INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition

Zheng Lian[1,3], Jianhua Tao[1,2,3], Bin Liu[1], Jian Huang[1,3], Zhanlei Yang[4], Rongjun Li[4]

[1]National Laboratory of Pattern Recognition, CASIA, Beijing
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing
[4]Huawei Technologies Co., LTD., Beijing

{zheng.lian, jhtao, liubin, jian.huang}@nlpr.ia.ac.cn, {yangzhanlei1, lirongjun3}@huawei.com

Abstract

Emotion recognition remains a complex task due to speaker variations and low-resource training samples. To address these difficulties, we focus on the domain adversarial neural networks (DANN) for emotion recognition. The primary task is to predict emotion labels. The secondary task is to learn a common representation where speaker identities can not be distinguished. By using this approach, we bring the representations of different speakers closer. Meanwhile, through using the unlabeled data in the training process, we alleviate the impact of low-resource training samples. In the meantime, prior work found that contextual information and multimodal features are important for emotion recognition. However, previous DANN based approaches ignore these information, thus limiting their performance. In this paper, we propose the context-dependent domain adversarial neural network for multimodal emotion recognition. To verify the effectiveness of our proposed method, we conduct experiments on the benchmark dataset IEMOCAP. Experimental results demonstrate that the proposed method shows an absolute improvement of 3.48% over state-of-the-art strategies.

Index Terms: emotion recognition, domain adversarial learning, speaker-independent representations, contextual information, multimodal features

1. Introduction

Emotion recognition is an important research topic for interactive intelligence systems with broad applications in many tasks, such as customer service [1, 2], social media analysis [3, 4, 5] and education [6, 7]. The task of emotion recognition requires understanding the way that humans express their emotions, and classifies each utterance into one of a fixed set of categories.

Despite its importance, emotion recognition remains a complex task due to the following challenges: (1) Since existing datasets are relatively small-scale [8, 9], the first challenge is how to learn a good representation that captures the emotion signals with limited training samples; (2) Since emotion recognition systems are greatly affected by speaker variations [10], the second challenge is how to learn robust emotion representations across different speakers; (3) Since multimodal features and context information are vitally important for emotion recognition [11, 12], the third challenge is how to effectively utilize these information in emotion recognition.

The key challenge in emotion recognition is how to learn good emotion representations with limited training data. The publicly available datasets [8, 9] have relatively small number of total utterances. To deal with this problem, previous works [13, 14] used unsupervised learning to convert original features into more compressed representations, thus capturing intrinsic structures of the data. One common method is to train autoencoder [13] and its variations (such as adversarial autoencoders (AAEs) [14] and variational autoencoders (VAEs) [15]). However, it is unclear whether these compressed representations preserve emotion components. In fact, prior work found that emotion components could be lost after feature compression [16].

Another key challenge is how to learn robust representations that remain invariant under different speakers. Previous works focused on data split strategies, thus ensuring no speaker overlap in the training set and the testing set [11, 17]. For example, the IEMOCAP dataset [9] contains five sessions and each session has different actors. Hazarika et al. [17] utilized utterances from the first four sessions for training and others for testing. However, it is unclear whether these methods can actually learn speaker-independent representations.

To address above difficulties, domain adversarial neural network (DANN) based methods [18, 19] have been proposed recently, and achieved promising results for emotion recognition. These methods contain three key modules: the feature encoder, the emotion category classifier and the speaker identity classifier. Firstly, a gradient reversal layer [20] is inserted between the feature encoder and the speaker classifier. Through optimizing speaker and emotion classifiers, these methods [18, 19] are able to learn representations that preserve emotion components and remain invariant under different speakers. Secondly, these methods [18, 19] can utilize the unlabeled data in the training process, thus alleviating the impact of limited training samples.

The third challenge is how to effectively use multimodal features and context information in emotion recognition [11, 12]. Firstly, human perceive emotions not only through the current utterance, but also from the contextual information in its surroundings [21]. Secondly, due to the complexity of emotion recognition, the single modality is difficult to meet the demand, and multimodal features should be considered. However, previous DANN based approaches [18, 19] ignore these information, thus limiting their performance in emotion recognition.

