Inferring Emotion from Large-scale Internet Voice Data: A Semi-supervised Curriculum Augmentation based Deep Learning Approach

Suping Zhou[1], Jia Jia[12*], Zhiyong Wu[12], Zhihan Yang[12], Yanfeng Wang[3], Wei Chen[3], Fanbo Meng[3], Shuo Huang[1], Jialie Shen[4], Xiaochuan Wang[3]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Beijing National Research Center for Information Science and Technology (BNRist)
The Institute for Artificial Intelligence, Tsinghua University
[2] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[3]Sogou Corporation, Beijing, China, [4]Queen's University Belfast, Belfast, U.K
jjia@mail.tsinghua.edu.cn

Abstract

Effective emotion inference from user queries helps to give a more personified response for Voice Dialogue Applications(*VDAs*). The tremendous amounts of VDA users bring in diverse emotion expressions. How to achieve a high emotion inferring performance from large-scale Internet Voice Data in VDAs? Traditionally, researches on speech emotion recognition are based on acted voice datasets, which have limited speakers but strong and clear emotion expressions. Inspired by this, in this paper, we propose a novel approach to leverage acted voice data with strong emotion expressions to enhance large-scale unlabeled internet voice data with diverse emotion expressions for emotion inferring. Specifically, we propose a novel semi-supervised multi-modal curriculum augmentation deep learning framework. First, to learn more general emotion cues, we adopt a curriculum learning based epoch-wise training strategy, which trains our model guided by strong and balanced emotion samples from acted voice data and sub-sequently leverages weak and unbalanced emotion samples from internet voice data.Second, to employ more diverse emotion expressions, we design a Multi-path Mix-match Multimodal Deep Neural Network(MMMD), which effectively learns feature representations for multiple modalities and trains labeled and unlabeled data in hybrid semi-supervised methods for superior generalisation and robustness. Experiments on an internet voice dataset with 500,000 utterances show our method outperforms (+10.09% in terms of F1) several alternative baselines, while an acted corpus with 2,397 utterances contributes 4.35%. To further compare our method with state-of-the-art techniques in traditionally acted voice datasets, we also conduct experiments on public dataset IEMOCAP. The results reveal the effectiveness of the proposed approach.

Introduction

Driven by fast development of deep learning techniques, smart Voice Dialogue Application (VDAs) brings great convenience to our daily life. Effective emotion inference from user queries can be very helpful to understand users' real meanings and intents, and subsequently provide anthropomorphic answers. While the tremendous amounts of uncertain speakers with a great diversity of dialects, expression preferences in VDAs benefit training in the deep learning framework, it is still very difficult to accurately and comprehensively infer emotion due to the weak and unbalanced emotion expressions. As reported by related works, the state of the art techniques always suffer from low and instable performance when facing large-scale internet voice data from VDAs(Wu et al. 2016) (Zhou et al. 2018b).

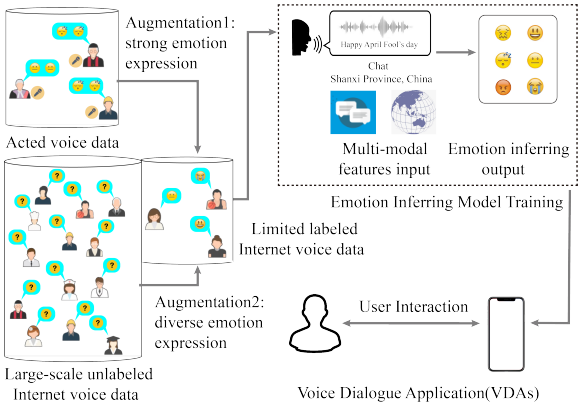


Figure 1: The logical workflow of our framework.

Driven by demand from real applications, considerable research development in speech emotion recognition has been witnessed in previous decades. (Aguilar et al. 2019) propose a hierarchical multimodal model including Modality-based attention for multimodal emotion recognition. (Zhang et al. 2019) propose a f-Similarity Preservation Loss for deep metric learning with soft labels, and apply the proposed methods on the task of cross-corpus speech emotion recognition. While these works based on acted voice data achieve good performances inner dataset due to strong and clear emotion expressions, they have limited generalization and robustness

*Corresponding author: J. Jia (jjia@mail.tsinghua.edu.cn)
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

