, domain adaptation of statistical classifiers has been shown to be well suited for a problem where the data distribution in the target domain is different from the one in the source domain. In this case, the target domain contains the test instances, and the source domain constrains training instances which are under a different distribution with the target domain data. One general approach to address the domain adaptation problem is to assign more weight to those training examples that are not most similar to the test data, and less weight to those that poorly reflect the distribution of the target (test) data. This idea of weighting the input data based on the test data is known as importance weighting. The goal is to estimate importance weights, denoted β, from training examples and test examples by taking the ratio of their densities $\beta(x) = p_{te}(x)/p_{tr}(x)$ where $p_{te}(x)$ and $p_{tr}(x)$ are test and training input densities. Kanamori et al. proposed unconstrained least-squares importance fitting (uLSIF) to estimate the importance weights by a linear model [7]. Tsuboi et al. modelled the importance function by a linear (or kernel) model, which resulted in a convex optimisation problem with a sparse solution, called KLIEP [8].

Kernel mean matching (KMM) could recently be shown to lead to significant improvement in acoustic emotion recognition when Hassan et al. firstly considered to explicitly compensate for acoustic and speaker differences between training and test databases [9]. The KMM was proposed to deal with sampling bias in various learning problems [10], which allows to directly estimate the resampling weights by matching training and test distribution feature means in a reproducing kernel Hilbert space. We employ KMM for comparison in the experiments later in this paper, where more details on KMM are given. The objective function is given by the discrepancy term between the two empirical means

$$\left\| \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i \Phi(x_i^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \beta_i \Phi(x_i^{te}) \right\|^2, \quad (1)$$

where Φ are the mapping functions.

Using $K_{ij} := k(x_i^{tr}, x_j^{tr})$ and $\kappa_i := \frac{n_{tr}}{n_{te}} \sum_{j=1}^{n_{te}} k(x_i^{tr}, x_j^{te})$, the (1) above becomes the quadratic problem for finding suitable β:

$$argmin_\beta \tfrac{1}{2}\beta^T K\beta - \kappa^T\beta$$
$$s.t.\, \beta_i \in [0, B] and \left| \sum_{i=1}^{n_{tr}} \beta_i - n_{tr} \right| \leq n_{tr}\epsilon, \quad (2)$$

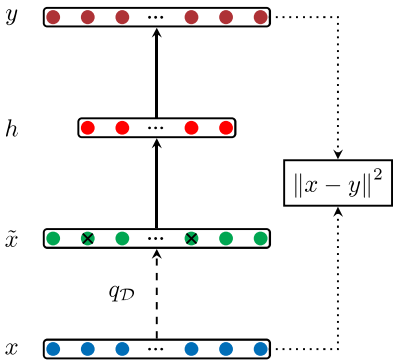

Fig. 1: A denoising autoencoder architecture. An input x is corrupted (via $q_{\mathcal{D}}$) to $\tilde{x}$.

where the upper limit of importance weight $B > 0$ and $\epsilon > 0$ are tuning parameters, and k is the kernel function. Since KMM optimisation is formulated as a convex quadratic programming problem, it leads to a unique global solution.

III. Proposed Methodology

Let us now turn to the proposed method by first introducing the background to start from.

A. Denoising Autoencoders

A denoising autoencoder (DAE) – a more recent variant of the basic autoencoder consisting of only one hidden layer – is trained to reconstruct a clean 'repaired' input from its corrupted version [11]. In doing so, the learner must capture the structure of the input distribution in order to reduce the effect of the corruption process [3]. It turns out that more robust features are learnt compared with the basic autoencoder. Deep neural networks use it, as an element, to find common data representation from the input [4], [12], [13]. We show the architecture of the DAE in Fig. 1.

Formally, an input example $x \in \mathbf{R}^n$ is first corrupted into $\tilde{x}$ by means of a corrupting function $\tilde{x} \sim q_{\mathcal{D}}(\tilde{x}|x)$, which could be masking corruption, additive Gaussian noise or salt-and-pepper noise in images.