In this paper, we propose the context-dependent DANN for multimodal emotion recognition. Different from previous works for low-resource problems [13, 14] , our method is able to learn representations that preserve emotion components and remain invariant under different speakers, thus improving recognition performance in low-resource conditions. Different from previous DANN approaches [18, 19], our method utilizes multimodal features and context information for emotion recogni-

http://dx.doi.org/10.21437/Interspeech.2020-1705

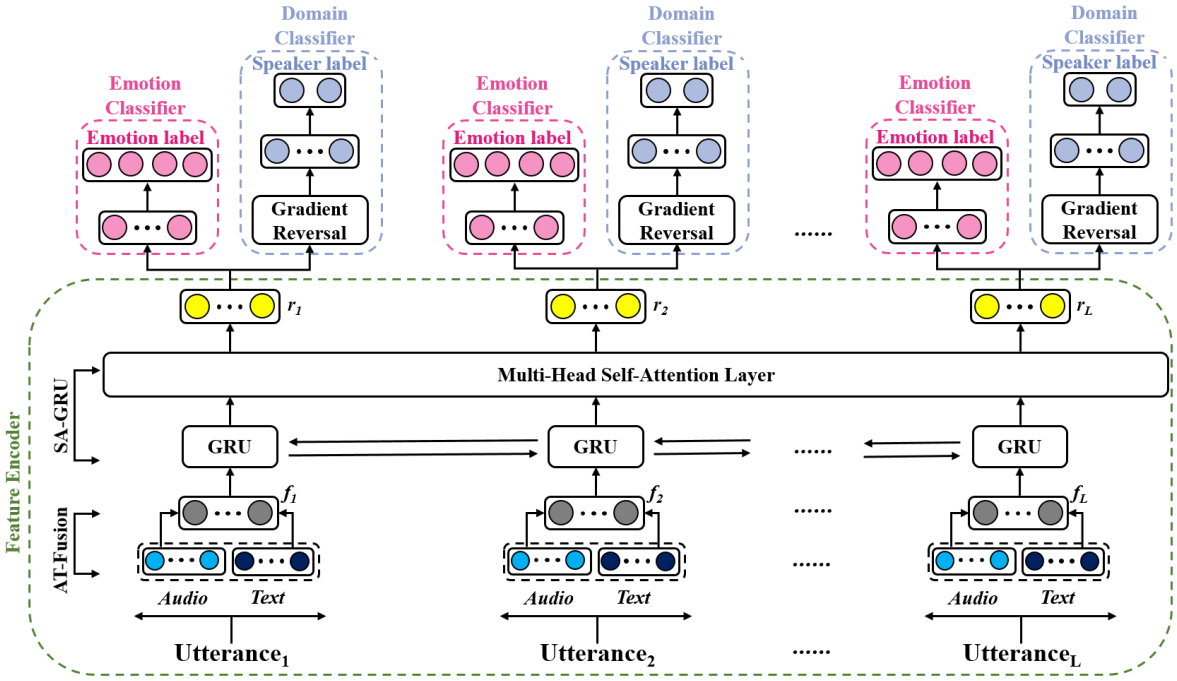

Figure 1: *Overall structure of the proposed framework.*

tion. The main contributions of this paper lie in three aspects: 1) We propose a novel DANN based framework for emotion recognition; (2) We observe that our method achieves promising performance with limited training data; (3) Experimental results on the popular benchmark datasets IEMOCAP demonstrate the effectiveness of our method. Our method shows an absolute improvement of 3.48% over state-of-the-art strategies.

The remainder of this paper is organized as follows: In Section 2, we formalize the problem statement and describe our proposed method in detail. The experimental datasets, setup, results and analysis are illustrated in Section 3. Finally, we give a conclusion of the proposed work in Section 4.

2. Proposed Method

In this paper, we propose a DANN based framework for emotion recognition. As shown in Figure 1, the proposed framework consists of three modules: (1) The feature encoder extracts context-dependent multimodal representations for each utterance; (2) The domain classifier learns to extract speaker-independent representations; (3) The emotion classifier ensures emotion components is preserved in these representations.