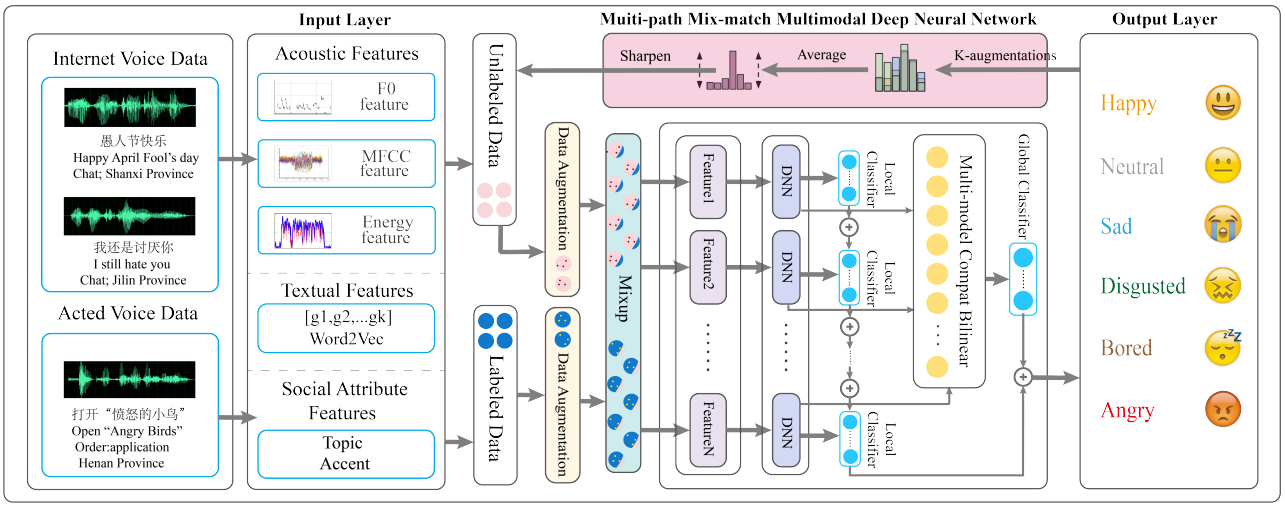


Figure 2: The workflow of our framework.

for real-world applications. Inspired by the contrast between the large scale internet voice data from VDAs and the acted emotional voice dataset, can we use one's strengths to compensate for the other's weakness? However, two main challenges remain unsolved: 1) how to effectively leverage acted voice dataset with strong and clear emotion expressions to enhance internet voice data? 2) how to utilize large-scale unlabeled data with diverse user emotion expressions to augment few labeled data.

In this paper, via leveraging a real-world voice dataset from Sogou Voice Assistant containing 500,000 utterances assigned with its corresponding speech-to-text information and social attributes (user location, query topic, etc. provided by [1]), we propose a novel semi-supervised multimodal curriculum augmentation deep learning framework. First, to learn more general emotion cues, we adopt a curriculum learning based epoch-wise training strategy. In each epoch, model is pre-trained by strong and balanced emotion samples from acted voice data and sub-sequently leverage weak and unbalanced emotion samples from internet voice data. Second, to employ more diverse emotion expressions, we design a Multi-path Mix-match Multimodal Deep Neural Network(MMMD). Specifically, to effectively learn feature representations for multiple modalities, we propose a Multi-path Multimodal Deep Neural Network(MMD) to integrate multiple modalities. Extending MMD to a semi-supervised model based on Mix-match, we introduce MMMD to achieve superior generalization and robustness. Experiments on an internet voice dataset with 500,000 utterances collected from Sogou Voice Assistant[1] show our method outperforms (+10.09% in terms of F1) several alternative baselines, while an acted corpus with 2,397 utterances contributes 4.35%. To further compare our method with state-of-the-art techniques in traditionally acted voice datasets, we also conduct experiments on public dataset IEMOCAP. The results reveal the effectiveness of the proposed approach. The illustration of our proposed novel approach is shown in Figure 1 and Figure 2.

Our main contributions are summarized as below.

- We successfully leverage the traditional acted voice dataset to enhance the emotion inferring results from the large-scale internet voice data in VDAs. The proposed curriculum learning based epoch-wise training strategy can well integrate the strong and clear emotion expressions with the users' diversity.
- To utilize large-scale unlabeled data to augment few labeled data, we propose a Multi-path Mixmatch multimodal deep learning method (MMMD). This semi-supervised framework enables us to learn effective feature representations for multiple modalities (acoustic, textual and social information) and promote the generalisation and robustness for emotion inferring in VDAs.

The rest of paper is organized as follows. Section 2 gives a comprehensive review on related works. Section 3 formulates the problem. Section 4 presents the key methodologies and core algorithms. Section 5 introduces the experimental configuration and core test results. Section 6 concludes the study with summary.

Related Work

Inferring Voice Emotion. In terms of emotion analysis for voice, previous works have focused primarily on extracting effective features and utilizing diverse types of learning methods(Neumann and Vu 2019)(Freitag et al. 2017)(Zhang et al. 2019). However, all these researches mainly focused on inferring emotions from acted corpora data, few have been done to address the problem for real-world large-scale internet voice data with weak emotion expressions and tremendous uncertain speakers. It is potential to transfer the emphasis on emotion recognition in the wild and assist this work through the augmentation of acted corpus.

