TRANSFERABLE POSITIVE/NEGATIVE SPEECH EMOTION RECOGNITION VIA CLASS-WISE ADVERSARIAL DOMAIN ADAPTATION

Hao Zhou, Ke Chen

School of Computer Science, The University of Manchester, Manchester, M13 9PL, U.K.

ABSTRACT

Speech emotion recognition plays an important role in building more intelligent and human-like agents. Due to the difficulty of collecting speech emotional data, an increasingly popular solution is leveraging a related and rich source corpus to help address the target corpus. However, domain shift between the corpora poses a serious challenge, making domain shift adaptation difficult to function even on the recognition of positive/negative emotions. In this work, we propose class-wise adversarial domain adaptation to address this challenge by reducing the shift for all classes between different corpora. Experiments on the well-known corpora EMODB and Aibo demonstrate that our method is effective even when only a very limited number of target labeled examples are provided.

Index Terms— speech emotion recognition, adversarial learning, supervised domain adaptation

1. INTRODUCTION

Speech emotion recognition [1] has attracted growing interest over the past two decades. It can be applied to many areas. For example, the recognition of positive/negative emotions can help improve the call center service and psychological disease diagnosis. However, it is highly difficult to collect a large volume of speech emotional data in a completely natural environment. Although considerable efforts have been made to build high-quality databases (corpora) of emotional speeches, data scarcity remains a bottleneck.

An increasingly popular solution to data scarcity is domain shift adaptation or transfer learning [2, 3]. It features leveraging a different but related, and information-rich source domain to improve the performance on the target domain, which we are really interested in but suffers from a lack of information. In the context of speech emotion recognition, a corpus collected in a specific way may be viewed as a domain. There are usually two cases of domain shift adaptation depending on whether label information in the target domain is available: unsupervised domain adaptation (UDA) and supervised domain adaptation (SDA). In either case, the key is eliminating domain shift, i.e. the difference of distributions, between the source and target domains. In speech emotion recognition, domain shift causes the well-known cross-corpora problem that the performance of a recognition system built on one corpus can be degraded significantly when being tested on a different corpus. Consequently, even recognition of positive/negative emotions can be challenging in the cross-corpora setting or in the scenario of domain shift adaptation. Moreover, the problem is even exacerbated when only very limited training data are available in the target domain. In this situation, it even poses a new challenge that the limited training data are insufficient to build up an effective speech emotion recognition system. Hence, the data in different domains/corpora have to be utilized.

So far there are several approaches [3, 4, 5] developed to address the domain shift issue, but most of them are applied to UDA that requires a great amount of unlabeled target-domain data be accessible during the training stage. This requirement, however, is often difficult to meet considering the difficulty of collecting speech emotional data. On the other hand, SDA can be applied as long as there are a small number of target-domain labeled data available, even if those training data are insufficient to build up an effective recognizer. For instance, a recent method named few-shot adversarial domain adaptation (FADA) [6] emerges as a promising SDA approach. The FADA works by aligning distributions across domains via adversarial learning in a discriminative manner. This strategy, however, may encounter a difficulty caused by very high intra-class variability in each corpus considering that the source and target corpora can differ with respect to the speech recording environment, the eliciting way of emotions (natural or acted), the speakers (subjects), the communication language, the pre-defined emotional states, and the emotion annotation scheme, etc. Hence, how to deal with very high intra-class variability in domain shift adaptation becomes an obstacle in transferable speech emotion recognition.

In this paper, we propose the class-wise adversarial domain adaptation (CADA) method to address the intra-class variability issue in domain shift adaptation towards improving the performance on the target domain for the positive/negative emotion recognition. Given a related source domain for which sufficient labeled data are provided, CADA employs adversarial learning in a generative manner to align each class pair between the source and target domains and hence is more effective for knowledge transfer in domain shift adaptation.

The experiments on the well-known corpora EMODB and Aibo demonstrate that CADA is effective even when only a very limited number of target labeled examples are provided.

