Speaker-Invariant Emotion Recognition with Adversarial Learning

Bryan Leow Xuan Zhen, L. L. Chamara Kasun, Ahn Chung Soo, Jagath C. Rajapakse

School of Computer Science and Engineering, Nanyang Technological University, Singapore

{xleow002, chamarakasun, csahn, asjagath}@ntu.edu.sg

Abstract

Performances of current methods for speech emotion recognition are dependent on whether the speakers are present in the training dataset. In this paper, we propose adversarial learning (AL) network for speaker-invariant emotion recognition (SIER) task. SIER is achieved in a network consisting of an encoder, an emotion classifier, and a speaker classifier, and implementing an adversarial learning strategy to learn representations that are invariant for speaker characteristics. We argue that the representation realized by AL network are independent of the speakers and therefore are conducive for SIER. Earlier, adversarial learning has been used for SIER and we present an improved encoder that demonstrated state-of-the-art performances for SIER on Emo-DB and RAVDESS datasets.

Index Terms: convolutional neural networks, domain adversarial training, recurrent neural networks, speaker independent emotion recognition.

1. Introduction

Speech emotion recognition (SER) aims at recognition of speakers' emotion from speech signals. SER algorithms typically begin with the extraction of features from speech and are then followed by classification of speech signals into emotions. Features extracted include high-level features such as those specificed by Interspeech Para-linguistics Challenge 2010 [1] and low-level features such as spectrogram features. Recently, deep learning architectures have become popular for SER and shown that they can directly learn from low-level features without resorting to time-consuming processes of creating high-level features. Direct use of Mel-scale spectrograms have shown state-of-the-art accuracy in speech recognition task with deep learning [2].

Performances of typical SER methods are dependent on whether the speaker is present or absent in the training set. Ideally, SER algorithms are to perform irrespective of whether speaker is new to the algorithm. In this paper, we focus on Speaker-Invariant Emotion Recognition (SIER) where the training and testing data come from different speakers. The aim of SIER is to create a model for speech data spoken from a set of speakers and apply the model to predict emotions from speech from a different set of speakers. SIER is more challenging than SER as SIER algorithms must learn a speaker invariant representation where speakers' identity is removed.

Typically for SER tasks, the training data is collected on an environment (source domain) different from the environment (target domain) where the predictions are performed. If source and target domain distributions are different, SER algorithms perform poorly during testing. Ideally, SER methods need to learn representations that are common to either domain. In order to address the issue of variability between training and testing data, adversarial learning approaches, inspired by Generative Adversarial Network (GAN) [3], have been proposed for SER. One approach is to use GAN to generate synthetic data to augment the training set [4, 5, 6]. The models trained with the augmented training data have shown to perform better than models trained with only original data. Alternatively, domain adversarial training can be used to resolve the disparity between training and testing environments or domains [7]. Domain adversarial networks using AL was trained with source data and to predict on target data which is from another source different from the source data [8, 9]. Adversarial has also been investigated in removing speaker variability in speech by learning speaker invariant representations [10, 11].

In this paper, we propose an adversarial learning (AL) network that learns representations independent of speaker characteristics for SIER task. The proposed AL network consists of an encoder that learns speaker-invariant representations, an emotion classifier to predict emotion labels, and a speaker classifier that helps remove speaker variability. Our work extends the work initiated by [10, 11] by proposing a novel decoder to learn speaker-invariant representations. By minimizing the accuracy of speaker classifier while maximizing the accuracy of emotion prediction, the proposed AL network is agnostic to speaker characteristics. Our encoder uses 2-dimensional (2D) Convolutional Neural Network (CNN) layers followed by bidirectional Gated Recurrent Units (biGRU). We also perform experiments to determine the required depth of the proposed 2D CNN biGRU encoder.

The popularity deep learning approaches to SER can be attributed to their usage of the Mel-scale to compute features such as MFCC where equal distance in pitch sound equally distant to the listener. This is important because humans do not perceive frequency on a linear scale and can distinguish low frequencies at a better resolution than high frequencies, i.e., a human can tell apart a 500Hz and 1000Hz signal but cannot tell the difference between a 1Khz and 1.5kHz signal. Hence, MFCC aim to capture this non-linear perception of sound with the frequencies spaced according to perception. MFCC represents speech as a set of frequencies that vary along with time. The number of frequencies captured can be determined by the user. However, previous adversarial learning for SIER used a 1-D neural architecture that had learned short-term temporal features capturing only statistical features such as energies of frequency components [11]. In order to incorporate multiple frequency components for SIER, we utilize 2D sprectral features given by MFCC coefficients gathered over a time interval.