Curriculum Learning. Curriculum learning is training strategy to learn from simple to complex and proved to

[1]http://yy.sogou.com

achieve great improvements in generalization and speed of convergence(Bengio et al. 2009). It is natural to apply curriculum learning to emotion recognition since we learn to perceive emotions gradually from infant to adulthood. Previous work on speech emotion recognition has utilize curriculum learning to solve the problem of Crowd-sourced Labels and achieve improvements (Lotfian and Busso 2019). In this paper, for a new task to learn more general emotion cues from strong emotion samples, we design a curriculum learning based epoch-wise training strategy, in which classifier is guided by strong and balanced emotion samples first and sub-sequently leverage weak and unbalanced emotion samples.

Semi-supervised Learning. Autoencoders have always been a common way to make better use of unlabeled data in speech emotion recognition.(Deng et al. 2017)(Jia et al. 2018). Some Generative and Adversarial Networks(Semi-VAE (Zhou et al. 2018b), DCGAN(Chang and Scherer 2017), ADDoG(Gideon, McInnis, and Provost 2019) are also utilized to make improvements. However, these works mainly adopt single Semi-supervised learning(SSL) method. (Berthelot et al. 2019) propose a hybrid method named Mix-match which combines several ideas and components from the current dominant paradigms for SSL. This idea of hybrid SSL methods achieves great success in several image classification and facial expression recognition tasks. In this paper, we introduce the hybrid SSL methods in our framework to solve the diversity of user dialects and expression preferences and to use the large-scale unlabeled data of VDAs to learn some universal pattern to enhance the performance of classification in emotion.

Problem Formulation

Given a set of utterances V. For each utterance $v \in V$, we denote $v = \{x^a, x^t, x^c\}$. 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. x^c represents the social attributes features of each utterance, which is a N_c dimensional vector. In addition, X^a is defined as a $|V| * N_a$ feature matrix with each element x^a_{ij} denoting the jth acoustic feature of v_i. The definition of X^t and X^c is similar to X^a.

The study involves two emotion datasets V_t and V_e, corresponding to two different recording environments. V_e refers to acted dataset with strong and clear emotion expressions. V_t refers to real world voice dataset with large amount of speakers. Specifically, we divide V_t into two sets V_t^L(labeled data) and V_t^U(unlabeled data).

Definition. **Emotion.** Some previous researches (Ren et al. 2014a) (Jia et al. 2018) discover that, emotion categories about human-mobile interaction are different from theories about facial expression(Ekman and Friesen 1969). According to their findings, we adopt {*Neutral*, *Sadness*, *Disgust*, *Anger*, *Happiness*, *Boredom*} as the emotion space and denote it as E_S, where $S = 6$.

Problem. **Learning task.** Given utterances sets V_t and V_e, we aim to infer the emotion for every utterance $v \in V_t$:

$$f : (V_t^L, V_t^U, V_e) \Rightarrow E_S \quad (1)$$

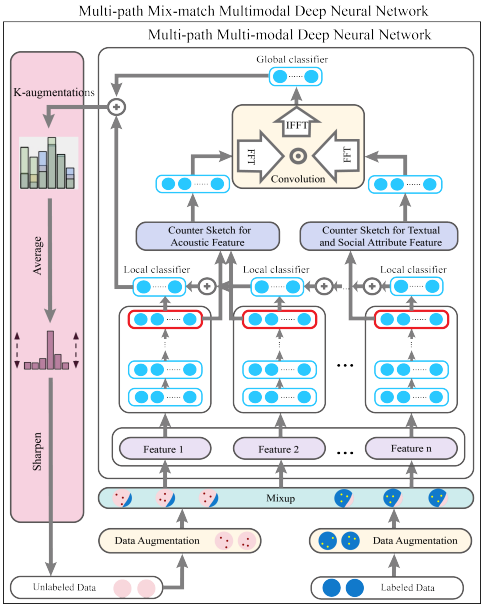


Figure 3: The structure of MMD and MMMD.

Methodology

In order to leverage acted voice data with strong emotion expressions to assist large-scale unlabeled internet voice data with diverse emotion expressions for emotion inferring, we formulate it to three tasks: 1) we design a supervised Multi-path Multimodal Deep Neural Network(MMD) to effectively learn feature representation for multiple modalities. 2)we design a Multi-path Mix-match Multimodal Deep Neural Network(MMMD) to employ more diverse emotion expressions from large-scale internet voice data. 3) we adopt a curriculum augmentation based epoch-wise training strategy to learn more general emotion cues from acted voice data with strong emotion. The structures of the proposed MMD and MMMD are shown in Fig. 3 and the epoch-wise training strategy is shown in Fig. 4.

Multi-path Multimodal Deep Neural Network

We adopt a multi-path solution to model the complex intra-modality relationship which balances both the independencies and dependencies of multi-modal features. Specifically, in our task, we first divide the raw features into groups based on different low-level descriptors(LLDs), and different modalities, such as mean for low-level acoustic features. Then, each group of feature of different modalities are feeded into different classifiers, which is called *local classifier*. With the approach, the problem of high-dimensional inputs can be effectively avoided. Then, we merge the highest hidden layers of each *local classifier* to generate a global representation by an effective approach, Multimodal Compact Bilinear pooling (MCB) (Fukui et al. 2016), to get a good joint representation.