Then, in response to the corrupted example $\tilde{x}$, the hidden representation $h(\tilde{x}) \in \mathbf{R}^m$ is

$$h(\tilde{x}) = f(W_1\tilde{x} + b_1), \quad (3)$$

where $f(z)$ is a non-linear activation function, typically a logistic sigmoid function $f(z) = 1/(1 + \exp(-z))$ applied component-wise, $W_1 \in \mathbf{R}^{m \times n}$ is a weight matrix, and $b_1 \in \mathbf{R}^m$ is a bias vector.

The network output maps hidden representation h back to a reconstruction $y \in \mathbf{R}^n$:

$$y = f(W_2 h(\tilde{x}) + b_2), \quad (4)$$

where $W_2 \in \mathbf{R}^{n \times m}$ is a weight matrix, and $b_2 \in \mathbf{R}^n$ is a bias vector. The weights W_1 and W_2 may be tied, i. e., $W_2 = W_1^T$, if it is necessary.

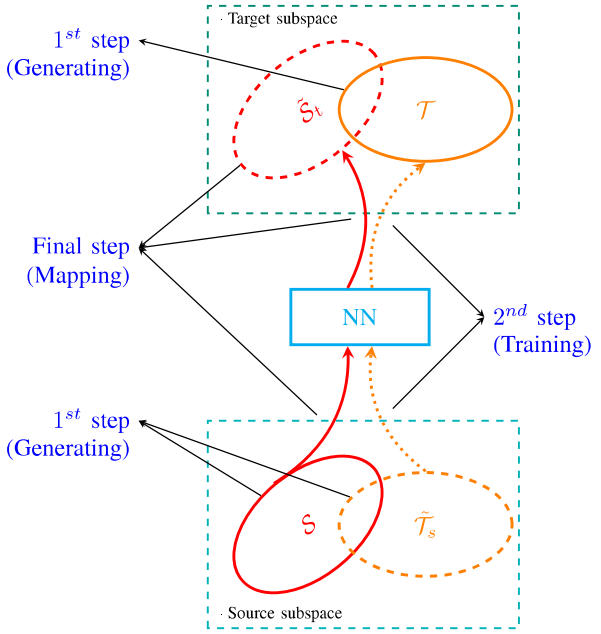

Fig. 2: Overview of the proposed feature transfer learning method. The function "NN" refers to a regression neural network with one hidden layer. The sets $\mathcal{S}$ and $\tilde{\mathcal{S}}_t$ are high-order features in the source space and in the target space. The sets $\mathcal{T}$ and $\tilde{\mathcal{T}}_s$ are high-order features in the target space and in the source space.

Given an input set of examples $\mathcal{X}$, a DAE training consists in finding parameters $\theta = \{W_1, W_2, b_1, b_2\}$ that minimise the reconstruction error, which corresponds to minimising the following objective function:

$$\mathcal{J}(\theta) = \sum_{x \in \mathcal{X}} \|x - y\|^2 \tag{5}$$

The minimisation is usually realised either by stochastic gradient descent or more advanced optimisation techniques such as the L-BFGS or conjugate gradient method. In this paper, we also add a weight-decay regularisation term into the objective function to avoid overfitting. Note that, if the number of hidden units m is less than the number of input units n, then the network is forced to learn a compressed representation of the input.

B. Feature Transfer Learning Using DAEs and NNs

It has been observed widely that autoencoders can automatically capture useful features hidden in data, which potentially have greater predictive power. Such features are often used in building a deep hierarchy of features, within the contexts of supervised, semi-supervised, or unsupervised modelling [12], [14]–[16].

Fig. 2 depicts an overview of our proposed method, which is composed of the following three steps. In this work, we first prepare two different DAEs for the source domain data and the target domain data so as to capture the domain-individuality information, which leads to generating the features in high-order subspace via encoding original features for the source or target data from the input layer to the hidden layer of the corresponding DAEs (see (3)). In addition, we also generate the high-order features for the target data in source subspace, which is built by the DAE for the source domain. As a result, we have the high-order features of the source data in the source subspace $\mathcal{S}$, the features of the target data in the target subspace $\mathcal{T}$, and the features of the target data in the source subspace $\tilde{\mathcal{T}}_s$.

