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

Suping Zhou1, Jia Jia1,2*, Zhiyong Wu1,2, Zhihan Yang1,2, Yanfeng Wang3, Wei Chen2, Fanbo Meng2, Shuo Huang1, Jialie Shen4, Xiaochuan Wang3

1Department 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
3Sogou Corporation, Beijing, China, 4Queen’s University Belfast, Belfast, U.K
jjia@mail.tsinghua.edu.cn

Abstract

Effective emotion inference from user queries helps to give a more personalized 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 subsequently 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 unstable performance when facing large-scale internet voice data from VDAs (Wu et al. 2016) (Zhou et al. 2018b).

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.
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 subsequently 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 (MMD). Specifically, to effectively learn feature representations for multiple modalities, we propose a Multi-path Multimodal Deep Neural Network (MMD). 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 Assistant1 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 Yu 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

---

1http://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 utilized 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 subsequently 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 $v = \{x^a, x^t, x^e\}$. $x^a$ 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. $x^e$ represents the social attributes features of each utterance, which is a $N_e$ dimensional vector. In addition, $X^a$ is defined as $\{|V| \times N_a\}$ feature matrix with each element $x_{ij}^a$ denoting the $j$th acoustic feature of $v_i$. The definition of $X^t$ and $X^e$ is similar to $X^a$.

The study involves two emotion datasets $V_l$ and $V_u$, corresponding to two different recording environments. $V_l$ refers to acted dataset with strong and clear emotion expressions. $V_l$ refers to real world voice dataset with large amount of speakers. Specifically, we divide $V_l$ into two sets $V_l^L$ (labeled data) and $V_l^U$ (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 $V_l$ and $V_u$, we aim to infer the emotion for every utterance $v \in V_l$: $f : (V_l^L, V_l^U, V_u) \Rightarrow E_S$. (1)

![Figure 3: The structure of MMD and MMMD.](image)

**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 (MCP) (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 multiplying, 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,

\[
\Phi(x_1 \otimes x_2, h, s) = \Phi(x_1, h, s) \ast \Phi(x_2, h, s) \tag{2}
\]

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

\[
y[h[i]] = y[h[i]] + s[i] \cdot x[i] \tag{3}
\]

where \(i\) denotes the \(i\)th feature of a vector. \(s \in \{-1,1\}^n\) and \(h \in \{1,\ldots,d\}^n\) are two randomly initialized vectors. Additionally, according to the convolution theorem states, for \(y_1 \in \mathbb{R}^{d_1}\) and \(y_2 \in \mathbb{R}^{d_2}\), \(y_1 \odot y_2\) can be rewritten as \(\text{FFT}^{-1}(\text{FFT}(y_1) \ast (\text{FFT}(y_2)))\), where \(\ast\) 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 classifiers generate groups of high-level acoustic hidden features \(h_{a1}, h_{a2}, \ldots, h_{a l_1}\) and concatenated textual and social features \(h_{tc1}, h_{tc2}, \ldots, h_{tc l_2}\), 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 \(h_{tc}\).

\[
h_a = \text{Concatenate}([h_{a1}, h_{a2}, \ldots, h_{al_1}]) \tag{4}
\]

\[
h_{tc} = \text{Concatenate}([h_{tc1}, h_{tc2}, \ldots, h_{tc l_2}]) \tag{5}
\]

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

\[
h = \text{MCB}(h_a, h_{tc}) \tag{6}
\]

Then, this \(h\) is applied to train a global classifier. Moreover, the local classifiers and global classifier 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 acoustic and textual features and generate the \(K\) augmentations from unlabeled data \(V_i^U\). We produce a guessed label \(q\) by

\[
q = \frac{1}{K} P(y|V_i^U; \theta) \tag{7}
\]

\(P(y|V_i^U; \theta)\) refer to a model which produces a distribution over class labels \(y\) for input \(V_i^U\) with parameters \(\theta\). 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.

\[
\text{Sharpen}(p, T) = p^T / \sum_{j=1}^{L} p_j^T \tag{8}
\]

where \(T\) is a hyper-parameter related to temperature to control the entropy of the label distribution and \(p\) is the ground truth label distribution. \(L\) is the number of label 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 \(\lambda\) from the Beta distribution and fuse the labeled and unlabeled data by the weight of \(\lambda\).

\[
\lambda \sim \text{Beta}(0.75,0.75) \tag{9}
\]

\[
\lambda' = \max(\lambda, 1-\lambda) \tag{10}
\]

\[
x' = \lambda' x_1 + (1-\lambda') x_2 \tag{11}
\]

\[
p' = \lambda' p_1 + (1-\lambda') p_2 \tag{12}
\]

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 \(\lambda\) is close to either 0 or 1, so that the data mixed up mostly resemble \(x_1\) or \(x_2\).