2.1. Problem Definition

A dialogue $U = \{(u_1, e_1, s_1), (u_2, e_2, s_2), ..., (u_L, e_L, s_L)\}$ contains L pairs of $(u_i, e_i, s_i) \in (\mathcal{U}, \mathcal{E}, \mathcal{S})$. Here, u_i is the i^{th} utterance in the dialogue, which is produced by the speaker s_i with the emotion e_i. $\mathcal{U}, \mathcal{E}, \mathcal{S}$ are the sets of whole utterances, emotion labels and all speakers, respectively.

Let there be M emotion-labeled dialogues, and each dialogue j contains L_j pairs of $(u_i, e_i, s_i) \in (\mathcal{U}, \mathcal{E}, \mathcal{S})$. Let there be N unlabeled dialogues and each dialogue j contains L_j pairs of $(u_i, s_i) \in (\mathcal{U}, \mathcal{S})$. Like previous experimental settings [17], speaker identities are always available. The task is to predict the emotion label for each utterance in these unlabeled dialogues.

2.2. Context-dependent Multimodal Feature Encoder

The feature encoder contains two key components: the Audio-Text Fusion component (AT-Fusion) for multi-modalities fusion and the Self-Attention based Gated Recurrent Unit (SA-GRU) for contextual feature extraction.

Multi-modalities Fusion (AT-Fusion): Different modalities have different contributions in emotion recognition. To aggregate the salient information over each modality, we utilize the attention mechanism for multi-modalities fusion. Specifically, we first extract acoustic features $a_i \in \mathbb{R}^{d_a \times 1}$ and lexical features $t_i \in \mathbb{R}^{d_t \times 1}$ from each utterance u_i. Here, d_a and d_t represent feature dimensions of acoustic features and lexical features, respectively. Then we equalize the dimensions of these features to size d using two fully-connected layers. AT-Fusion takes these features as inputs, and outputs the attention vector $\alpha_{fuse} \in \mathbb{R}^{1 \times 2}$ over two modalities. Finally, the fusion representation $f_i \in \mathbb{R}^{d \times 1}$ is generated as follows:

$$u_i^{cat} = Concat(W_a a_i, W_t t_i) \tag{1}$$

$$\alpha_{fuse} = softmax(w_F^T tanh(W_F u_i^{cat})) \tag{2}$$

$$f_i = u_i^{cat} \alpha_{fuse}^T \tag{3}$$

where $w_a \in \mathbb{R}^{d \times d_a}$, $w_t \in \mathbb{R}^{d \times d_t}$, $W_F \in \mathbb{R}^{d \times d}$ and $w_F \in \mathbb{R}^{d \times 1}$ are trainable parameters. Here, $u_i^{cat} \in \mathbb{R}^{d \times 2}$.

This multimodal representation is generated for utterances in the conversation U, marked as $F = [f_1, f_2, ..., f_i, ..., f_L]$.

Contextual Feature Extraction (SA-GRU): SA-GRU uses the bi-directional GRU (bi-GRU), in combination with the self-attention mechanism [22] to amplify the important contextual evidents for emotion recognition. Specifically, multimodal representations F are given as inputs to the bi-GRU. Outputs of this layer form $H = [h_1, h_2, ..., h_i, ..., h_L]$, where $H \in \mathbb{R}^{L \times d}$. Then H is fed into the self-attention network. It consists of a multi-head attention to extract the cross-position information.

Each head $head_i \in \mathbb{R}^{L\times(d/h)}, i \in [1,h]$ (h is the number of heads) is generated using the inner product as follows:

$$head_i = softmax((HW_i^Q)(HW_i^K)^T)((HW_i^V) \quad (4)$$

where $W_i^Q \in \mathbb{R}^{d\times(d/h)}$, $W_i^K \in \mathbb{R}^{d\times(d/h)}$ and $W_i^V \in \mathbb{R}^{d\times(d/h)}$ are trainable parameters.