2. RELATED WORK

Some pioneer works have systematically evaluated cross-corpora speech emotion recognition with a number of high-quality databases [7, 8]. These works, unfortunately, do not involve any adaptation techniques for reducing the domain shift. [3] first treats the mismatch in emotional data as covariate shift and proposes compensating for that shift by classical importance-weighting at the instance level. At the feature level, some autoencoder-based transfer learning methods [4, 5] have developed to seek a shared feature representation so that the knowledge can be transferred between the domains. All of these methods, however, are usually applied to unsupervised rather than supervised domain adaptation, which demands a lot of data in target domain that may not be easy to collect in reality.

Regarding supervised domain adaptation in speech emotion recognition, some works have verified that a few labeled data from the target domain can be hugely helpful, but the adaptation is achieved by simple fine-tuning [9]. A sophisticated technique is by [10]. But still based on autoencoder, this technique needs a relatively large number of target examples for adaptation. Recently adversarial learning [11, 12] gains a great popularity on domain shift adaptation. A significant trait about adversarial learning is that instead of directly measuring the similarity of different domains, it introduces a domain discriminator that distinguishes the source from the target domain, and a feature representation is then learned to be domain invariant by fooling the domain discriminator. As the state-of-the-art supervised domain adaptation approach, FADA [6] conducts adversarial learning in a discriminative way on generated pairs which mix multi-class training examples by considering the combination of classes in different domains. Although FADA yields good performance in different applications, it does not work well in speech emotion recognition as the method cannot deal with the high intra-class variability effectively in domain shift adaptation. In order to overcome this weakness, our CADA decomposes the domain shift problems on the basis of each common class in the source and target domains, leading to more effective adversarial learning. In addition, by contrast to FADA which employs typical Siamese networks feeding on paired examples, CADA can be implemented easily with a slightly modified multilayer perceptron (MLP).

3. OUR CADA APPROACH

Formally, given the source domain D_s and target domain D_t (we use s and t to refer to the source and target domain, respectively), where the source domain follows the distribution $P(X^s, Y^s)$ and the target domain $P(X^t, Y^t)$ (in our context, X denotes the input speech and Y the emotional class), the goal of domain shift adaptation is to learn a classification function f that minimizes the misclassification error $L_y(f(X^t), Y^t)$ by using all the data available in two domains. Under the setting of supervised domain adaptation, $D_s = \{(x_i^s, y_i^s)\}_{i=1}^N$ and $D_t = \{(x_i^t, y_i^t)\}_{i=1}^M$ ($M << N$).

A typical domain shift adaptation method usually works under the assumption $P(Y^s|X^s) = P(Y^t|X^t)$. By learning a feature space ϕ such that $P(\phi(X^s)) = P(\phi(X^t))$, it ideally leads to $P(Y^s|\phi(X^s)) = P(Y^t|\phi(X^t))$, which means the classifier can be shared by both domains. However, in speech emotion recognition, the underlying assumption that $P(Y^s|X^s) = P(Y^t|X^t)$ may be less solid because of the high intra-class variability between the source and target domains, and this further affects the learning process for the desired feature space.

To tackle this weakness, class-wise adversarial domain adaptation (CADA) is aimed at seeking a feature space ϕ for $P(\phi(X^s)|y_i) = P(\phi(X^t)|y_i)$, $y_i \in Y$ instead of $P(\phi(X^s)) = P(\phi(X^t))$. With the assumption that $P(Y^s) = P(Y^t)$, by Bayesian theory, we wish to have

$$\begin{aligned} P(y_i|\phi(X^s)) &= \frac{P(\phi(X^s)|y_i)P(y_i)}{\Sigma_i P(\phi(X^s)|y_i)P(y_i)} \\ &\approx \frac{P(\phi(X^t)|y_i)P(y_i)}{\Sigma_i P(\phi(X^t)|y_i)P(y_i)} \\ &= P(y_i|\phi(X^t)) \end{aligned} \tag{1}$$

where $\approx$ is carried out for domain shift adaptation.