The MCB method not only inherits the advantage of bilinear models, allowing all elements of both acoustic, textual and social features to interact with each other by matrix mul-

tiplying, but also overcomes its disadvantage of high memory consumption and an excessive amount of parameters. The original form of the multi-modal fusion in bilinear models is outer product. To project the outer product to a lower dimensional space and avoid computing the outer product directly, we transform it in the count sketch approach,

$$\Phi(x_1 \bigotimes x_2, h, s) = \Phi(x_1, h, s) * \Phi(x_2, h, s) \quad (2)$$

where $*$ donates the convolution operator and $\bigotimes$ donates outer product. $x_1 \in R^{n_1}$ and $x_2 \in R^{n_2}$ are two vectors to calculate their outer product. For the Count Sketch function $\Phi(x, h, s)$, it projects an input vector $x \in R^n$ to an output $y \in R^d$,

$$y[h[i]] = y[h[i]] + s[i] \cdot x[i] \quad (3)$$

where $[i]$ denotes the ith feature of a vector. $s \in \{-1, 1\}^n$ and $h \in \{1, ..., d\}^n$ are two randomly initialized vectors. Additionally, according to the convolution theorem states, for $y_1 \in R^{d_1}$ and $y_2 \in R^{d_2}$, $y_1 \odot y_2$ can be rewritten as $FFT^{-1}(FFT(y_1) \odot (FFT(y_2))$, where $\odot$ refers to element-wise product. y_1 and y_2 are the Count Sketch results of x_1 and x_2.

In our model, the multi-path *local classifers* generate groups of high-level acoustic hidden features $h_a^1, h_a^2, ..., h_a^{l_1}$ and concatenated textual and social features$h_{tc}^1, h_{tc}^2, ..., h_{tc}^{l_2}$, where l_1 and l_2 refers to the numbers of the acoustic and textual-social paths. We first concatenate the vectors from the same modality and produce two main feature vector h_a and h_{tc}.

$$h_a = Concatenate([h_a^1, h_a^2, ..., h_a^{l_1}]) \quad (4)$$

$$h_{tc} = Concatenate([h_{tc}^1, h_{tc}^2, ..., h_{tc}^{l_2}]) \quad (5)$$

Instead of concatenating the two vectors to get subsequent predictions, we combine them by MCB method into a single vector of feature $h \in R^d$, where d is a hyperparameter affecting the amount of the information compressed.

$$h = MCB(h_a, h_{tc}) \quad (6)$$

Then, this h is applied to train a *global classifer*. Moreover, the *local classifers* and *global classifer* are trained simultaneously through a single objective function. Finally, we get the weighted emotion prediction from all of the local classifiers and global classifier.

Multi-path Mix-match Multimodal Deep Neural Network

To employ more diverse emotion expressions from large-scale internet voice data, we employ hybrid semi-supervised methods from the idea of MixMatch(Berthelot et al. 2019) which achieve success in computer vision.

First, we perform the SSL method of **data augmentation** by adding a Gaussian noise to the multi-modal features and generate the K augmentations $\hat{v}_{t,k}^U$ from unlabeled data ${v_t}^U$. We produce a guessed label q by

$$q = \frac{1}{K} \sum_{k=1}^{K} P(y|\hat{v}_{t,k}^U; \theta) \quad (7)$$

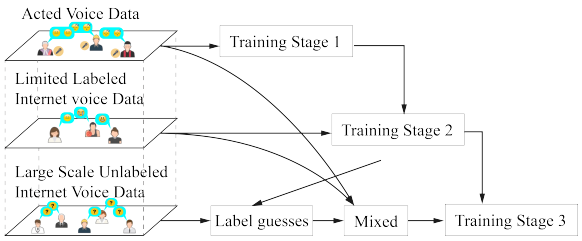


Figure 4: Learning Strategy in an epoch.

$P(y|{V_t}^U; \theta)$ refer to a model which produces a distribution over class labels y for input ${V_t}^U$ with parameters θ. In our work, this generic model is our proposed supervised component MMD.

To encourage the model to give predictions at **entropy minimization** in semi-supervised learning, the sharpening function is applied to reduce the entropy of the label distribution in each path.

$$Sharpen(q, T) = q^{\frac{1}{T}} / \sum_{j=1}^{L} q_j^{\frac{1}{T}} \quad (8)$$

where T is a hyper-parameter related to temperature to control the entropy of the label distribution and q is the guessed label distribution produced as Eq.7. L is the number of emotion categories. For lower-entropy predictions we need to keep T under 1 and close to 0. We apply the sharpen function to model's predictions for the augmented unlabeled data, to encourage the model polarizing its predictions.

