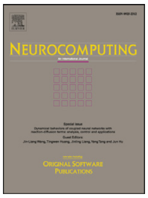
Neurocomputing 257 (2017) 79–87

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



An unsupervised deep domain adaptation approach for robust speech recognition

Sining Sun, Binbin Zhang, Lei Xie*, Yanning Zhang

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 13 November 2016
Accepted 29 November 2016
Available online 7 February 2017

Keywords:
Domain adaptation
Robust speech recognition
Deep neural network
Deep learning

ABSTRACT

This paper addresses the robust speech recognition problem as a domain adaptation task. Specifically, we introduce an unsupervised deep domain adaptation (DDA) approach to acoustic modeling in order to eliminate the training–testing mismatch that is common in real-world use of speech recognition. Under a multi-task learning framework, the approach jointly learns two discriminative classifiers using one deep neural network (DNN). As the main task, a label predictor predicts phoneme labels and is used during training and at test time. As the second task, a domain classifier discriminates between the source and the target domains during training. The network is optimized by minimizing the loss of the label classifier and to maximize the loss of the domain classifier at the same time. The proposed approach is easy to implement by modifying a common feed-forward network. Moreover, this unsupervised approach only needs labeled training data from the source domain and some unlabeled raw data of the new domain. Speech recognition experiments on noise/channel distortion and domain shift confirm the effectiveness of the proposed approach. For instance, on the Aurora-4 corpus, compared with the acoustic model trained only using clean data, the DDA approach achieves relative 37.8% word error rate (WER) reduction.

1. Introduction

The increasing availability of multimedia big data, including various genres of speech, is fostering a new wave of multimedia analytics that aim to effectively access the content and pull meaning from the data. Automatic speech recognition (ASR), which transcribes speech into text, serves as a necessary preprocessing step for multimedia analytics. With the help of big data, supercomputing infrastructure and deep learning [1], the speech recognition accuracy has been dramatically lifted during the past years [2]. Besides the Gaussian mixture model–hidden Markov model (GMM–HMM) architecture that dominates the acoustic modeling in speech recognition for many years, artificial neural networks have been historically used as an alternative model but with limited success [3]. Only recently, neural network has re-emerged as an effective tool for acoustic modeling because of the power of big data and effective learning method [4,51]. The DNN–HMM architecture has come to the central stage in speech recognition [2,5], replacing the GMM–HMM architecture. We have witnessed the success of various types of (deep) neural networks (DNNs) not only in speech recognition, but also in visual data processing, data mining and other areas [6–10].

Speech is a typical big data: not just in volume, but also noisy and heterogeneous. In practice, we desire a *robust* speech recognizer that is able to handle noisy data. For many machine learning tasks, including ASR, we usually assume that the training data and the testing data have the same probability distributions. However, real-world applications often fail to meet this hypothesis [5,11]. In speech recognition, both the GMM–HMM and DNN–HMM systems are Bayesian classifiers by nature. Theoretical investigation has shown that the training–testing mismatch notoriously leads to increase of errors in Bayesian classification [11]. There are many reasons that lead to the mismatch such as environmental noises, channel distortions [12] and room reverberations [13, [14]. To improve the environmental robustness of a speech recognizer, a common and efficient approach is multi-condition training [15] that uses the contaminated noisy data, together with the clean data, in the acoustic modeling training. But it is impossible to cover all kinds of real-world conditions and the mismatch still exists. Therefore, environment robustness is still a big challenge remain unsolved. On the other hand, real-world speech data is heterogeneous. Speech in different domains, e.g., broadcast news, lectures, meeting recordings and conversations, has different char-

* Corresponding author.
E-mail addresses: snsun@nwpu-aslp.org (S. Sun), xielei21st@gmail.com (L. Xie).

http://dx.doi.org/10.1016/j.neucom.2016.11.063
0925-2312/© 2017 Elsevier B.V. All rights reserved.

acteristics. This causes another mismatch that apparently decrease the speech recognition performance [14].

In order to eliminate the training–testing mismatch, a large number of robust speech recognition methods have been proposed, which in general fall into two categories: feature-space approaches and model-space approaches [16,17]. Most approaches need some prior knowledge about the mismatch. For example, noise characteristics have to be known beforehand or clean-noisy speech pairs[1] are needed [18]. Model adaptation is a typical model-space approach that is quite useful in noise robustness. The acoustic model, e.g., GMM-HMMs, is adapted using the new data either in a supervised manner [19] or an unsupervised manner [20]. For feature-space approaches, it is common to combine information about speaker, environment and noise, such as using *i*-vector [21], to acoustic features.

