The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Inferring Emotion from Conversational Voice Data: A Semi-Supervised Multi-Path Generative Neural Network Approach

Suping Zhou,[1] Jia Jia,[1]* Qi Wang,[1] Yufei Dong,[3] Yufeng Yin,[1] Kehua Lei[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
[2]Academy of Arts & Design, Tsinghua University, Beijing, China
[3]University of Science & Technology Beijing, Beijing, China
jjia@mail.tsinghua.edu.cn

Abstract

To give a more humanized response in Voice Dialogue Applications (VDAs), inferring emotion states from users' queries may play an important role. However, in VDAs, we have tremendous amount of VDA users and massive scale of unlabeled data with high dimension features from multimodal information, which challenge the traditional speech emotion recognition methods. In this paper, to better infer emotion from conversational voice data, we propose a semi-supervised multi-path generative neural network. Specifically, first, we build a novel supervised multi-path deep neural network framework. To avoid high dimensional input, raw features are trained by groups in local classifiers. Then high-level features of each local classifiers are concatenated as input of a global classifier. These two kinds classifiers are trained simultaneously through a single objective function to achieve a more effective and discriminative emotion inferring. To further solve the labeled-data-scarcity problem, we extend the multi-path deep neural network to a generative model based on semi-supervised variational autoencoder(semi-VAE), which is able to train the labeled and unlabeled data simultaneously. Experiment based on a 24,000 real-world dataset collected from Sogou Voice Assistant[1](SVAD13) and a benchmark dataset IEMOCAP show that our method significantly outperforms the existing state-of-the-art results.

1 Introduction

The increasing popularity of Voice Dialogue Applications(VDAs), such as Siri[2], brings great convenience to our daily life. As the same words said in different emotion can convey quite different messages, inferring emotion from these conversational voice data of queries can assist to understand the true meaning of users as well as provide more humanized responses.

Traditionally, in speech emotion recognition, there are two kinds of major frameworks. One is HMM-GMM framework based on dynamic features (Schuller, Rigoll, and Lang 2003) , another one is support vector machines (SVM) based on high-level representations(Schuller et al. 2009). Recently, more and more attention has been paid to deep learning for speech emotion recognition which results in a better performance than the traditional framework (Kim, Lee, and Provost 2013). Totally, two kinds of frameworks based on deep learning have been proposed for speech emotion recognition. Some researches are based on utterance-level features which usually extract high-level representations from low-level descriptors(LLDs) and then utilize deep neural network (DNN) for classification (Xia and Liu 2017). Meanwhile, instead of high-level statistics representation, some other researches utilize frame-level representation or raw signal as input to neural network for an end-to-end training (Zhang et al. 2016; Trigeorgis et al. 2016). Also, some works have focused on unsupervised learning for speech emotion recognition (Ren et al. 2014; Wu et al. 2016; Ghosh et al. 2016; Chang and Scherer 2017), which utilize the unlabeled data to help improve the performance. Generally, deep learning have made a great contribution to speech emotion learning.

However, for inferring emotion from conversational voice data, these works based on deep learning have limitations in the following two aspects. 1) for utterance-level framework, features generated from lots of statistics functions usually concatenated without selection before input to nets, which lead to difficulty for getting satisfied training performance because of high dimension features. Although some dimensionality reduction techniques are explored to solve this problem (Jin et al. 2014; Liu et al. 2017), finding a satisfied strategy is not easy because of some information must be loss from raw features; 2) Previous works primarily focus on datasets with limited amount of labeled data, such as IEMOCAP database (Busso et al. 2008). While in VDAs, we have tremendous amount of users and massive scale conversational voice data which raise difficulty to manually label, so how to make use of labeled data and unlabeled data jointly is a quite crucial factor. As for the works focused on unsupervised learning, their frameworks usually contain two steps: representation learning based on unlabeled data and classifier training based on labeled data. However, these models, overfit easily due to the small amount labeled data

*Corresponding author: J. Jia (jjia@mail.tsinghua.edu.cn)