Next, a regression neural network (NN), consisting of one hidden layer, is used to discover the difference between the source subspace and the target subspace. At this point, the NN is trained to minimise the error for the target data between the high-order features $\mathcal{T}$ in the target subspace and its 'other' version $\tilde{\mathcal{T}}_s$ in the source subspace. Specifically, given a target example in the target subspace $h_T \in \mathcal{T}$ and the respective version in the source subspace of it $h_{\tilde{T}_s} \in \tilde{\mathcal{T}}_s$, the NN learns by solving the following optimisation problem:

$$\mathcal{J}_{\mathcal{NN}}(\theta_{NN}) = \sum_{\substack{h_T \in \mathcal{T} \\ h_{\tilde{T}_s} \in \tilde{\mathcal{T}}_s}} \left\| h_T - g(h_{\tilde{T}_s}) \right\|^2, \tag{6}$$

where

$$\begin{aligned} g(h_{\tilde{T}_s}) &= f(W_2^{NN}(f(W_1^{NN} h_{\tilde{T}_s} + b_1^{NN})) + b_2^{NN}), \\ \theta_{NN} &= \{W_1^{NN}, W_2^{NN}, b_1^{NN}, b_2^{NN}\}. \end{aligned} \tag{7}$$

Here $f(z)$ is a non-linear activation function. Similarly with the DAE, the parameters W_1^{NN}, W_2^{NN} are the weights and b_1^{NN}, b_2^{NN} are the bias terms. Note that, the size of the input layer is the same as the output layer in the special architecture of the NN.

Finally, we transfer the features of the source data $\mathcal{S}$ from the source subspace to the target subspace by means of the trained NN, which leads to a new form of the source data $\tilde{\mathcal{S}}_t$ in the target domain. In the end, the new form of the source data $\tilde{\mathcal{S}}_t$ and the features of the target data $\mathcal{T}$ will be taken to build a standard supervised classifier for speech emotion recognition in the following exemplary use-case. In the ongoing, our proposed method is referred to as DAE-NN for simplicity.

IV. Experiments

Let us now investigate effectiveness of the proposed linked DAE-NN transfer learning on a well defined and standardised real-world Machine Learning task. This example stems from the field of (cross-corpus) acoustic speech emotion recognition. Most previous approaches do not consider reduction of the differences between corpora before building emotion recognition models, but demonstrate the difficulty in cross-corpus processing, cf. e. g., [17], [18]. Recently, [5] used sparse autoencoders to successfully transfer useful knowledge from other corpora to the target one benefiting, however, from the labels of the target set. [9] considers important weights to shift the separating hyperplane of Support Vector Machines

(SVMs) in such a way as to take into consideration the more important training data, however, without considering a cross-corpus scenario. In the ongoing, we provide experimental results for a challenging real-life task by using other disjoint corpora as training set based on the proposed DAE-NN feature transfer learning and demonstrate its superiority over related approaches.

A. Selected Task and Data

To investigate the performance of the proposed method, we consider the INTERSPEECH 2009 Emotion Challenge two-class task [19]. It is based on the spontaneous FAU Aibo Emotion Corpus (FAU AEC), which contains recordings of 51 children at the age of 10–13 years interacting with the pet robot Aibo in German speech. The children were made believe that the Aibo was responding to their commands, whereas the robot was actually remote-controlled in a Wizard-of-Oz manner and did not respond to their commands. Hence, the database contains induced emotionally-coloured speech. The details of the challenge's two-class task 'negative' versus 'idle' emotion are given in Table II. For the experiments to follow, we always evaluate the emotion recognition model on the test set of the FAU AEC as was used in the challenge. Transfer learning can be considered very promising in this case, as speech resources of emotional children speech and naturalistic emotion display are extremely rare – in FAU AEC's spoken language (German) we are not aware of an alternative one. Thus, to benefit from richer availability of adult speech resources or such in related languages such as English (also Germanic language family), one can use these in training – however, after transferring them to the target domain (FAU AEC).