Then outputs of each $head_i \in \mathbb{R}^{L\times(d/h)}, i \in [1,h]$ are concatenated together as final values $R \in \mathbb{R}^{L\times d}$. As contextual representations R is generated for all utterances in the conversation U, it can also be represented as $R = [r_1, r_2, ..., r_i, ..., r_L]$, where $r_i \in \mathbb{R}^d, i \in [1, L]$.

2.3. Domain Adversarial Neural Networks

DANN has two classifiers – the emotion classifier and the domain classifier. Both classifiers share the feature encoder (in Session 2.2) that determines the representations of the data used for classification. This approach introduces a gradient reversal layer [20] between the domain classifier and the feature encoder. This layer passes the data during forward propagation and inverts the sign of the gradient during backward propagation. Therefore, DANN attempts to minimize the emotion classification error and maximize the domain classification error. By considering these two goals, the model ensures a discriminative representation for the emotion recognition, while making the samples from different speakers indistinguishable.

Let there be M emotion-labeled dialogues, and each dialogue i contains L_i utterance. Define u_j is the j^{th} utterance in the dialogue i, which is uttered by the speaker s_j with the emotion e_j. After feature encoder (in Session 2.2), r_j is the context-dependent multimodal feature for u_j. We train the emotion recognition task with emotion-labeled dialogues (in Session 2.1). The performance of emotion classifier is optimized by minimizing the cross entropy loss L_y:

$$L_y = \sum_{i=1}^{M}\sum_{j=1}^{L_i} -logP(e_j|r_j) \quad (5)$$

In addition, we train the domain classifier with M labeled and N unlabeled dialogues (in Session 2.1). The performance of domain classifier is optimized by minimizing the cross entropy loss L_d:

$$L_d = \sum_{i=1}^{M+N}\sum_{j=1}^{L_i} -logP(s_j|r_j) \quad (6)$$

To combine these two objective functions together, we flip the sign of L_d to do a gradient reversal and minimize the weighted overall loss sums. The final objective loss function is written as:

$$L = L_y - \lambda L_d \quad (7)$$

where $\lambda \in [0, 1]$ is a hyper-parameter that controls the trade off between two losses.

3. Experiments and Discussion

3.1. Corpus Description

We perform experiments on the IEMOCAP dataset [9], a benchmark dataset for emotion recognition. It contains audio-visual conversations spanning 12.46 hours of various dialogue scenarios. There are five sessions and two distinct professional actors are grouped in a single session. All the conversations are split into small utterances, which are annotated using the following categories: *anger*, *happiness*, *sadness*, *neutral*, *excitement*, *frustration*, *fear*, *surprise* and *other*. To compare our method with state-of-the-art methods [11, 17], we consider the first four categories, where *happiness* and *excitement* categories are merged into the single *happiness* category. Thus 5531 utterances are involved (*happiness*: 1636, *neutral*: 1084, *anger*: 1103, *sadness*: 1708). The number of utterances and dialogues of each session are listed in Table 1.

Table 1: *The data distribution of the IEMOCAP dataset.*

Session	1	2	3	4	5
No.utterance	1085	1023	1151	1031	1241
No.dialogue	28	30	32	30	31

3.2. Data Representation

Acoustic features: We extract utterance-level acoustic features using the openSMILE [23] toolkit. Specifically, we utilize the Computational Paralinguistic Challenge (ComParE) feature-set introduced by Schuller et al. [24]. Totally, 6373-dimensional features are extracted, including energy, spectral, MFCCs, and their statistics (such as mean, root quadratic mean).

Lexical features: We use word embeddings to represent the lexical information. Specifically, we employ deep contextualized word representations using the language model ELMo [25]. These word vectors are trained on the 1 Billion Word Benchmark [26]. Compared with traditional word vectors [27], these representations have proven to capture syntax and semantics aspects as well as the diversity of the linguistic context of words. To extract utterance-level lexical features, we calculate mean values of word representations in the utterance. Finally, 1024-dimensional utterance-level lexical features are extracted.