[1]http://yy.sogou.com
[2]http://www.apple.com/ios/siri/

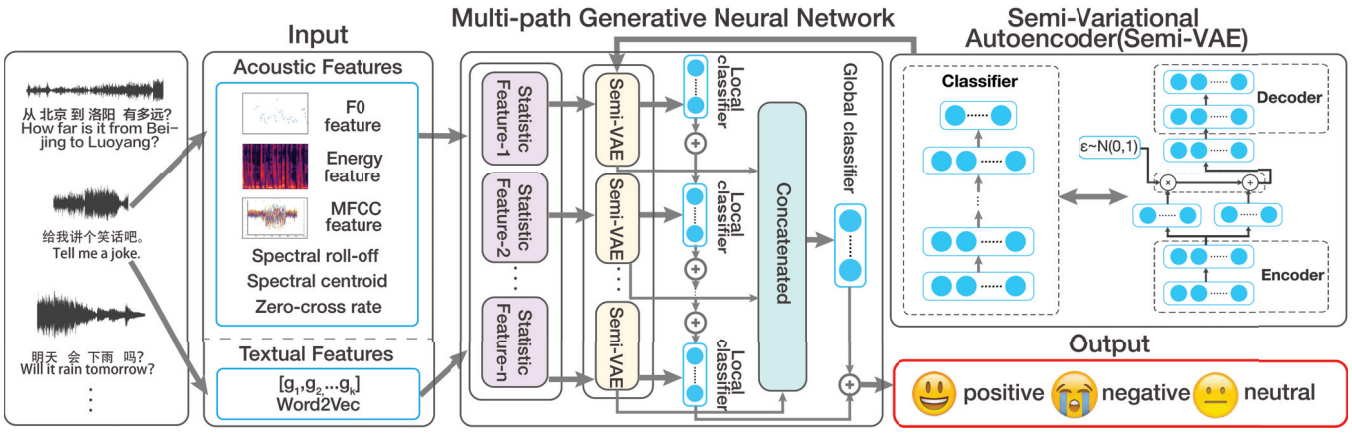

Figure 1: The workflow of our framework.

for classifier training which greatly limit the performances. Therefore, how to utilize those unlabeled data to increase the speech emotion inference accuracy is still a challenge.

In this paper, employing a real-world VDA data, we propose a novel semi-supervised learning scheme with multi-path generative neural network (MGNN) to solve the limitations mentioned above for speech emotion recognition(shown in Figure 1). Specifically, first, we propose a novel supervised multi-path deep neural network framework. Rather than learn a single classifier through the whole features for our task, the whole features are divided into groups based on different LLDs and statistics functions, such as mean of MFCC features. Then, the high-level representation features learned from local classifiers are concatenated to feed into a global classifier. More importantly, local classifiers and global classifier are trained simultaneously through a single objective function. Second, we extend the multi-path deep neural network to a generative model based on semi-supervised variational autoencoder(semi-VAE)(Kingma et al. 2014). Employed as the local part of multi-path method, semi-VAE utilize labeled data to train a classifier and exploit unlabeled data to strengthen the classifier simultaneously. Experiments on benchmark dataset IEMOCAP and real-world voice dataset from Sogou Voice Assistant(SVAD13)(a Chinese Siri) show that our framework is efficient. Especially for IEMOCAP, our method outperform (+11.6% in terms of unweighted accuracy of acoustic and textual feature)the existing state-of-the-art methods.

Our main contributions are summarized as below.

- First, we propose a novel supervised multi-path deep neural network framework. Unlike the existing works that employ the whole features as input and train them in a single classifier, the proposed framework train raw features by groups in local classifiers to avoid high dimensional. Then high-level features of each local classifiers are concatenated as input of a global classifier. More importantly, these two kinds classifiers are trained simultaneously through a single objective function to achieve a more effective and discriminative emotion inferring.
- Second, extending the multi-path deep neural network to a generative model based on semi-VAE, we introduce a semi-supervised multi-path generative framework. By utilizing labeled data and unlabeled data as a joint process for training, to some extent, it help solve the problem of over-fitting easily caused by the small amount labeled data in classifier training.

The rest of paper is organized as follows. Section 2 lists related works. Section 3 formulates the problem. Section 4 presents the methodologies. Section 5 introduces the experiment dataset and results. Section 6 is the conclusion.

2 Related Works

In this section, we briefly review previous methods which are most related to our work including speech emotion recognition and semi-supervised learning.

Speech emotion recognition. Existing works can be sorted into two aspects: utterance-level feature based and frame-level feature based. Researches on utterance-level feature based approaches generate high-level representation from LLDs with lots of statistics functions, and then as input of DNN for classification (Xia and Liu 2017). For frame-level feature based approaches, (Mirsamadi, Barsoum, and Zhang 2017) propose local attention based recurrent neural networks for speech emotion recognition. (Zhang et al. 2016) propose multi-modal learning scheme based on convolutional network for audio-visual emotion recognition. (Trigeorgis et al. 2016) utilize convolutional recurrent network to learn high-level representations for recognition.

semi-supervised learning. As autoencoders have always been a common way to make better use of unlabeled data. (Ghosh et al. 2016) utilize stacked autoencoder to form high-level representation with an unsupervised way, and adopt BLSTM for classification. (Chang and Scherer 2017) learn representations based DCGAN for speech emotion recognition. In these works, two independent parts usually contained: representation learning based on unlabeled data and classifier training based on labeled data, which overfit easily due to the small amount labeled data for classifier training. To our best knowledge, seldom work make use of la-

beled data and unlabeled data as a joint process in training for speech emotion recognition. It is worth noting that great success has been achieved based semi-supervised learning approach, such as variational autoencoder (VAE) (Kingma and Welling 2014), in the area of computer vision and natural language processing. These works prove that the semi-supervised learning based is more efficient than the traditional approach mentioned above.