To integrate adversarial learning and supervised learning into one process, we introduce a domain-class discriminator which can not only distinguish the classes but also distinguish the domains. To illustrate this idea, we take a binary classification task as an example. The modified discriminator (or classifier) classifies any instance into one of four categories: d_1 indicating Class 1 from source domain, d_2 Class 2 from source domain, d_3 Class 1 from target domain, and d_4 Class 2 from target domain. In the testing stage, we perform classification with these 4 categories and treat either the prediction d_1 or d_3 as Class 1, and either d_2 or d_4 as Class 2. This categorization scheme can be straightforwardly popularized to the cases involving more classes.

As shown in Fig. 1, CADA can be implemented based on an MLP which comprises two components, the feature encoder G_e parameterized by θ_e and the predictor G_p parameterized by θ_p. Both θ_e and θ_p are trained to minimize the typical cross entropy loss function

$$L_d = -\sum_i^{N+M} d^{x_i} \log G_p(G_e(x_i, \theta_e), \theta_p) \tag{2}$$

where d^{x_i} is the category of x_i. Meanwhile, θ_e is trained to

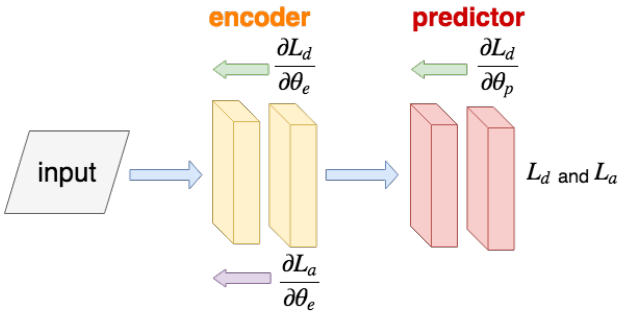

Fig. 1. The proposed class-wise adversarial learning domain adaptation structure comprises an encoder and a predictor (either may contain more than one hidden layer), parameterized by θ_e and θ_p respectively. The training process consists of two stages. In the first stage, both the encoder and predictor are trained based on the loss function L_d defined in Eq.(2). In the next stage, the predictor is fixed and only the encoder is trained based on the loss function L_a defined in Eq.(3). Through the two-stage learning, domain shift is reduced.

minimize the following loss function

$$\begin{aligned} L_a = -\{ & \sum_{x \in X_{d_1}} d_3 \log G_p(G_e(x, \theta_e), \theta_p) \\ & + \sum_{x \in X_{d_2}} d_4 \log G_p(G_e(x, \theta_e), \theta_p) \\ & + \sum_{x \in X_{d_3}} d_1 \log G_p(G_e(x, \theta_e), \theta_p) \\ & + \sum_{x \in X_{d_4}} d_2 \log G_p(G_e(x, \theta_e), \theta_p)\} \end{aligned} \tag{3}$$

where X_{d_i} denotes all the examples belonging to d_i ($i \in \{1, 2, 3, 4\}$). This loss function is designed to encourage the confusion of the equivalent classes in different domains. Specially, we want the model to believe the examples of certain class in one domain also belong to the equivalent class in the other domain. For instance, the first term on the right side of Eq.(3) suggests that the examples from d_1 are also from d_3. This principle is applied to all categories we defined. While FADA [6] performs adversarial learning on newly-generated data pairs without considering specific class information, minimizing Eq.(3) allows the adversarial learning to operate on each specific common class across the domains, i.e. class-wise adversarial learning. For clarity, the CADA learning process is summarized in Algorithm 1.