Thus, in our experiments, two further publicly available and popular databases, namely the Airplane Behavior Corpus (ABC) [20], and the Speech Under Simulated and Actual Stress (SUSAS) set [21] are chosen as training sets, which are highly different from the target set FAU AEC in terms of speaker age (adults vs. children as in FAU AEC), spoken language (English vs. German as in FAU AEC), type of emotion (partially acted vs. naturalistic as in FAU AEC), degree of spontaneity and phrase length, type of recording situation, and naturally annotators and subjects. For comparability with FAU AEC, we have to map the diverse emotion classes onto the valence axis in the dimensional emotion model. The mapping defined for the cross-corpus experiments is used to generate labels for binary valence from the emotion categories in order to generate a unified set of labels. This mapping is given in Table I. In addition, Table II summarises the three chosen databases and shows the existing difference in them.

B. Acoustic Features

To keep in line with the INTERSPEECH 2009 Emotion Challenge [19], we decided to use its standardised feature set of 12 functionals applied to 2×16 acoustic Low-Level Descriptors (LLDs) including their first order delta regression coefficients as shown in Table III. In detail, the 16 LLDs are zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and Mel-frequency cepstral coefficient (MFCC) 1–12. Then, 12 functionals – mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and ranges as well as two linear regression coefficients with their mean square error (MSE) – are applied on the chunk level. Thus, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes. To ensure reproducibility as well, the open source openEAR toolkit [22] was used with the pre-defined challenge configuration. The features were normalised in the range between 0 and 1.

TABLE I: Emotion categories mapping onto negative and positive valence classes for the three chosen databases.

Corpus	Negative	Positive
FAU AEC	angry, emphatic, reprimanding, touchy	joyful, motherese, neutral, rest
ABC	aggressive, intoxicated, nervous, tired	cheerful, neutral, rest
SUSAS	high stress, screaming, fear	medium stress, neutral

TABLE III: Overview of the standardised feature set provided by the INTERSPEECH Emotion Challenge.

LLDs (16×2)	**Functionals** (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel, position, range
(Δ) MFCC 1–12	linear regression: offset, slope, MSE

C. Experimental Setup and Evaluation Metrics

As classifier, we use linear SVMs as was used in the official baseline of the challenge with a fixed penalty factor $C = 0.5$ as the basic supervised learner. The toolkit LIBLINEAR [23] is applied in the experiments, and the type of solver s is fixed to 2.

In the DAE-NN learning process, we applied the 3rd party software minFunc implementing L-BFGS to optimise the parameters of DAEs and NNs [24]. For training of the DAE, we inject Gaussian noise with a variance of 0.01 to generate the corrupted input. The number of hidden units m was fixed to 200 for the two DAEs and the NN, and the weight decays values λ were set to 0.0001. The number of epochs for the DAEs is set to 250 and the number for NNs is decreased to 50. To choose the hyper-parameters, we search for the minimum validation error (sum of squared error) in a pre-defined range.

We evaluate the performance of the baselines and DAE-NN systems using unweighted average recall (UAR) as was the competition measure in the challenge. It equals the sum of the recalls per class divided by the number of classes, and better reflects overall accuracy in the given case of presence of class imbalance. We validate statistical significance of the results according to a one-sided z-test.

Note that the FAU AEC official test partition is always used

TABLE II: Summary of the three chosen databases.

Corpus	Age	Language	Speech	Emotion	# Valence		# All	h:mm	#m	#f	Rec	Rate
					-	+						kHz
FAU AEC	children	German	variable	natural	3 358 / 2465	6 601 / 5 792	9 959 / 8 257	9:20	21	30	normal	16
ABC	adults	German	fixed	acted	213	217	430	1:15	4	4	studio	16
SUSAS	adults	English	fixed	natural	1 616	1 977	3 593	1:01	4	3	noisy	8

*Number of instances per binary valence (**# Valence**, Negative (**-**), Positive (**+**)), and overall number (**# All**) – for FAU AEC divided into official training and test set by "/". Total audio time (**h:mm**). Number of female (**#f**) and male (**#m**) subjects. **Rec**ording conditions (studio/normal/noisy).*

as the 'target' test set, its official training partition is partially used as 'target' training set, and the further two corpora are exclusively used as additional 'source' training sets.