3 Problem Formulation

Given a set of utterances S, we divide it into two sets S_l (labeled data) and S_u (unlabeled data). For each utterance $x \in S$, we denote $x = \{x^a, x^t\}$. x^a represents the acoustic features of each utterance, which is a n_a dimensional vector. x^t represents the textual features of each utterance, which is a n_t dimensional vector. In addition, X^a is defined as a $|S| * n_a$ feature matrix with each element x^a_{ij} denoting the jth acoustic feature of v_i. The definition of X^t is similar to X^a.

Definition. **Emotion.** We adopt {*Positive*, *Neutral*, *Negtive*} as the emotion space and denote it as E_S, where $S = 3$.

Problem. **Learning task.** Given utterances set V, we aim to infer the emotion for every utterance $x \in S$:

$$f : (S_l, S_u, X^a, X^t) \Rightarrow E_S \tag{1}$$

4 Multi-path Generative Neural Network

In this work, a semi-supervised multi-path generative neural network framework is proposed for inferring emotion from real-world conversational voice data. Specifically, first, to take the feature independency nature into account and avoid high dimensional input, we adopt a supervised multi-path deep neural network(MDNN) framework. Raw features are trained by groups in local classifiers. Then high-level features of each local classifiers are concatenated as input of a global classifier. These two kinds classifiers are trained simultaneously through a single objective function to achieve a more effective and discriminative emotion inferring. Second, to solve the labeled-data-scarcity problem, we extend the MDNN framework to a semi-supervised multi-path generative neural network(MGNN) framework based on semi-VAE, which utilize labeled data and unlabeled data as a joint process for training. The structures of MDNN and MGNN are shown in Figure 2 and Figure 3.

4.1 Multi-path Deep Neural Network

As discussed above, traditionally, for speech emotion recognition, high-level representations generated by applying statistics functions on low-level descriptors (LLDs) with simply concatenating are used as input of deep neural network. These not only ignore the independent nature of each feature but also cause the feature dimension to be too high, which restrict the performance to a great extent.

To solve these, we propose a novel supervised framework for speech emotion recognition, called multi-path deep neural network(MDNN) shown in Figure 2, which trains group-features as local classifiers and their concatenated high-level representation features as global classifier simultaneously through a single objective function.

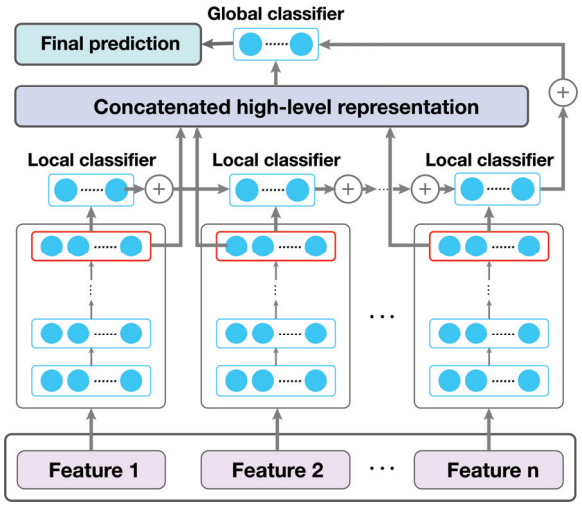

Figure 2: The structure of Multi-path Deep Neural Network.

Rather than learn a single classifier through the whole sample features for our task, the raw features are divided into groups based on different LLDs and statistics functions, such as mean or standard deviation of MFCC features, to learn multiple classifiers. Therefore, each grouped features are trained to obtain the corresponding classifier, we called these are 'local classifiers'. Through 'local classifiers', each grouped features are trained for classification independently, which effectively avoid the problem that the dimensions too high.

It is worth noting that although 'local classifiers' take the independent nature of features into account, they ignore relevance between different features. To solve this problem, we merge the highest hidden layers of 'local classifier' to generate a global representation for training a 'global classifier', which can model the relevance between different feature effectively. More importantly, we optimize the framework through the single objective function as following, which can make us train 'local classifiers' and 'global classifier' simultaneously:

$$\begin{aligned}\mathcal{L}(\Theta, \Phi; \mathbf{x}) &= (1-\lambda)\mathcal{L}_g(\theta_g, \phi_g; \mathbf{x}) \\ &+ \lambda \sum_{i=1}^{N} \mathcal{L}_l(\theta_{l,i}, \phi_{l,i}; \mathbf{x}) \\ &= (1-\lambda)\mathcal{H}(p_{\theta_g}(x), q_{\phi_g}(x)) \\ &+ \lambda \sum_{i=1}^{N} \mathcal{H}(p_{\theta_{l,i}}(x), q_{\phi_{l,i}}(x))\end{aligned} \tag{2}$$