Then, we fuse the data with ground truth label and guessed label by applying a regularization SSL method **MixUp**. We get a parameter λ from the Beta distribution and fuse the labeled and unlabeled data by the weight of λ.

$$\lambda \sim Beta(0.75, 0.75) \quad (9)$$

$$\lambda' = max(\lambda, 1 - \lambda) \quad (10)$$

$$x' = \lambda' x_1 + (1 - \lambda') x_2 \quad (11)$$

$$p' = \lambda' p_1 + (1 - \lambda') p_2 \quad (12)$$

where (x, p) represent a pair of data and its label, and we utilize data from (x_1, p_1) and (x_2, p_2) to compute the new pair(x', p'). The Beta distribution with weight (0.75,0.75) ensures the λ is close to either 0 or 1, so that the data mixed up mostly resemble x_1 or x_2.

To apply the above MixUp method in our work, we concatenated the data X' augmented from labeled data $X = V_t^L \cup V_e$ and the data U' augmented from the unlabeled data V_t^U to produce data W. Then we randomly seize data from W by the length of X' as W_1 and the rest of W as W_2. X'' and ${V_t}^{U''}$ are generated by Mixup as follows.

$$X' = Augment(X) \quad (13)$$

$${V_t}^{U'} = Augment({V_t}^U) \quad (14)$$

$$W = Shuffle(Concatenate(X', {V_t}^{U'})) \quad (15)$$

$$X'' = MixUp(X', W_1) \quad (16)$$

$${V_t}^{U''} = MixUp({V_t}^{U'}, W_2) \quad (17)$$

X'' and ${V_t}^{U''}$ are then concatenated as new training data for next epoch training.

Training Strategy Design

It is natural to apply curriculum learning to our task since we learn to perceive emotions gradually from infant to adulthood. In our work, as shown in Fig. 4, we adopt a epoch-wise training strategy to learn more general emotion cues from acted voice data with strong emotion, which train our model guided by strong and balanced emotion samples and sub-sequently leverage weak and unbalanced emotion samples. Specifically, in each epoch, we first train the data from acted voice data V_e. Then we train the data from the labeled internet voice data ${V_t}^L$. Thirdly, we training the model by mixed labeled X'' and unlabeled data${V_t}^{U''}$.

Finally, we induce the combined loss function $\mathcal{L}$.

$$\mathcal{L}_{V_e} = \frac{1}{|V_e|} \sum_{x,p \in V_e} H(p, P(y|x;\theta)) \tag{18}$$

$$\mathcal{L}_{{V_t}^L} = \frac{1}{|{V_t}^L|} \sum_{x,p \in {V_t}^L} H(p, P(y|x;\theta)) \tag{19}$$

$$\mathcal{L}_X = \frac{1}{|X''|} \sum_{x,p \in X''} H(p, P(y|x;\theta)) \tag{20}$$

$$\mathcal{L}_U = \frac{1}{L|{V_t}^{U''}|} \sum_{u,q \in {V_t}^{U''}} ||q - P(y|u;\theta)||_2^2 \tag{21}$$

$$\mathcal{L} = \mathcal{L}_{V_e} + \mathcal{L}_{{V^L}_t} + \mathcal{L}_{\mathcal{X}} + \lambda_U \mathcal{L}_U \tag{22}$$

where p is the true distribution of one-hot label and q is the guessed label for unlabeled data. $H(p,q)$ is the cross-entropy between distributions p and q. $P(y|x;\theta)$ and $P(y|u;\theta)$ are the approximating distribution getting from our supervised component MMD. λ_U is a hyperparameter.

Experiments

Dataset

Internet Voice Dataset(IVD). Based on Sogou Voice Assistant [1] (Chinese Siri), provided by Sogou Corporation, we collect 7,534,064 Mandarin utterances recorded from 405,510 users by 2013. The corresponding text, query topic and user's accent information are attached to utterances. First, we randomly select 500,000 utterances as the unlabeled data in our proposed framework. Then, we randomly select 2,946 utterances and manually label them with the same labeling method as in (Zhou et al. 2018b). The emotion distributions of labeled utterances are: *Neutral: 49.3%*, *Happiness: 16.5%*, *Disgust: 11.0%*, *Boredom: 8.7%*, *Anger: 9.8%* and *Sadness: 4.6%*. We utilize the 500,000 unlabeled data and 2,946 labeled data for our semi-supervised learning approach.

Acted Voice Dataset(AVD). To enhance the IVD, we establish an acted Chinese voice data with strong emotion including 2397 utterances. The texts of utterances are originally selected from STC conversation dataset (Zhou et al. 2018a) (Shang, Lu, and Li 2015) constructed from Weibo. Specifically, texts are manually selected according to whether expressing emotion easily and adapting to daily communication. Each utterance has averagely 14 characters. Then we invited 27 volunteers(20 women and 7 men) to read the selected text with reference emotion label. Specifically, volunteers are allowed to modify the text and emotional labels according to their own speaking habits and preferences. Then we invite two well-trained annotators to label the utterances. Only when two annotators and the volunteer who read the utterance have same opinion about the emotion labeling, the utterance and its label will be adopted. Finally, we get 2397 labeled strong emotional expression utterances. The emotion distributions of labeled utterances are: *Neutral: 14.0%*, *Happiness: 23.8%*, *Disgust: 17.4%*, *Anger: 22.6%* and *Sadness: 22.2%*.