Deep learning architectures consisting of Deep Neural Networks (DNN), CNN, and Recurrent Neural Networks (RNN) have been used for SER. A typical speech encoder consists of several DNN or CNN layers followed by RNN layers made up of Long Short-Term Memory (LSTM) units or GRUs. Inspired by the success of 2D CNN LSTM for SER [12], we propose 2D CNN biGRU architecture as the encoder for AL network. Further, we empirically determine the required depth for SIER. The CNN captures local contextual information from spectral features and biGRU captures the temporal dependencies. Previously, 1D CNN [11] and Time-delayed Neural Networks (TDNN) [10] have been used for SIER, which did not capture the necessary contextual information from speech. We demonstrate that 2D CNN biGRU encoder achieves state-of-the-art performances on two benchmark datasets for SIER.

2. Related work

SIER tasks typically have a limited number of training data due to the difficulty of collecting and labeling speech emotion data. Hence, GAN [3] approaches have been introduced to SIER tasks to generate data for augmentation. Traditional GAN architectures use a random Gaussian vector as a seed to generate data, assuming a Gaussian generative model for speech. However, due to the high variability of speech emotion data, speech emotion data cannot be modelled as coming from a simple Gaussian distribution [5]. Hence, a linearly mixed speech emotion data are used to generate data in GAN for SER. Another approach of generating speech emotion data is to use an autoencoder to generate latent vector [6]. And use this latent vector added with random Gaussian noise as a seed to a GAN to generate speech emotion data. This approach has been shown of capable of generating speech emotion belonging to a specific category such as anger or happy.

Adversarial learning has also been attempted to resolve the issue of variability of data for SER [7]. Speech emotion data can be collected from different domains such as with German and English speakers. Machine learning algorithms trained on German speech emotion data have difficulty in predicting emotion of English speech emotion data. Domain adversarial training has been applied to remove the variation of data, caused due to the variability across different domains [8]. This is achieved by using a encoder, emotion classifier and domain classifier with gradient reversal. Domain classifier aims to remove the variability of the domain. This same concept of learning a domain invariant representation has been extended to learning a speaker invariant representation by changing the domain classifier to speaker classifier with gradient reversal [10, 11].

3. Methodology

The proposed AL network architecture is illustrated in figure 1, which consists of three components: (i) an encoder made out of 2D CNN and biGRU layers, which generates speaker invariant representations; (ii) an emotion classifier that predicts emotional labels; and (iii) a speaker classifier with gradient reversal, which removes speaker variability.

3.1. Encoder

The network takes MFCC features of speech in a time window size of T . Input speech emotion data $x = (x(t))$ where $x(t)$ denotes the features extracted at time t, $x \in \mathbb{R}^{T\times n}$ and n denotes the number of features. The encoder consist of a one or more 2D CNN layers followed by one biGRU layer. The 2D CNN aims to learn short-term frequency features as well as temporal information as speech contain multiple frequencies. 2D CNN consist of (i) a 2D convolutional layer; (ii) a batch normalization layer; and (iii) a max pooling layer.

Let w_k^1 denote the filter weight connected to k the feature map at the convolutional layer. The output h_k^1 of the k the feature is given by

$$h_k^1 = x \circledast w_k^1 \tag{1}$$

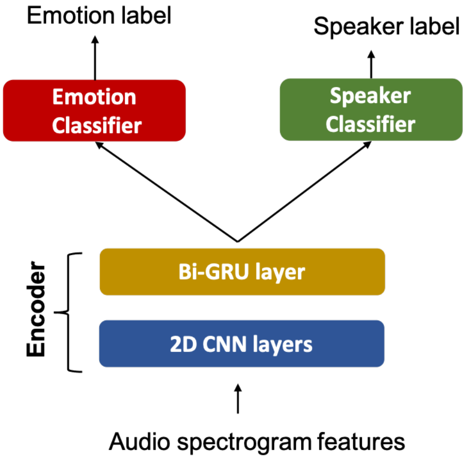

Figure 1: *Proposed adversarial learning (AL) network architecture. The encoder consists of 2D CNN layers and a bidirectional GRU (biGRU) layer. Emotion and speaker classifier consist of two hidden layer fully connected neural network layers followed by a softmax layer.*