where $\mathcal{L}_g(\cdot)$ is the cost function for 'global classifier' and $\mathcal{L}_l(\cdot)$ is the cost function for 'local classifier'. p_θ is the true distribution of one-hot label and q_ϕ is the approximating distribution. N is the number of 'local classifiers' and λ is the weight coefficient that between 0 and 1. Specially, for $\lambda = 0$, only the 'global classifier' are included in the framework, and for $\lambda = 1$, only 'local classifiers' are included. $\mathcal{H}(\cdot)$ is a function that returns cross-entropy between an approximat-

ing distribution and a true distribution that can be written mathematically:

$$\mathcal{H}(p(x), q(x)) = -\sum_{x} p(x) \log(q(x)) \tag{3}$$

The objective function Eq. 2, can be optimized based on back-propagation algorithm.

4.2 Multi-path Generative Neural Network

In VDAs, we have tremendous amount of users and massive scale conversational voice data which raise difficulty to manually label. However, previous works in speech emotion recognition primarily focus on datasets with limited amount of labeled data. Although some works focus on unsupervised learning which also utilize the unlabeled data to help pre-training model, their frameworks usually train unlabeled data and labeled data separately which overfit easily due to the small amount labeled data for classifier training. Therefore, how to make use of labeled data and unlabeled data jointly is still challenge.

To solve these, we extent the multi-path deep neural network to a semi-supervised framework named multi-path generative neural network(MGNN, shown in Figure 3), which train labeled data and unlabeled data simultaneously based on semi-VAE. As shown in Figure 3, rather than employ DNN, we employ a semi-VAE as the building block for constructing the multi-path neural network. Similar to multi-path deep neural network, group-based features are used as the input of semi-VAEs.

Typically, a Variational Autoencoder(VAE) is a deep generative model (Kingma and Welling 2014) which contains both a probabilistic encoder and decoder with an observed data set $\mathbf{S}$. For $\mathbf{x} \in S$, the encoder network $q_\phi(\mathbf{z} \mid \mathbf{x})$, as an approximation to the true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$, is an inference model to acquire a latent distribution $\mathbf{z}$ from observed datapoint $\mathbf{x}$. The decoder network $p_\theta(\mathbf{x} \mid \mathbf{z})$ is a generative model to get a distribution over the possible corresponding $\mathbf{x}$, given a latent variable $\mathbf{z}$. The whole process can be written mathematicallly as:

$$\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}), \mathbf{x} \sim p_\theta(\mathbf{x} \mid \mathbf{z}) \tag{4}$$

According to (Kingma and Welling 2014), based on SGVB estimator and reparameterization trick, the parameters, θ and ϕ, can be trained simultaneously based on (deep) neural networks (most commonly DNNs or CNNs).

As an extension of VAE, semi-supervised learning method based on semi-VAE was proposed in (Kingma et al. 2014). Besides an encoder and a decoder, a classifier is consisted in the semi-VAE framework and these three parts are trained simultaneously with a single objective function. For the auto-encoder part, two objective functions are utilized for optimization with labeled data and unlabeled data.

Give a labeled data $\mathbf{x}$ and its label y, the evidence lower bound with corresponding latent variable $\mathbf{z}$ is:

$$\begin{aligned} \log p_\theta(\mathbf{x}, y) &\geq E_{q_\phi(\mathbf{z}|\mathbf{x},y)}[\log p_\theta(\mathbf{x} \mid y, \mathbf{z})] + \log p_\theta(y) \\ &\quad - D_{KL}[q_\phi(\mathbf{z} \mid \mathbf{x}, y) \mid\mid p_\theta(\mathbf{z})] \\ &= -\mathcal{L}(\mathbf{x}, y) \end{aligned} \tag{5}$$

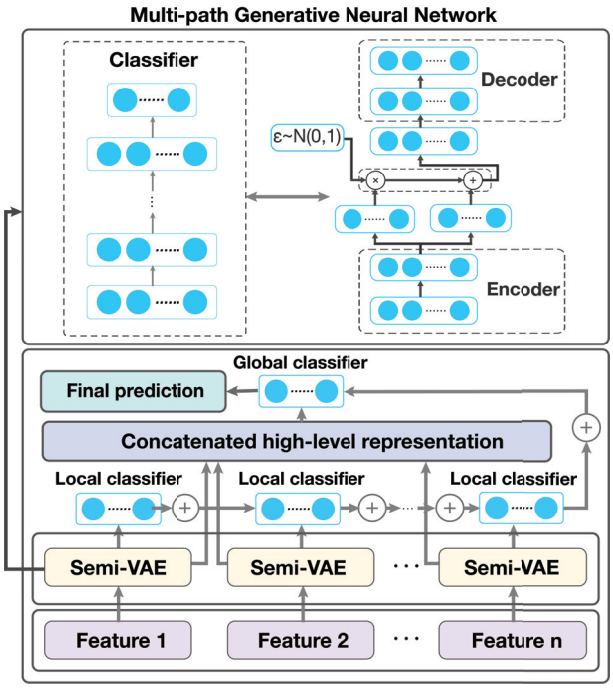

Figure 3: The structure of multi-path generative neural network.