IEMOCAP. The IEMOCAP (Busso et al. 2008) is a widely-used English acted speech emotion database. To compare with the state-of-the-art supervised method, we form a four-class emotion classification dataset by merging the excitement category into the happy category. The emotion distributions of labeled utterances are: *Happy: 29.6%*, *Anger: 19.9%*, *Sad: 19.6%* and *Neutral: 30.9%*. There are in total 5531 utterances.

Feature Extraction

Acoustic Feature. OpenSMILE toolkit (Eyben et al. 2010) is used to extract acoustic features for both three datasets. Totally, we obtain 1,582 statistic acoustic features, which are also utilized in the Interspeech 2010 Paralinguistic Challenge (Schuller et al. 2010).

Textual Feature. As for the textual information of Chinese utterances in IVD and AVD. First, we utilize Thulac Tool (Li and Sun 2009) to segment words . Then 300-dimensional word2vec (Mikolov et al. 2013) vectors are trained by 31.2 million word corpora from our IVD. As for the textual information of English utterances in IEMOCAP, we utilize the publicly available 300-dimensional word2vec vectors, which are trained on 100 billion words from Google News(Mikolov et al. 2013). Then, 4200-dimensional utterance-level textual features are extracted according to the statistic functions (max, min, mean, range, std, disp, kurtosis,skewness,iqr1-2/2-3/1-3,quartile1/2/3) over the LLDs.

Social Feature. For social attribute feature in Real-world voice Data, we define 7 query topic types $\{$*Chat*, *Consultation*, *Joke*, *Entertainment*, *Operation*, *Search* and *Other*$\}$ as type features and user query locations as the accent features.

Experimental Setup

Parameter Settings. Softmax function is adopted for prediction layers, while the other layers use eLU activation. The model is optimized by Adam (Kinga and Adam 2015) with a mini-batch size of 256. Each local classifier has two hidden fully connected layers with 128 and 64 neurons. The hyper-parameter d in MMD is 2048.

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

	Method	Neutral	Sadness	Disgust	Anger	Happiness	Boredom	Average
F1-Measure	DNN	0.7038	0.3260	0.2264	0.3893	0.4313	0.1814	0.3764
	SAE	0.6688	0.3407	0.2384	0.4000	0.4280	0.2406	0.3861
	MixMatch	0.6759	0.3451	0.2265	0.4087	0.4393	0.2531	0.3914
	MMMD-w/o-avd	0.6966	0.3521	0.2310	0.4186	0.4453	0.2445	0.3980
	MMMD-w-avd	0.6898	0.3706	0.2388	0.4026	0.4563	0.2719	0.4050
	epoch-wise MMMD	0.6874	0.3976	0.2511	0.4115	0.4618	0.2772	**0.4144**

Table 1: The F1-Measure of inferring emotion in different classification models.

	Method	A(%)		T(%)		A+T(%)	
		UA	WA	UA	WA	UA	WA
RNN	[ICASSP, 2017]	58.8	**63.5**	-	-	-	-
MDNN	[AAAI, 2018]	62.7	61.8	**66.9**	65.8	76.7	75.2
AE-ACNN	[ICASSP, 2019a]	59.54	-	-	-	-	-
CNN-LSTM	[ICASSP, 2019b]	53.23	53.43	59.40	59.63	65.9	64.97
Attention-GMU	[ACL, 2019]	59.76	-	-	-	71.69	-
MMD	Our Method	**63.7**	62.2	66.06	**66.37**	**77.0**	**76.6**

Table 2: The performance on IEMOCAP with different features and comparison with the state of the art. A:acoustic. T:text.

We design two experiments:

Experiment 1. To evaluate the effectiveness of our proposed epoch-wise-MMMD, we compare the performance with some semi-supervised baseline methods:

DNN: Learning a Deep neural network(Ren et al. 2014b) merely in labeled IVD.

SAE: Learning a DNN with labeled and unlabeled IVD pre-trained with Stacked Autoencoder(SAE)(Vincent et al. 2010).

Mixmatch: Learning a DNN with labeled and unlabeled IVD augmented with Mixmatch(Berthelot et al. 2019).

MMMD without acted voice data(MMMD-w/o-avd): Training labeled and unlabeled IVD with our proposed MMMD.

MMMD with acted voice data(MMMD-w-avd): Learning MMMD with AVD, labeled and unlabeled IVD. The data samples are trained without curriculum and in random turn.

Our proposed epoch-wise-MMMD: Learning MMMD with AVD, labeled and unlabeled IVD in epoch-wise strategy.