Algorithm 1 CADA learning algorithm

1: Initialize θ_e and θ_p randomly
2: Re-label training examples in both source and target domains in terms of d_i, $i \in \{1, 2, 3, 4\}$
3: **while** not convergent **do**
4: Update θ_e and θ_p by minimizing Eq.(2).
5: Update θ_e by minimizing Eq.(3).
6: **end while**

4. EXPERIMENT

4.1. Datasets

We evaluate our proposed CADA with the two well-known speech emotion datasets Aibo [13] and EMODB [14] for positive/negative emotion recognition. Despite the same language (German), these two corpora are collected in very different ways. They differ with respect to, at least, the recording environments, speakers (adults or children), the eliciting way of emotions (acted or spontaneous), and the pre-defined emotion class stereotypes. These factors can generate both large domain shift between the two corpora and high intra-class variability for each corpus. Considering the large size of Aibo, we treat the two parts, Aibo-Ohm and Aibo-Mont, separately in our experiments. Details on these three corpora are listed in Table 1.

Table 1. Datasets

Corpus	Type	Speakers	Class and Size
EMODB	acted	10 adults	Negative (anger, sadness, etc.): 385 Positive (happiness, neutral): 150
Aibo-Ohm	spontaneous	26 children	Negative (angry, touchy, etc.): 3358 Positive (joyful, neutral, etc.): 6601
Aibo-Mont	spontaneous	25 children	Negative (angry, touchy, etc.): 2465 Positive (joyful, neutral, etc.): 5792

4.2. Experimental Settings

To simulate the cross-corpora setting, we use one corpus as the source domain and a different corpus as the target domain. Since both Aibo-Ohm and Aibo-Mont have much more data than EMODB, we consider Aibo-Ohm or Aibo-Mont as the source domain, and EMODB as the target domain, which is consistent with the general experiences to use an information-rich domain as the source. For the transferable positive/negative emotion recognition, the goal is to achieve the best performance on the target domain with few labeled data in the target domain. The performance is measured by unweighted accuracy (UA); i.e. the accuracy per class averaged by the class number.

We further set the following baselines for comparison: 1) *all-source*: we test using the trained source model without any target information (without any adaptation); 2) *all-target*: we use the entire target dataset for traditional supervised learning; 3) *label-target*: we make use of only a few labeled target data (no source domain knowledge used), and this baseline is only introduced when there are over 10 examples per class in the target domain. We use the 5-fold cross validation (under the speaker-dependent condition) where 10% training examples are preserved for early stopping, and the mean and

Table 2. Results when using Aibo-Ohm as source domain. Two baselines are *all-source*: 55±2, and *all-target*: 81±3.

Examples	2	4	6	8	10	12
fine-tune	55±4	57±4	59±4	60±5	61±4	62±4
FADA	54±1	54±2	54±1	55±2	55±1	55±2
CADA	**58±4**	**59±4**	**62±3**	**63±4**	**63±4**	**64±3**

Table 3. Results when using Aibo-Mont as source domain. Two baselines are *all-source*: 51±1, and *all-target*: 81±3.

Examples	2	4	6	8	10	12
fine-tune	55±4	55±4	57±5	58±4	59±5	61±4
FADA	50±1	50±2	51±2	52±2	52±2	53±1
CADA	**57±3**	**59±3**	**60±3**	**60±4**	**61±3**	**63±4**

standard deviation of UA are reported. In addition, we consider two typical adaption methods in our comparison: fine-tuning [9], an effective method in speech emotion recognition for transfer learning (with an MLP in our experiment); and FADA [6], a state-of-the-art supervised domain shift adaptation method. For all the adaptation methods, the used target-domain examples are randomly selected from random speakers with the same setting. 20 trials have been conducted for each adaptation method in our experiment and the mean and the standard deviation of UA are reported.