for unlabeled data, the evidence lower bound is:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \sum_{y} q_\phi(y \mid \mathbf{x})(-\mathcal{L}(\mathbf{x}, y)) \\ &\quad - D_{KL}(q_\phi(y, \mathbf{z} \mid \mathbf{x}) \mid\mid p_\theta(\mathbf{z})) \\ &= \sum_{y} q_\phi(y \mid \mathbf{x})(-\mathcal{L}(\mathbf{x}, y)) + \mathcal{H}(q_\phi(y \mid \mathbf{x})) \\ &= -\mathcal{U}(x) \end{aligned} \tag{6}$$

Then,the final objective function is:

$$\begin{aligned} \mathcal{J} = &\sum_{(\mathbf{x},y)\in\mathcal{S}_l} \mathcal{L}(\mathbf{x}, y) + \sum_{\mathbf{x}\in\mathcal{S}_u} \mathcal{U}(\mathbf{x}) \\ &+ \alpha E_{(\mathbf{x},y)\in\mathcal{S}_l}[-\log q_\phi(y \mid \mathbf{x})] \end{aligned} \tag{7}$$

Specifically, the first item is the loss of labeled data in auto-encoder part defined as Eq. 5, the second item is the loss of unlabeled data in auto-encoder part defined as Eq. 6, and the third item is the loss for classifier part to learn the distribution $q_\phi(y \mid \mathbf{x})$ from the labelled data. α is a hyper-parameter that controls the weights of classification loss and we use $\alpha = 0.1 * |S|$ in all experiments. S_l and S_u are the dataset for labeled and unlabeled data respectively.

As shown in Figure 3, for MGNN, we utilize the semi-VAEs as the building blocks. The objective of MGNN is minimize the function as following:

$$\begin{aligned} \mathcal{J}_{MGNN} &= (1-\lambda) E_{(\mathbf{x},y)\in\mathcal{S}_l}[-\log q_\phi(y_{global} \mid \mathbf{x})] \\ &\quad + \lambda \sum_{i=1}^{N} \mathcal{J}_{local,i} \end{aligned} \tag{8}$$

Table 1: The number of utterances for each emotion category.

Emotion	Happy	Anger	Sad	Neutral	Total
Utterances	1636	1103	1084	1708	5531
(%)	29.6	19.9	19.6	30.9	-

where $\mathcal{J}_{local,i}(\cdot)$ is the cost function for 'local classifiers' , which is defined as Eq. 7. N is the number of 'local classifiers'. The first term is the loss of classifier part produced by 'global classifier'. λ is the weight coefficient between 0 and 1. Specially, for $\lambda = 0$, only the 'global classifier' are included in the framework, and for $\lambda = 1$, only 'local classifiers' are included. The whole framework can be optimized by AEVB algorithm with SGVB estimator (Kingma and Welling 2013).

5 Experiments

5.1 Dataset Details

SVAD13 We establish a corpus of voice data from Sogou Voice Assistant [1] (Chinese Siri)(SVAD13) containing 24,000 Mandarin utterances recorded in 2013. Every utterance is assigned with its corresponding speech-to-text information provided by Sogou Corporation. Due to the massive scale of our dataset, manually labeling the emotion for every utterance is not practical. Thus we randomly select 2,000 utterances from the dataset and invite three well-trained people to annotate the emotion. The annotators are asked to label the emotion by listening to the utterances and reading corresponding words simultaneously. The annotators stop and discuss when they can't reach consensus. If they still cannot reach an agreement, the utterance will be discarded. The emotion distributions of these utterances are: *Neutral: 43.75%*, *Positive: 18.5%*, *Negtive: 38.2%*. Besides, randomly selected 24000 unlabeled data are employed to do model training in our experiment.

IEMOCAP The IEMOCAP (Busso et al. 2008) database have been widely used for evaluating speech emotion recognition systems. It contains approximately 12 hours of audio-visual conversations in English, and the conversations are manually segmented into utterances. The categorical labels of the utterances are as follows: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and others. In our experiment, to compare with the state of the art as mentioned beforemerging the happiness and excitement categories as the happy category, we form a four-class emotion classification dataset containing {happy, angry, sad and neutral}. Table 1 presents the detail utterance number of each category. There are in total 5531 utterances.