Experiment 2. To compare our method with other currently state-of-art supervised approaches, we conduct Experiment 2. The works we compared have similar conditions as our work(emotion classes, evaluation metrics, consider current utterance instead of conversations). More importantly, since the IEMOCAP is acted and has all labeled data, we utilize the supervised component MMD of our proposed method in comparing. The comparison methods are as follows:

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

[AAAI, 2018] This paper proposes a multi-path deep neural framework while raw features are trained by groups in local classifiers and high-level features fused into global classifier. All classifiers are trained simultaneously. (Zhou et al. 2018b)

[ICASSP, 2019a] This paper learns integrating representations by an unsupervised autoencoder into an attentive convolutional neural network(ACNN) with multi-view learning emotion classifier to improve the speech emotion recognition accuracy.(Neumann and Vu 2019)

[ICASSP, 2019b] This paper use a LSTM network to detect emotion from acoustic features and a multi-resolution CNN to detect emotion from word sequences.(Cho et al. 2019)

[ACL, 2019] This paper presents a hierarchical multimodal model including modality- and context-based attention mechanisms for multimodal emotion recognition. Meanwhile, it adopts multi-view learning for acoustic-only emotion recognition. (Aguilar et al. 2019)

Performance

Performance on Experiment 1. Table 1 shows the performance of our proposed semi-supervised MMMD framework, epoch-wise training strategy and other baselines. First, to evaluate the effectiveness of our proposed semi-supervised framework MMMD, we conduct the experiments of SAE, MixMatch and MMMD to compare their capacity of exploiting the unlabeled data. The results show that the MixMatch approach which combines hybrid semi-supervised learning methods, outperforms the baseline SAE with 1.37% of F1 relatively. And our proposed MMMD-w/o-avd which combine hybrid semi-supervised learning methods based on MixMatch and supervised component MMD further improves the F1 by 3.08% comparing to SAE relatively. It verifies that our proposed semi-supervised framework MMMD is a more effective way to leverage large-scale IVD in VDAs. Furthermore, to evaluate the effectiveness of our proposed epoch-wise training strategy, we

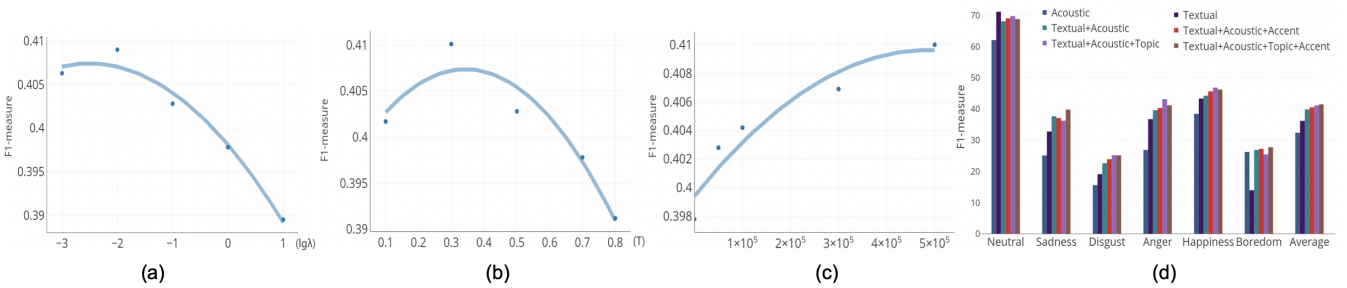


Figure 5: (a)Effects of *Lamda*. (b)Effects of *T*. (c) Performance with different amount of unlabeled data. (d)Feature contribution analysis.

conduct the experiments among MMMD-w/o-avd, MMMD-w-avd and epoch-wise MMMD. The MMMD-w-avd which leverage AVD to augment learning outperforms +1.76% the MMMD-w/o-avd which only utilize IVD relatively. The epoch-wise-MMMD with a epoch-wise learning strategy to leverage AVD further improves the F1 by 4.12% relatively. It verifies that our proposed epoch-wise-MMMD is a more effective way to leverage AVD with strong emotion.

Performance on Experiment 2. Table 2 shows the unweighted accuracy (UA) and weighted accuracy (WA) of competitive methods and the proposed supervised component MMD. While comparing the performance 'feature A+T', our proposed method outperforms all the state-of-the-art baseline methods. Especially, for the UA of the 'feature A+T', +11.1% compared with [ICASSP, 2019b] using CNN-LSTM and +5.3% compared with [ACL, 2019] using Attention-GMU. As for UA of 'feature A only', it shows that our propose MMD (63.7%) is +4.9% compared with [ICASSP, 2017] with RNN, +4.16% compared with [ICASSP, 2019a] using AE-ACNN. These strongly demonstrates the effectiveness of the supervised part of our proposed method.

Analysis

Parameter Analysis. 1)*Loss coefficient* λ. λ is the weight of semi-supervised loss in our model. Figure 5(a) shows the relation between F1-score performance and the magnitude of λ under logarithmic scale. The semi-supervised function starts to take effect when $lg(\lambda)$ increase from $-\infty$ to -2 (λ from 0 to 0.01), but as it increases more, the semi-supervised loss will disturb the main loss function and trigger a worse performance. 2)*Temperature* T. T controls the degree of the entropy of the predictions. Lower T means lower entropy, and encourages the model to give a more precise answer. As the results shown in Figure5(b), the F1 score increases when T lows down from 1 to 0.3. However as T is closer to 0, the restrain will limit model's ability to prediction.