D. Models for Comparison

We compare the following methods to evaluate our proposed approach in the context of the current state-of-the-art:

- **Target**: in this reference 'method' we randomly (repeated ten times to reduce singularity effects) pick a number of instances from the FAU AEC official training set to train an SVM, i. e., without the need of transferring in an intra-corpus scenario. For a fair comparison, the number is given by picking the same number of learning instances as is given by the ABC or SUSAS sets, respectively. In other words, this can be considered as baseline reference using exclusively target-type data, but each with the same amount of training instances as will later be used coming from non-target data.
- **w/o**: uses the source corpora ABC or SUSAS to train the standard (SVM) classifier directly, i. e., without (w/o) using any methods to reduce the mismatch between source and target data. This is the 'classical' cross-corpus testing.
- **KMM**: utilises the KMM method (see Section II) on the ABC and SUSAS database for covariate shift adaptation. We choose the 'tuning parameters' in KMM following [9], [10]. It is thus the first reference with application of transfer learning.
- **DAE**: employs denoising autoencoders for representation learning in order to match training examples to test examples, which was successfully applied to the transfer learning challenge and domain adaptation [6], [25], and may be considered as close reference from a method point of view, as a DAE is also used, yet, without the linking between source and target domain during transferring as is proposed in the current paper.
- **DAE-NN**: finally uses the proposed DAE-NN to compensate for the mismatch between the features on the training and test sets, then trains standard SVMs using the compensated features and labels in the training set.

E. Results

In the case of a cross-corpus scenario, we train acoustic emotion recognition models on ABC or SUSAS while testing on the FAU AEC test set. We run the experiments ten times for each training set, and evaluate the performance by UAR. When using ABC or SUSAS, averaged UAR over the ten trials is visualised in Figure 3, including the error bars, and given quantitatively in Table IV for reference comparison. As can been seen, the **DAE-NN** method always achieves larger average UAR for ABC and SUSAS when compared to the **Target** and **w/o** cases. It also exceeds the UAR achieved by the **DAE** method and the covariate shift adaptation **KMM**.

TABLE IV: Average UAR over ten trials: **Target**, **w/o**, covariate shift adaptation KMM, **DAE**-based representation learning, and the proposed **DAE-NN** method related to training with ABC and SUSAS.

UAR [%]	**Target**	**w/o**	**KMM**	**DAE**	**DAE-NN**
ABC	58.32	55.28	62.52	56.20	**63.63**
SUSAS	62.41	57.32	60.41	62.08	**64.73**

More specifically, for the small database ABC (composed of only 430 instances), one can easily see that the standard method (**w/o**) only obtains an average UAR around the chance level (55.28 %) due to the inherent mismatch between the ABC used for training and the FAU AEC test set. The accuracy obtained by the **DAE** method reaches 56.20 %, the covariate shift adaptation **KMM** can boost the accuracy to 62.52 %. However, with **DAE-NN** one reaches 63.63 %, which yields 1.11 % absolute improvement when compared to **KMM**. This improvement has a high statistical significance at the 0.001 level compared with the baseline **w/o** and even the one of **Target**, i. e., even when using 430 target domain instances.

In comparison with ABC, SUSAS's average UAR in 'classical do-nothing' cross-corpus testing (**w/o** method) is also close to chance level. Here, however, it is worth noting that the average UAR achieved by training with an equivalent number of target domain instances as found in SUSAS (i. e., 3.6 k instances, **Target** method) increases sharply to 62.41 % because of the eight times larger size of SUSAS than ABC leading to eight times more instances chosen from the FAU AEC training set. Unlike ABC, SUSAS cannot obtain a great benefit from the covariate shift adaptation **KMM** but can achieve a comparable performance by DAE. Again, the **DAE-NN** method gives the highest average UAR of 64.73 %, which is again surprisingly even exceeding average UAR obtained by the **Target** 'method'. Compared with all four reference methods, the superiority of the proposed **DAE-NN** method passes the significant test at the 0.001 level.