5.2 Feature Extraction

Acoustic Feature We utilize openSMILE toolkit (Eyben, Wöllmer, and Schuller 2010) to extract acoustic features for SVAD13 and IEMOCAP. Totally, we obtain 1,582 statistic acoustic feature, which is the same as the acoustic features used in the INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al. 2010).

Textual feature As for the textual information in real-world chinese database SVAD13. Thulac Tool (Li and Sun 2009), an efficient Chinese word segmentation is used to get words of an utterance. Then word embeddings is learned with word2vec (Mikolov et al. 2013). Specifically, we use the whole 31.2 million chinese word corpora collected from the 7.5 million utterance from SVAD13 as the training corpora for word2vec. As for the textual information in IEMOCAP, we adopt the publicly available 300-dimensional word2vec vectors, which are trained on 100 billion words from Google News(Mikolov et al. 2013) to represent word vector. Then, we extract 4200-dimensional utterance-level textual features according to the statistic functions (mean, std, disp, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3, skewness, kurtosis) over the LLDs.

5.3 Experimental setup

Evaluation metrics. In all the experiments, we evaluate the performance in terms of F1-measure (Powers 2011), Unweighted accuracy(UA), Weighted accuracy(WA) (Rozgic et al. 2012). The results reported in SVAD13 are based on 5-fold cross validation. To compare with the state of the art, the results reported in IEMOCAP are based on 10-fold leave-one-speaker-out(LOSO) cross-validation.

$$F1\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

$$UA = \sum_{i} \frac{correct\ utterances\ for\ emotion\ i}{utterances\ for\ emotion\ i} \tag{10}$$

$$WA = \frac{correct\ utterances}{utterances} \tag{11}$$

5.4 Performance of MDNN

To evaluate the effectiveness of our proposed supervised multi-path neural network approach, we compare the performance of emotion classification with some baseline methods based on SVAD13 and IEMOCAP. The comparison methods are as follows:

Deep neural network (DNN)(Ren et al. 2014): The whole features are putted together into one classifier to get prediction.

Global Deep neural network(global): Features are grouped into different local classifiers based on DNN, and we utilize the concatenated high-level representation features of each local classifiers as the input of a global classifier to get the final result.

Local Deep neural network(local): Features are also grouped into different local classifiers based on DNN for training while the final predict is calculated by the mean of all single local classifiers.

Our proposed multi-path Deep neural network(multi-path): Local classifiers and global classifiers are trained simultaneously through a single objective function. The final result is calculated by combining results of local classifiers and global classifiers.

Table 2: Comparison for different supervised method on SVAD13 and IEMOCAP with different features. A:acoustic. T:text.

	Method	SVAD13				IEMOCAP			
		DNN	global	local	multi-path	DNN	global	local	multi-path
UA	A	49.3	48.6	53.3	**53.7**	58.4	59.8	61.6	**62.7**
	T	55.9	56.0	58.4	**58.5**	59.9	60.3	66.8	**66.9**
	A+T	57.5	59.0	59.8	**61.3**	72.8	70.0	75.8	**76.7**
WA	A	51.3	51.0	54.4	**54.6**	57.6	57.8	60.9	**61.8**
	T	57.4	58.0	60.8	**61.2**	58.8	59.1	65.5	**65.8**
	A+T	58.5	61.1	59.5	**61.7**	71.1	68.3	74.6	**75.2**
F1	A	48.3	48.4	52.3	**52.8**	57.6	57.9	60.7	**61.9**
	T	55.4	55.8	58.4	**58.7**	58.3	58.7	65.6	**65.8**
	A+T	56.8	58.8	58.5	**60.5**	71.8	68.7	75.1	**75.6**

Table 3: The performance on IEMOCAP dataset with different features and comparison with the state of the art. A:acoustic. T:text.

	Method	A(%)		T(%)		A+T(%)	
		UA	WA	UA	WA	UA	WA
SVM	[APSIPA ASC, 2012]	60.9	60.8	48.6	48.5	67.4	67.4
SVM	[ICASSP, 2015]	-	-	-	-	69.2	-
RNN	[ICASSP, 2017]	58.8	**63.5**	-	-	-	-
CNN	[ICDM, 2016]	61.3	-	59.3	-	65.1	-
LSTM	[ACL, 2017]	57.1	-	**73.1**	-	75.6	-
MDNN	Our Method	**62.7**	61.8	66.9	**65.8**	**76.7**	**75.2**

As shown in Table 2, the performance of 'local' method including grouped features and separate predictions are generally better than DNN which can be proved by the obvious improvement of UA, WA and F1. For 'feature A', 'local' method(52.3%) is +4% than DNN(48.3%) for SVAD13 in F1 measures. However, the 'global' method, which is also with grouped features but predicted by the concatenated high-level representation, have a better performance +2% compared with DNN in SAVD13, while DNN performs better in IEMOCAP. Therefore, 'global' method may be not stable enough for different dataset. Furthermore, taking 'feature A', 'feature T' and 'feature A+T' all into consideration, multi-path has better performance than any previous methods. Specifically, for UA, about 4.4% enhancement has been accomplished in 'feature A' in SVAD13. As for WA, the performance of multi-path is almost 3.2% higher than DNN(61.7 VS 58.5 for 'feature A+T'). It verifies that training local classifiers and global classifier simultaneously through a single objective function is a more effective way to take the feature independency and relavance into consideration and improve the emotion recognition performance.