3.3. Experimental Setup

As for the feature encoder (in Figure 1), AT-Fusion contains two fully-connected layer, mapping acoustic and lexical features into size $d = 100$. SA-GRU contains a bi-GRU layer (50 units for each GRU component) and a self-attention layer (100 dimensional states and 4 attention heads). To optimize the parameters, we use the Adam optimization, starting with an initial learning rate of 0.0001. We train our models for 100 epochs with a batch size of 20. Dropout [28] with $p = 0.2$ and $L2$ regularization with weight 0.00001 are also utilized to alleviate over-fitting problems. In our experiments, each configuration is tested 20 times with varied weight initializations. To compare our method with other advanced approaches [29, 30], weighted accuracy (WA) is chosen as our evaluation criterion. WA is a weighted mean accuracy over different emotion classes with weights proportional to the number of utterances in a particular emotion class.

3.4. Classification Performance of the Proposed Method

Two systems are evaluated in the experiments. In additional to the proposed system, one comparison systems are also implemented to verify the effectiveness of our proposed method:

(1) Our system (***Our***): It is our proposed method. For the emotion classifier, we train the classifier with the labeled data. For the domain classifier, we train the classifier with both the labeled and unlabeled data. Specifically, we find that we can gain the best performance by setting λ in Eq. (7) to be 1.

Table 2: *Experimental results of WA(%) for two systems under different training settings.*

	TS_1234	TS_123	TS_134	TS_234	TS_23
Cmp	81.06	80.82	79.85	78.89	77.60
Our	**81.14**	**82.68**	**82.27**	**82.43**	**81.39**
Δ	+0.08	+1.86	+2.42	+3.54	+3.79

(2) Comparison system (***Cmp***): It comes from ***Our***, but ignoring the domain classifier. Specifically, we only optimize the emotion classifier by setting λ in Eq. (7) to be 0.

Furthermore, to explore the impact of the amounts of labeled samples in the training set, five training settings are discussed, including ***TS_1234***, ***TS_123***, ***TS_134***, ***TS_234*** and ***TS_23***. These training settings follow the same naming way. For example, ***TS_123*** represents that the training data contains the labeled data from Session 1$\sim$3, and the unlabeled data from Session 4$\sim$5. ***TS_1234*** represents that the training data contains the labeled data from Session 1$\sim$4, and the unlabeled data from Session 5. As Session 5 is always unlabeled under these training settings, we evaluate the emotion recognition performance on Session 5. Experimental results of WA are listed in Table 2.

To verify the effectiveness of our proposed method, we compare the performance of ***Our*** and ***Cmp***. Experimental results in Table 2 demonstrate that our proposed method is superior to ***Cmp*** in all cases. Compared with ***Cmp***, our proposed method can learn speaker-independent representations. It ensures our model to focus on emotion-related information, while ignoring the difference between speaker identities. Therefore, our method increases generalization to unseen speakers, thus improving performance of emotion recognition.

To show the impact of the amount of training samples, we compare the performance under different training settings. Experimental results in Table 2 demonstrate that when we reduce training samples, ***Cmp*** has 0.2%$\sim$3.5% performance decrement. Without enough training samples, ***Cmp*** faces the risk of over-fitting, thus leading to performance decrement on the unlabeled data. Interestingly, we notice that our method gains 0.2%$\sim$1.5% performance improvement when we reduce training samples. Meanwhile, we compare the performance of ***Our*** and ***Cmp*** under different training settings. We observe that the margin of improvement increases with small amounts of training data. These phenomenons reveal that our method can utilize unlabeled samples properly. Therefore, our method achieves better performance than ***Cmp*** in low-resource conditions.

3.5. Comparison to State-of-the-art Approaches

To verify the effectiveness of the proposed method, we further compare our method with other currently advanced approaches. Experimental results of different methods are listed in Table 3.

Compared with our proposed method, these approaches [18, 19] also utilize DANN for emotion recognition. However, these methods ignore the contextual information and multimodal information in the training process. Experimental results in Table 3 demonstrate that our method is superior to [18, 19] with a large margin. This serves as strong evidence that considering contextual information and multimodal information in DANN can improve the performance of emotion recognition.