Data Scalability Analysis. To verify the effectiveness of the scale of unlabeled data that our model benefits from, we inspect different scale of unlabeled data from 0 to 500,000. We adopt the epoch-wise MMMD to calculate the performance. In Figure 5(c), as the scale of unlabeled data increases, the model gradually reaches higher performance, ensuring the strong capacity of our proposed model to take advantage of the large-scale data.

Feature Contribution Analysis. We discuss the contributions of different modality features. The F1-measure results for 6 emotion categories and their average are shown in Figure 5(d). Specifically, for all these we adopt the epoch-wise MMMD to calculate the performance in the IVD. The performance of 'Textual Only' outperforms 'Acoustic Only' by 3.81%, which indicates that the textual information can contribute more to the emotion recognition in VADs. Moreover, 'Textual + Acoustic' which contains both Textual information and acoustic information performs best than single modality, indicating the necessity to consider both acoustic and textual information into account. When we take social attribute 'Accent' and 'Topic' into consideration, the average f1-score improve by 0.658% and 1.2% correspondingly, which reveal the potential of social and environmental attributes in assisting the emotion recognition in IVD. Moreover, 'Textual + Acoustic + Accent + Topic' which contains both acoustic, Textual and social information performs best, demonstrates that considering multi-modalities simultaneously can be more effective to infer emotional utterances.

Conclusion

In this paper, we propose a novel semi-supervised multimodal curriculum augmentation deep learning framework to infer emotion for large-scale Internet voice data. To effectively utilize the strong and clear emotion from acted corpus to enhance internet voice data, we design a curriculum learning based epoch-wise training strategy, which trains our model guided by strong and balanced emotion samples from acted voice data and sub-sequently leverages weak and unbalanced emotion samples from internet voice data. Then to take advantage of the large-scale unlabeled data of Real-world dataset, we introduce a Multi-path Mix-match Multimodal Deep Neural Network(MMMD), which effectively trains labeled and unlabeled data in hybrid semi-supervised methods for superior generalisation and robustness. Our approach turns out to be effective in real-world speech emotion inferring, which can provide more intelligent response in real-world VDA applications.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, the state

key program of the National Natural Science Foundation of China (NSFC) (No.61831022) and Tiangong Institute for Intelligent Computing, Tsinghua University. We would like to sincerely thank Huawei Technologies Co., Ltd. for their support and guidance to the work.

References

Aguilar, G.; Rozgic, V.; Wang, W.; and Wang, C. 2019. Multimodal and Multi-view Models for Emotion Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 991–1002.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 5050–5060.

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. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2746–2750. IEEE.

Cho, J.; Pappagari, R.; Kulkarni, P.; Villalba, J.; Carmiel, Y.; and Dehak, N. 2019. Deep neural networks for emotion recognition combining audio and transcripts. *arXiv preprint arXiv:1911.00432* .

Deng, J.; Xu, X.; Zhang, Z.; Frühholz, S.; and Schuller, B. 2017. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(1): 31–43.

Ekman, P.; and Friesen, W. V. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1(1): 49–98.

Eyben, F.; Bock, S.; Schuller, B.; and Graves, A. 2010. Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. In *ISMIR*, 589–594.

Freitag, M.; Amiriparian, S.; Pugachevskiy, S.; Cummins, N.; and Schuller, B. 2017. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research* 18(1): 6340–6344.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* .

Gideon, J.; McInnis, M.; and Provost, E. M. 2019. Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG). *IEEE Transactions on Affective Computing* .

Jia, J.; Zhou, S.; Yin, Y.; Wu, B.; Chen, W.; Meng, F.; and Wang, Y. 2018. Inferring Emotions From Large-Scale Internet Voice Data. *IEEE Transactions on Multimedia* 21(7): 1853–1866.

Kinga, D.; and Adam, J. 2015. A method for stochastic optimization Int. In *Conf. on Learning Representations (ICLR,)*.

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

Lotfian, R.; and Busso, C. 2019. Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(4): 815–826.

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.

Neumann, M.; and Vu, N. T. 2019. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7390–7394. IEEE.

Powers, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* .

Ren, Z.; Jia, J.; Cai, L.; Zhang, K.; and Tang, J. 2014a. Learning to infer public emotions from large-scale networked voice data. In *International Conference on Multimedia Modeling*, 327–339. Springer.

Ren, Z.; Jia, J.; Guo, Q.; Zhang, K.; and Cai, L. 2014b. 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.; 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*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364* .

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.

Zhang, B.; Kong, Y.; Essl, G.; and Provost, E. M. 2019. f-Similarity Preservation Loss for Soft Labels: A Demonstration on Cross-Corpus Speech Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5725–5732.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, S.; Jia, J.; Wang, Q.; Dong, Y.; Yin, Y.; and Lei, K. 2018b. Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*.