Comparison to the state-of-art method To demonstrate the comparability and the adaptability of our proposed MDNN, we also compare the performance on the public dataset IEMOCAP (Busso et al. 2008) with some state-of-art methods. The comparison methods are as follows:

[APSIPA ASC, 2012] (Rozgic et al. 2012) propose a ensemble of trees of binary SVM classifiers to address the sentence-level multimodal emotion recognition problem.

[ICASSP, 2015] (Jin et al. 2015) This paper generate different kinds of acoustic and lexical features in utterance level and combine them via early fusion and late fusion to recognize emotion with a SVM classifier.

[ICDM, 2016] (Poria et al. 2016) feed features extracted by deep convolutional neural networks(CNN) into a multiple kernel learning classifier to do multimodal emotion recognition.

[ICASSP, 2017] (Mirsamadi, Barsoum, and Zhang 2017) This paper study automatically discovering emotionally relevant speech features using a deep recurrent neural network(RNN) and a local attention base feature pooling strategy.

[ACL, 2017] (Poria et al. 2017) This paper propose a LSTM-based model to capture contextual information between utterance-level features in the same video.

Table 3 shows the unweighted accuracy(UA) and weighted accuracy(WA) of competitive methods and our proposed MDNN. While comparing the performance 'feature A+T', our proposed method outperforms all the baseline methods that are state-of-the-art. Especially, for the UA of the 'feature A+T', +7.5% compared with [ICASSP , 2015] using SVM, +11.6% compared with [ICDM, 2016] using CNN and +1.1% compared with [ACL, 2017] using LSTM. As for UA of 'feature A only', it shows that MDNN (62.7%) is +1.8% compared with [APSIPA ASC, 2012] using SVM, +3.9% compared with [ICASSP, 2017] with RNN. These strongly demonstrate the effectiveness of the supervised part of our proposed method.

5.5 Performance of MGNN

To demonstrate the semi-supervised part of our proposed method, we make comparisons among the performance of different autoencoder pretraining strategy Stacked Autoencoder(SAE) (Vincent et al. 2010), variational autoencoder

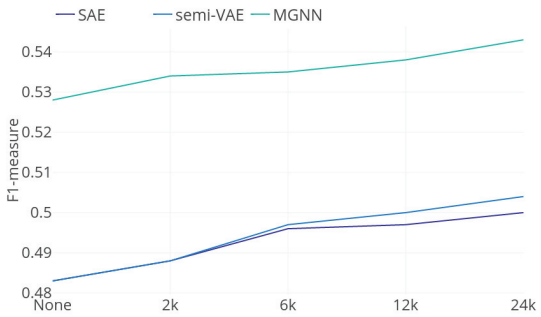

Figure 4: Performance of the semi-supervised methods with different amount of unlabeled data on SVAD13 dataset of acoustic features.

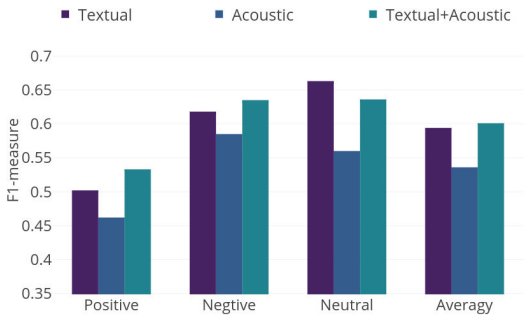

Figure 5: Feature contribution analysis.

(VAE)(Kingma and Welling 2014) with different amount of unlabeled data on SVAD13 dataset. As show in Table 4, VAE outperform SAE on different unlabel data size. These verifies VAE which train the labeled and unlabeled data simultaneously be a more effective learning method to solve the problem of emotion inferring from the real-world voice data with limited labeled data. Furthermore, we find that the MGNN which employ semi-VAE as the local part of multi-path framework, has the best performance. These proof that our proposed MGNN method which combine semi-VAE and multi-path method do have an effective impact on the real-world voice dataset emotion inferring.