For preprocessing and feature extraction, we use GeMAPS (62 features) [15] which has shown a comparable discriminative power as some large standard feature sets but with a much smaller size. We perform normalization by standardizing the feature values to the range $[-1, 1]$. Model selection suggests that an MLP of one hidden layer of 256 ReLu units is a proper base model and our CADA is implemented based on this model as well (c.f. Fig. 1). The model is constructed with TensorFlow and trained with the mini-batch of size 64 and the default learning rate via AdamOptimizer.

4.3. Results Analysis

Tables 2 and 3 report the performance when very few examples (from 2 to 12 with interval of 2) in the target domain are used. Due to the nature of supervised domain adaptation, the effectiveness of adaptation highly depends on the informativeness of the labeled target data which are used in the training process. That explains the relatively large standard deviation in two tables. All the adaptation methods achieve better performance than the baseline *all-source*, suggesting that domain shift adaptation is effective. It is clearly seen from Tables 2 and 3 that our CADA outperforms two state-of-the-art adaptation methods and FADA is inferior to others.

Figs. 2 and 3 show the performance of different adaptation methods and the MLP with the *label-target* setting by using different numbers of target labeled data. It is evident from Figs. 2 and 3 that the performance of those adaptation methods is better than that of the *label-target* setting when there are a few labeled data in the target domain. In particular, our CADA always outperforms other adaptation methods and the *label-target* setting up to 50/40 labeled examples in the target domain.

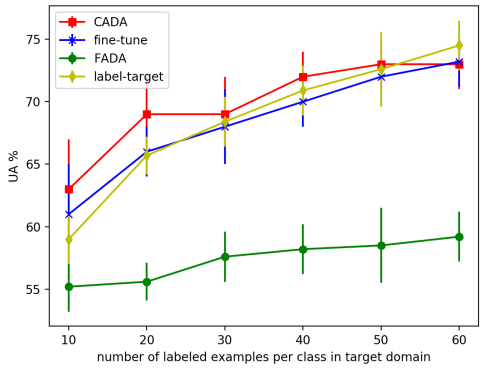

Fig. 2. Results when using Aibo-Ohm as source domain

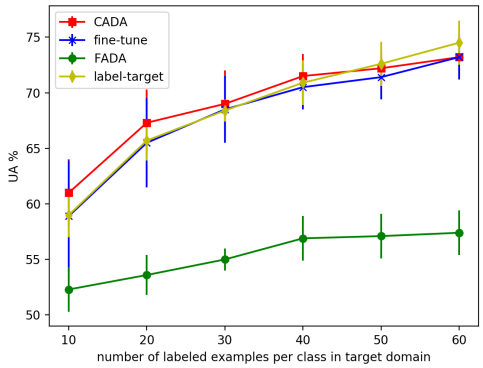

Fig. 3. Results when using Aibo-Mont as source domain

5. CONCLUSION

We have presented a new domain shift adaptation method, CADA, for transferable speech emotion recognition. Our experiments demonstrate that CADA is more effective than the direct use of those training examples in the target domain to build up a system and other adaptation methods when there are few training examples in the target domain. Although CADA was only evaluated on the positive/negative emotion recognition, it is straightforward to be extended to multi-class speech emotion recognition across multiple corpora and other application areas. In our ongoing work, we are going to address the computational issues in multi-class adaptation and extend CADA to unsupervised domain shift adaptation.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

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

[4] Jun Deng, Zixing Zhang, Florian Eyben, and Björn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[5] Jun Deng, Rui Xia, Zixing Zhang, Yang Liu, and Björn Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4818–4822.

[6] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2017, pp. 6673–6683.

[7] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[8] Florian Eyben, Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, et al., "Cross-corpus classification of realistic emotions–some pilot experiments," in *Proc. LREC workshop on Emotion Corpora, Valettea, Malta*, 2010, pp. 77–82.

[9] Mohammed Abdelwahab and Carlos Busso, "Supervised domain adaptation for emotion recognition from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5058–5062.

[10] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[12] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," *arXiv preprint arXiv:1702.05464*, 2017.

[13] Stefan Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*, University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[14] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[15] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.