where $\circledast$ denote the convolution operation. The convolution layer output is processed by a batch-normalization layer, which is processed by dropout and exponential linear units (eLU). The batch-normalization layer output h_k^2 for k feature map is given by

$$h_k^2 = \texttt{eLU}\left(\texttt{dropout}\left(w_k^2 \cdot \texttt{BN}(h_k^1) + b_k^2\right)\right) \tag{2}$$

where $\cdot$ denotes the element-wise multiplication, BN denotes batch-normalization function that normalizes the data by subtracting the mean and dividing by the standard deviation over a batch. w_k^2 and b_k^2 denote learnable parameters. dropouts and eLU denote dropout operation and eLU activation function. Max-pooling layer output h_k^2 is given by

$$h_k^3 = \texttt{pool}\left(h_k^2\right) \tag{3}$$

Let $h_k^3 \in \mathbb{R}^{T\times n^3}$ where n_3 is the 2D CNN output features.

The Bi-directional Gated Recurrent Unit (biGRU) learns the long-term temporal relationship of the short-term frequency and temporal features and in this work we used a biGRU as the recurrent neural network. To achieve this, the convolution layer output $h^3 = \left(h_k^3\right) \in \mathbb{R}^{K\times T_c\times n^3}$ is reshaped as $h^3(t) \in \mathbb{R}^{K\times n^3}$. The RNN output $h^4(t)$ is given by:

$$h^4(t) = \texttt{biGRU}(h^3(t), h^4(t-1)) \tag{4}$$

where $h^4(t) \in \mathbb{R}^{n^4}$ and n^4 is the number of hidden neurons in bi-GRU. The output of the stats pooling layer $h^5 \in \mathbb{R}^{2*n^4}$ is given by:

$$h^5 = \texttt{stats\_pool}(h^4) \tag{5}$$

where stats_pool function calculates the mean and standard deviation of h^4 along the time dimension and concatenates along the feature dimension.

3.2. Emotion and Speaker Classifiers

Emotion and speaker classifiers consist of two fully connected layers and each layer performs batch-normalization and dropout, and is processed with ReLU activation function. Output layer is a softmax layer.

Emotion classifier predicts the emotion labels y_e from the stats pooling layer output h^5 as:

$$\begin{aligned} h^6 &= \texttt{ReLU}\left(\texttt{dropout}\left(W^6\texttt{BN}\left(V^6h^5 + c^6\right) + b^6\right)\right) \\ h^7 &= \texttt{ReLU}\left(\texttt{dropout}\left(W^7\texttt{BN}\left(V^7h^6 + c^7\right) + b^7\right)\right) \\ y_e &= \texttt{softmax}\left(W^8h^7 + b^8\right) \end{aligned} \tag{6}$$

Speaker classifier predicts speaker labels y_s from the stats pooling layer output h^5 as:

$$\begin{aligned} h^9 &= \texttt{relu}\left(\texttt{dropout}\left(W^9\texttt{BN}\left(V^9h^5 + c^9\right) + b^9\right)\right) \\ h^{10} &= \texttt{relu}\left(\texttt{dropout}\left(W^{10}\texttt{BN}\left(V^{10}h^9 + c^{10}\right) + b^{10}\right)\right) \\ y_s &= \texttt{softmax}\left(W^{11}h^{10} + b^{11}\right) \end{aligned} \tag{7}$$

3.3. Adversarial Learning

We minimize the cross entropy loss of the emotion classifier as:

$$J_e = -E_x[d_e log(y_e)] \tag{8}$$

where d_e is the emotion labels and E_x is the expectation over data x. The cross-entropy loss J_s of speaker classifier is given by

$$J_s = -E_x[d_s log(y_s)] \tag{9}$$

where d_s are the speaker identity labels.

During training, the parameters of the underlying deep feature mapping are optimised to minimise the loss of the emotion classifier J_e and to maximise the loss of the speaker classifier J_s. The latter is enabled with the use of the gradient reversal that leaves the features unchanged during forward propagation and reverses the gradient by multiplying it by a negative scalar λ during the backpropagation, thus working in an adversarial manner to the domain classifier and encourages the emergence of speaker-invariant features.