To apply the above SSL methods in our work, we concatenated the data \(X'\) augmented from labeled data \(X = V_i^L \cup V\) and the data \(U'\) augmented from the unlabeled data \(V_i^U\) 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 \(V_i^{U''}\) are generated by Mixup as follows.

\[
X'' = \text{Augment}(X) \tag{13}
\]

\[
V_i^{U''} = \text{Augment}(V_i^U) \tag{14}
\]

\[
W = \text{Shuffle}((\text{Concatenate}(X', V_i^{U'}))) \tag{15}
\]

\[
X'' = \text{MixUp}(X', W_1) \tag{16}
\]

\[
V_i^{U''} = \text{MixUp}(V_i^{U'}, W_2) \tag{17}
\]

\(X''\) and \(V_i^{U''}\) 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 an 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 $V_a$. Then we train the data from the labeled internet voice data $V_{IL}$. Thirdly, we training the model by mixed labeled $X''$ and unlabeled data $V_U''$.

Finally, we induce the combined loss function $L$.

$$L_{V_a} = \frac{1}{|V_a|} \sum_{x, p \in V_a} H(p, P(y|x; \theta))$$  \hspace{1cm} (18)

$$L_{V_{IL}} = \frac{1}{|V_{IL}|} \sum_{x, p \in V_{IL}} H(p, P(y|x; \theta))$$  \hspace{1cm} (19)

$$L_{X} = \frac{1}{|X|} \sum_{x, p \in X''} H(p, P(y|x; \theta))$$  \hspace{1cm} (20)

$$L_{U} = \frac{1}{|V_{U''}|} \sum_{u, q \in V_{U''}} \left| q - P(y|u; \theta) \right|^2$$  \hspace{1cm} (21)

$$L = L_{V_a} + L_{V_{IL}} + L_{X} + \lambda_U L_U$$  \hspace{1cm} (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. $\lambda_U$ is a hyperparameter.

Experiments

Dataset

**Internet Voice Dataset (IVD).** Based on Sogou Voice Assistant (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 extracted according to the statistic functions (max, min, mean, range, std, disp, kurtosis, skewness, iqr1-2/2-3/1-3, quartile 1/2/3) over the LLDs.

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.

**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 hyperparameter $d$ in MMD is 2048.
Table 1: The F1-Measure of inferring emotion in different classification models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Neutral</th>
<th>Sadness</th>
<th>Disgust</th>
<th>Anger</th>
<th>Happiness</th>
<th>Boredom</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>0.7038</td>
<td>0.3260</td>
<td>0.2264</td>
<td>0.3893</td>
<td>0.4313</td>
<td>0.1814</td>
<td>0.3764</td>
</tr>
<tr>
<td>SAE</td>
<td>0.6688</td>
<td>0.3407</td>
<td>0.2384</td>
<td>0.4000</td>
<td>0.4280</td>
<td>0.2406</td>
<td>0.3861</td>
</tr>
<tr>
<td>MixMatch</td>
<td>0.6759</td>
<td>0.3451</td>
<td>0.2265</td>
<td>0.4186</td>
<td>0.4453</td>
<td>0.2445</td>
<td>0.3980</td>
</tr>
<tr>
<td>MMMD-w/o-avd</td>
<td>0.6966</td>
<td>0.3521</td>
<td>0.2310</td>
<td>0.4026</td>
<td>0.4563</td>
<td>0.2719</td>
<td>0.4050</td>
</tr>
<tr>
<td>epoch-wise MMMD</td>
<td>0.6874</td>
<td>0.3976</td>
<td>0.2511</td>
<td>0.4115</td>
<td>0.4618</td>
<td>0.2772</td>
<td>0.4144</td>
</tr>
</tbody>
</table>

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 Lambda. (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 $\lambda$. $\lambda$ is the weight of semi-supervised loss in our model. Figure 5(a) shows the relation between F1-score performance and the magnitude of $\lambda$ under logarithmic scale. The semi-supervised function starts to take effect when $\log(\lambda)$ increase from $-\infty$ to -2 ($\lambda$ 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 Figure 5(b), the F1 score increases when $T$ lowers 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


Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P. 2010. Stacked denoising autoencoders: Learning