5.6 Analysis

unlabeled data size To verify the effectiveness of the unlabeled data, we test different size of unlabeled data for MGNN on SVAD13. In Figure 4, as the scale of unlabeled data increase, the performance gets better gradually which imply that the semi-supervised learning continues to help improve the model's ability to infer emotion from conversational voice data.

feature contribution analysis We discuss the contributions of acoustic and textual features. The F1-measure for 3 emotion categories and their average are shown in Figure 5 on SVAD13. Specifically, for all these we adopt the MDNN to calculate the performance. As in Figure 5, the performance of 'Textual Only' is better than 'Acoustic Only' in SVAD13, which indicates that the textual information can contribute more to the emotion recognition in the real world VDAs. Moreover, 'T+A' which contains both Textual information and acoustic information performs best. Specifically, for the Positive emotion, 'T+A' +3.1% compared with 'Textual Only', +7.1% compared with 'Acoustic Only', and for Negative emotion, 'T+A' +1.7% compared with 'Textual Only', +5.0% compared with 'Acoustic Only'. These convince that utilize the two modalities simultaneously can be more effective to infer emotional utterances.

Table 4: Performance of the methods about different amount of unlabeled data with acoustic features on SVAD dataset.

Method	None	2k	6k	12k	24k
SAE	48.3	48.8	49.6	49.7	50.0
semi-VAE	48.3	48.8	49.7	50.0	50.4
MGNN	**52.8**	**53.4**	**53.5**	**53.8**	**54.3**

6 Conclusion

In this paper, to study the problem of inferring emotion from conversational voice data, we propose a semi-supervised multi-path generative neural network. Our main contribution are as follows: first, to take the feature independency nature into account and avoid high dimensional input, we adopt a supervised MDNN framework. Raw features are trained by groups in local classifiers. Then high-level features of each local classifiers are concatenated as input of a global classifier. These two kinds classifiers are trained simultaneously through a single objective function to achieve a more effective and discriminative emotion inferring. Second, to solve the labeled-data-scarcity problem, we extend the MDNN to a semi-supervised MGNN framework based on semi-VAE, which utilize labeled data and unlabeled data as a joint process for training. As shown in the experiment results based on real-world VDA data SVAD13 and a public dataset IEMOCAP, our proposed MGNN turns out to be effective in speech emotion inferring. Furthermore, our work can be well utilized in real-world applications. For instance, we can provide emotional response in the VDAs, which contributes to more humanized intelligent service.

7 Acknowledgments

This work is supported by National Key Research and Development Plan (2016YFB1001200),the Innovation Method Fund of China (2016IM010200), the National Natural, and Science Foundation of China (61370023 61602033) and Tiangong Institute for Intelligent Computing, Tsinghua University.

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

Chang, J., and Scherer, S. 2017. Learning representations of emotional speech with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1705.02394*.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462. ACM.

Ghosh, S.; Laksana, E.; Morency, L.-P.; and Scherer, S. 2016. Representation learning for speech emotion recognition. In *INTERSPEECH*, 3603–3607.

Jin, Y.; Song, P.; Zheng, W.; and Zhao, L. 2014. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 4808–4812. IEEE.

Jin, Q.; Li, C.; Chen, S.; and Wu, H. 2015. Speech emotion recognition with acoustic and lexical features. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 4749–4753. IEEE.

Kim, Y.; Lee, H.; and Provost, E. M. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 3687–3691. IEEE.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *Conference proceedings: papers accepted to the International Conference on Learning Representations (ICLR) 2014*.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 3581–3589.

Li, Z., and Sun, M. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics* 35(4):505–512.

Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; and Tan, G.-Z. 2017. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mirsamadi, S.; Barsoum, E.; and Zhang, C. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2227–2231. IEEE.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, 439–448. IEEE.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 873–883.

Powers, D. M. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Ren, Z.; Jia, J.; Guo, Q.; Zhang, K.; and Cai, L. 2014. Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, 1–4. IEEE.

Rozgic, V.; Ananthakrishnan, S.; Saleem, S.; Kumar, R.; and Prasad, R. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 1–4. IEEE.

Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G.; and Wendemuth, A. 2009. Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 552–557. IEEE.

Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; and Narayanan, S. S. 2010. The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Schuller, B.; Rigoll, G.; and Lang, M. 2003. Hidden markov model-based speech emotion recognition. In *International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings*, I–401–4 vol.1.

Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M. A.; Schuller, B.; and Zafeiriou, S. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 5200–5204. IEEE.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research* 3371–3408.

Wu, B.; Jia, J.; He, T.; Du, J.; Yi, X.; and Ning, Y. 2016. Inferring users' emotions for human-mobile voice dialogue applications. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, 1–6. IEEE.

Xia, R., and Liu, Y. 2017. A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing* 8(1):3–14.

Zhang, S.; Zhang, S.; Huang, T.; and Gao, W. 2016. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 281–284. ACM.