The overall loss that is minimized during learning is given by

$$J = J_e - \lambda J_s \tag{10}$$

where $\lambda = \frac{2}{1+exp(-\gamma p)} - 1$, γ is a positive annealing hyper parameter, and p is the percentage of training. Importance factor of the speaker classifier λ gradually increases from 0 to at most 1 with the training progress since the negative gradient can hamper the initial weight updates [7]. The rate and degree of speaker adversarial training occurs can be adjusted by changing γ, where a higher γ value will result in a higher rate and degree of speaker adversarial training.

4. Experiments and Results

In this section, we investigate the efficacy of the AL network encoder on two benchmark datasets for SIER: Emo-DB [13] and RAVDESS datasets [14]. The performances of the our network is compared with two existing methods that uses AL for SIER: (i)1D TDNN biLSTM [10] ; (ii) 1D CNN GRU [11].

All the experiments were carried out on a DGX server with two Xeon E5-2698 v4 clocked at 2.2 GHz, 512 GB RAM and eight Nvidia V100 32 GB graphic cards running Ubuntu 16.04. The scripts were written in Python using Pytorch package.

4.1. Datasets and feature extraction

Emo-DB dataset contain speech recording of 7 emotions representing anger, boredom, disgust, fear, happy, sad and neutral, from 10 German speaking persons. There were total of 535 speech samples. For SIER, we performed 5-fold leave-two-speakers-out for cross-validation and testing. That is, in each fold, 6 speakers were selected for training, 2 speakers for validation and 2 speakers for testing. The ratio of male to female speakers in training, validation, and testing was set to 1:1.

Table 1: *Leave-two-speaker-out cross-validation accuracy (%) of adversarial learning network with 2D CNN biGRU encoder for speaker-invariant emotion recognition on Emo-DB and RAVDESS datasets at different number of CNN layers and annealing parameter γ values.*

No. of CNN layers	γ	Emo-DB	RAVDESS
1	without AL	67.3±6.5	56.3±8.1
	1.25	72.1±8.4	60.6±10.3
	2.50	70.6±5.9	**61.2±8.4**
	3.33	**73.6±7.1**	60.6±10.6
2	without AL	68.3±9.6	56.3±9.8
	1.25	68.0±8.0	59.4±7.6
	2.50	71.0±7.6	60.9±8.5
	3.33	67.6±7.4	60.5±9.9
3	without AL	65.7±10.6	55.8±8.8
	1.25	65.1±7.2	53.7±6.7
	2.50	66.0±5.5	59.4±7.6
	3.33	66.6±5.6	58.9±8.5
4	without AL	59.6±5.3	54.0±11.3
	1.25	66.0±6.9	54.2±8.7
	2.50	62.9±9.6	54.1±9.3
	3.33	65.2±7.2	51.3±11.1

RAVDESS dataset contains speech recordings of 8 emotions representing anger, calm, surprised, fear, happy, sad, neutral and disgust, from 24 professional actors with north American English accent. There were total of 1440 speech samples. We performed 12-fold leave-two-speakers-out cross validation and testing by selecting 20 speakers for training, 2 for validation and 2 for testing. The ratio of male to female speakers in training, validation and testing was set to 1:1.

Speech recordings in both datasets were trimmed and filtered to remove silence and background noise. The recordings in each dataset were then padded with zeros to ensure that they are of the same length as the longest audio recording in the dataset. The pre-processed recordings were then converted to MFCC features with the following settings: frame size of 2048 frames, hop size of 512 frames, and Hann Window of 2048 frames as a windowing function to minimise spectral leakage. We only retained the first 20 MFCC coefficients.

4.2. Parameter initialization

Hyper-parameter selection for the neural network architecture was performed using the grid search method and the parameters that yielded the best accuracy were selected. We chose γ from $[1.25, 2.5, 3.33]$ and the number of layers from 1 to 4 in 2D CNN with biGRU encoder. We used convolutional filter size from $[5 \times 5, 2 \times 2]$, number of convolutional filters $[128, 64]$, number of hidden neurons in GRU $[256, 128]$ and the number of hidden neurons in the fully connected layer $[128, 64]$. We set the convolutional stride two 2 and the max pooling size 2×2 and stride pooling stride of 2.