It is worth noting that, the UAR from the DAE-NN for the SUSAS database is slightly larger than the one for the ABC. We believe that this can partially be attributed to the

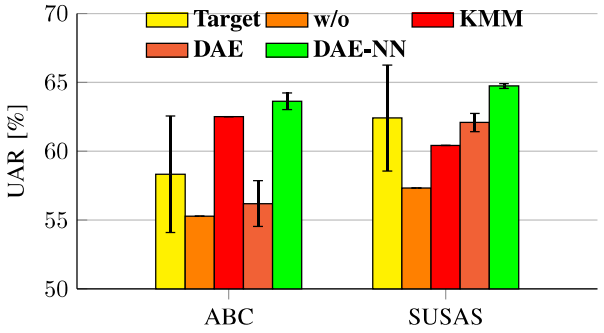

Fig. 3: Cross-corpus average UAR over ten trials using **Target**, **w/o**, the covariate shift adaptation **KMM**, the **DAE**-based representation learning, and the representation learning **DAE-NN** for ABC and SUSAS.

larger size of the SUSAS. Overall, the DAE-NN-based feature transfer learning could be shown as highly useful in reducing the difference for cross-corpus recognition.

V. Conclusions and Outlook

We proposed a feature transfer learning method, referred to as DAE-NN, to address a situation where training and test set come from different corpora. The method uses denoising autoencoders to build a subspace for the source domain and the target domain, and makes use of regression neural networks in order to reduce the mismatch between target data and source data on subspace feature level. The proposed method was successfully applied to a well-defined standard machine learning task: speech emotion recognition. Experimental results with three publicly available corpora demonstrate that the proposed method effectively and significantly enhances the emotion classification accuracy and competes well with other domain adaptation methods.

It is natural to believe that deep architectures are able to extract complex structure and build internal representation from rich inputs. Thus, we plan to expand the shallow architecture of DAE-NN to a deep architecture.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 338164 (European Research Council Starting Grant 'iHEARu').

The authors further acknowledge funding from the China Scholarship Council (CSC).

Responsibility lies with the authors.

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] G. Liu, Y. Lei, and J. H. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification," in *Proc. of INTERSPEECH*, Makuhari, Japan, 2010, pp. 482–485.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] I. Goodfellow, Q. Le, A. Saxe, H. Lee, and A. Ng, "Measuring invariances in deep networks," in *Proc. NIPS*, Vancouver, Canada, 2009, pp. 646–654.

[5] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. ACII* , Geneva, Switzerland, 2013, pp. 511–516.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. ICML*, Bellevue, U. S. A., 2011.

[7] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," in *Proc. of NIPS*, Vancouver, Canada, 2008, pp. 809–816.

[8] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation with model selectionand its application to covariate shift adaptation," in *Proc. of NIPS*, Vancouver, Canada, 2007, pp. 1433–1440.

[9] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1458–1468, 2013.

[10] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of ICML*, Helsinki, Finland, 2008, pp. 1096–1103.

[12] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition," in *Proc. of INTERSPEECH*, Lyon, France, 2013.

[13] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. of NIPS*, 2013, pp. 809–817.

[14] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. NIPS*, Vancouver, Canada, 2007.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 6822–6826.

[17] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. of ASRU*, Big Island, HY, 2011, pp. 523–528.

[18] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions – some pilot experiments," in *Proc. of LREC Workshop*, Valletta, Malta, 2010, pp. 77–82.

[19] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH*, Brisbane, U. K., 2009, pp. 2794–2797.

[20] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. of ICASSP*, Honolulu, HY, 2007, pp. 733–736.

[21] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. EUROSPEECH*, Rhodes, Greece, 1997.

[22] F. Eyben, M. Wollmer, and B. Schuller, "openEAR — Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. ACII*, Amsterdam, 2009, pp. 576–581.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] M. Schmidt. (2012) minFunc. [Online]. Available: http://www.di.ens.fr/ mschmidt/Software/minFunc.html

[25] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML*, Bellevue, U. S. A., 2011.