In this paper, we regard the robust speech recognition problem as a domain adaptation (DA) task [22]. Learning a discriminative classifier in the presence of the mismatch between training and testing distributions is known as *domain adaptation*. The essence of the domain adaptation and robust speech recognition is identical, that is, to eliminate the mismatch between the training data and the test data. We find that the speech features yield to different distributions if they come from different domains (such as clean and noisy speech conditions [16], and data sets with different genres). Specifically, if we train a DNN acoustic model using clean speech, we discovered that the feature distributions of clean and noisy speech yielded from this acoustic model are significantly different. Hence we would like to embed the domain information during the acoustic model training in order to obtain a “domain-invariant feature extractor”.

Our work is inspired by a recent DNN based unsupervised domain adaptation approach for image classification [23]. This *deep domain adaptation* (DDA) approach combines domain adaptation and deep feature learning within a single training process. Specifically, under a multi-task learning framework [52], the approach jointly learns one feature extractor and two discriminative classifiers using one single DNN: the feature extractor is trained to extract domain-invariant and classification-discriminative features; the label predictor predicts class labels and is used both during training and testing; a *domain predictor* discriminates between the source and the target domains during training. In order to obtain domain-invariant and classification-discriminative features, the feature extractor sub-network is optimized by minimizing the loss of the label predictor and maximizing the loss of the domain predictor at the same time, which is achieved by a special objective function we defined later. The parameters of two predictor sub-networks are optimized in order to minimize their losses on the training set. Compared with other unsupervised adaptation approaches, the DDA approach is easy to implement by simply augmenting a common feed-forward network with few standard layers and a simple new gradient reversal layer. Moreover, this approach only needs the labeled training data from the source domain and some unlabeled raw data of the new domain. Experiments show that the DDA approach outperforms previous state-of-the-art image classification approaches on several popular datasets [23].

In this study, we introduce the DDA approach to robust speech recognition. Applying DDA to speech recognition is not trivial. This is because speech recognition is a more challenging task as compared with image classification. We elaborate some of the major challenges as follows.

- *The large number of labels*: In the typical image classification task in [23], the number of classes are only dozens. In contrast, in speech recognition, the class labels are thousands of senones (i.e., phoneme states). The effectiveness of DDA on a large scale classification task like speech recognition desires an intensive study.
- *Decoding*: As compared with image classification, speech recognition is a rather complicated task with frame-level classification (classify each speech frame into senone labels) and decoding (Viterbi search from a large graph based on the classified frame labels). The accuracy gain in frame-level classification may not ensure consistent accuracy gain at the word level [24].
- *Deeper networks*: The neural networks in speech recognition usually have many hidden layers in order to learn highly nonlinear and discriminative features which are robust to irrelevant variabilities.

To bridge the gap, in this paper, we study how to integrate DDA into acoustic modeling and present a systematic analysis of the performance of DDA in robust speech recognition. Our study shows that the DDA approach can significantly boost the speech recognition performance in both noisy/channel distortion and domain-shift conditions.

The rest of this paper is structured as follows. Section 2 surveys the related work. Section 3 presents the framework of deep domain adaptation and studies how to use it in the speech recognition task. Experimental settings and results are discussed in Sections 4–6 and finally conclusions are drawn in Section 7.

2. Related work

As we just mentioned, robust speech recognition methods can be classified into two categories: feature-space approaches and model-space approaches [16,17]. Compared with model-space approaches, feature-space approaches do not need to modify or retrain the acoustic model. Instead, various operations can be performed in the acoustic features to improve the noise (or other distortions) robustness of the features. As for the model-space approaches, rather than focusing on the modification of features, the acoustic model parameters are adjusted to match the testing data.

2.1. Traditional methods

In the feature space, feature normalization is the most straightforward strategy to eliminate the training–testing mismatch. Popular strategies include cepstral mean subtraction (CMS) [25], cepstral mean variance normalization (CMVN) [26] and histogram equalization (HEQ) [27]. Obviously, speech enhancement methods [28,29] can be adopted to remove the noise before speech recognition. But the unavoidable distortions in the enhanced speech may cause another new mismatch problem.

Rather than updating the features, the acoustic model parameters can be compensated to match the testing conditions. A simple example of updating the models is to re-train them with the new data; or more popular, adding a variety of noise samples to clean training data, known as multi-style or multi-condition training [15,17]. However, due to the unpredictable nature of real-world noise, it is impossible to account for all noise conditions that may be encountered. Thus adaptive and predictive methods are proposed in the model-space. The adaptive methods update the model parameters when sufficient corrupted speech data are available. Popular methods include maximum a posteriori re-estimation (MAP) [30] and maximum likelihood linear regression (MLLR) [31]. In the predictive methods, a noise model is combined with the clean speech models to provide a corrupted speech acoustic model using some model of the acoustic environment. Parallel model combination (PMC) [32] and vector Taylor series (VTS) [33] fall into this category.

[1] Noisy speech may be generated manually by adding noises into clean speech.