Table 2: *Comparison of leave-two-speaker-out cross-validation accuracy (%) of different encoders for AL for speaker-invariant emotion recognition.*

Input features	Encoder	Emo-DB	RAVDESS
MFCC	TDNN biLSTM [10]	61.4±8.7	52.4±10.9
LogMFB, energy, pitch	1D CNN GRU [11]	44.2±4.7	30.5±4.6
MFCC	2D CNN biGRU	**73.6**±**7.1**	**61.2**±**8.4**

4.3. Adversarial learning with 2D CNN biGRU encoder

Table 1 shows the result of our 2D CNN biGRU encoder with different number of CNN layers. For each depth, we compare the performance of our 2D CNN biGRU with different value of γ, where a higher γ value resulted in a higher rate and degree of speaker adversarial training. Comparison were also made to our 2D CNN biGRU encoder without AL since the 2D CNN biGRU encoder without AL did not have a speaker classifier connected to it.

As seen Table 1, it is quite apparent that having lesser number of CNN in our 2D CNN biGRU encoders yield better result. Our encoder trained with AL significantly outperformed training without AL. However, the depth of the encoder was also critical. We note that the benefits of training our 2D CNN biGRU encoders with AL diminish as the number of CNN layers increase. This is especially true for the case with the RAVDESS dataset. When the number of 2D CNN layers is 4, the 2D CNN biGRU trained with AL does not seem offer much additional performance compared to without AL for RAVDESS. We found that one CNN layer depth is optimal for both datasets for AL for SIER.

4.4. Comparison with existing AL architectures

To understand how our 2D CNN biGRU encoder perform relative to other encoders utilising AL in SIER, comparisons to recent methods using AL [10, 11] were performed. To ensure fairness of comparison, we performed the same 5-fold leave-two-speakers-out cross-validation and testing and 12-fold-leave-two-speakers-out cross-validation and testing on Emo-DB and RAVDESS dataset, respectively, on the respective encoders. Table 2 shows that our proposed architecture achieves the best performance compared to existing encoders in both RAVDESS and Emo-DB dataset, confirming the superiority of encoders utilising 2D convolution filters for SIER tasks.

5. Conclusions

We presented an AL network to learn speaker-invariant representations for SER. The network learns representation maximizing emotion classification and minimizing the sensitivity for speaker characteristics. We demonstrated that proposed 2D CNN biGRU encoder outperformed existing 1D TDNN biLSTM and 1D CNN GRU and achieved state-of-the-art accuracy for SIER. Our AL network is useful when SER is to be trained on limited amount of speakers and be tested on unseen speakers.

6. Acknowledgements

We would like to thank Singtel Cognitive and Artificial Intelligence Lab for Enterprises Nanyang technological University Singapore for providing access to the GPU server.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proceedings of INTERSPEECH*, sep 2010.

[2] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep Speech - Scaling up end-to-end speech recognition." *CoRR abs/1602.00985*, vol. cs.CL, 2014.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[4] S. Sahu, R. Gupta, G. Sivaraman, C. Espy-Wilson, and W. AbdAlmageed, "Adversarial Auto-Encoders For Speech Based Emotion Recognition," in *Proceedings of INTERSPEECH*, 2017.

[5] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting Generative Adversarial Networks for Speech Emotion Recognition," in *Proceedings INTERSPEECH*, 2020, pp. 521–525.

[6] S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, "GAN-Based Data Generation for Speech Emotion Recognition," in *Proceedings INTERSPEECH*, 2020, pp. 3446–3450.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 1, p. 2096–2030, Jan 2016.

[8] M. Abdelwahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, 04 2018.

[9] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition," in *Proceedings of INTERSPEECH*, 2019, pp. 1656–1660.

[10] M. Tu, Y. Tang, J. Huang, X. He, and B. Zhou, "Towards adversarial learning of speaker-invariant representation for speech emotion recognition," 2019.

[11] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-Invariant Affective Representation Learning via Adversarial Training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7144–7148.

[12] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312 – 323, 2019.

[13] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of INTERSPEECH*, 2005, pp. 1517–1520.

[14] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.