Compared with our proposed method, these approaches [11, 17, 29, 30, 31, 32] also utilized acoustic features and lexical features for emotion recognition. Context-free systems [29, 30, 31, 32] inferred emotions based only on the current utterance in conversations. While context-based networks [11] utilized the LSTMs to capture contextual information from their surroundings. However, context-based networks [11] suffered from incapability of capturing inter-speaker dependencies. To model the inter-speaker dependencies, Hazarika et al. [17] used memory networks to perform speaker-specific modeling. However, the inter-speaker influence was hard to evaluated. To avoid this problem, the proposed method learns speaker-independent representations via DANN, attempting to reduce the impact of inter-speaker influence. Experimental results in Table 3 demonstrate the effectiveness of the proposed method. Our proposed method shows an absolute improvement of 3.48% over state-of-the-art strategies. This serves as strong evidence that our context-dependent domain adversarial neural network can yield a promising performance for multimodal emotion recognition.

Table 3: *The performance of state-of-the-art approaches and the proposed approach on the IEMOCAP database.*

Approaches	WA (%)
Abdelwahab et al. (2018) [19]	56.68
Li et al. (2019) [18]	58.62
Rozgić et al. (2012) [29]	67.40
Jin et al. (2015) [30]	69.20
Poria et al. (2017) [11]	74.31
Li et al. (2018) [31]	74.80
Hazarika et al. (2018) [17]	77.62
Li et al. (2019) [32]	79.20
Proposed method	**82.68**

4. Conclusions

In this paper, we propose a context-dependent domain adversarial neural network for multimodal emotion recognition. To evaluate the effectiveness of our proposed method, we conduct experiments on the IEMOCAP database. Experimental results demonstrate that our method enables the model to focus on emotion-related information and ignore the difference between speaker identities. This is why we achieve better performance on unseen speakers compared with the fully supervised learning strategy. In the meantime, our method can utilize unlabeled samples properly, and achieve promising results in low-resource conditions. Furthermore, we prove that considering contextual information and multimodal information in DANN can improve the performance of emotion recognition. Due to above advantages, this novel framework is superior to state-of-the-art strategies for emotion recognition.

Future investigations include a detailed analysis of the amounts of unlabeled samples for DANN. Besides unlabeled samples in IEMOCAP, unlabeled samples from other corpora (e.g., SEMAINE) should also be evaluated. Additionally, besides acoustic and lexical modalities, we aim to further improve the classification accuracy using the visual information.

5. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

6. References

[1] J. Auguste, D. Charlet, G. Damnati, F. Béchet, and B. Favre, "Can we predict self-reported customer satisfaction from interactions?" in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 7385–7389.

[2] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5876–5880.

[3] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[4] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 6818–6825.

[5] C. Huang, A. Trabelsi, and O. R. Zaïane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," in *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, 2019, pp. 49–53.

[6] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 4970–4977.

[7] L. Chen, C. Chang, C. Zhang, H. Luan, J. Luo, G. Guo, X. Yang, and Y. Liu, "L2 learners' emotion production in video dubbing practices," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 7430–7434.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Ninth European Conference on Speech Communication and Technology*, 2005.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[10] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997, pp. 1039–1042.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2017, pp. 873–883.

[12] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 2017, pp. 1033–1038.

[13] C. Poultney, S. Chopra, Y. L. Cun *et al.*, "Efficient learning of sparse representations with an energy-based model," in *Proceedings of the Advances in Neural Information Processing Systems*, 2007, pp. 1137–1144.

[14] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5099–5103.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[16] C. Busso and S. S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[17] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 2122–2132.

[18] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7144–7148.

[19] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[20] 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, pp. 2096–2030, 2016.

[21] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion Review*, vol. 3, no. 1, pp. 8–16, 2011.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the Interspeech*, 2013.

[25] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 2227–2237.

[26] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," in *Proceedings of the Interspeech*, 2014, pp. 2635–2639.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the 1th International Conference on Learning Representations (ICLR)*, 2013.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[30] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4749–4753.

[31] R. Li, Z. Wu, J. Jia, J. Li, W. Chen, and H. Meng, "Inferring user emotive state changes in realistic human-computer conversational dialogs," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 136–144.

[32] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 5060–5066.