ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414654

MULTIMODAL CROSS- AND SELF-ATTENTION NETWORK FOR SPEECH EMOTION RECOGNITION

Licai Sun[1,2]*, Bin Liu*[2]*, Jianhua Tao*[1,2,3]*, Zheng Lian*[1,2]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
sunlicai2019@ia.ac.cn, {liubin, jhtao, zheng.lian}@nlpr.ia.ac.cn

ABSTRACT

Speech Emotion Recognition (SER) requires a thorough understanding of both the linguistic content of an utterance (i.e., textual information) and how the speaker utters it (i.e., acoustic information). The one vital challenge in SER is how to effectively fuse these two kinds of information. In this paper, we propose a novel Multimodal Cross- and Self-Attention Network (MCSAN) to tackle this problem. The core of MCSAN is to employ the parallel cross- and self-attention modules to explicitly model both inter- and intra-modal interactions of audio and text. Specifically, the cross-attention module utilizes the cross-attention mechanism to guide one modality to attend to the other modality and update the features accordingly. Similarly, the self-attention module employs the self-attention mechanism to propagate information within each modality. We evaluate MCSAN on two benchmark datasets, IEMOCAP and MELD. Experimental results demonstrate that our proposed model achieves state-of-the-art performance on both datasets.

Index Terms— speech emotion recognition, multimodal fusion, self-attention, cross-attention

1. INTRODUCTION

Emotion plays an important role in human communication. Speech Emotion Recognition (SER) aims to endow machines with the ability to perceive emotion. SER has a wide range of applications in the human-computer interaction field [1]. Takes the ubiquitous voice assistants (such as Amazon's Alexa and Apple's Siri) as an example. It's necessary for them to infer the user's emotion and respond properly to enhance the user experience.

Most recent studies on SER only focus on acoustic information. Various deep learning models have been developed to extract emotion relevant information from either handcrafted acoustic features or raw speech signals. Such as Convolution Neural Network (CNN) [2, 3], Recurrent Neural Network (RNN) [4, 5], self-attention mechanism [6] and their combinations [7, 8, 9]. The textual information embedded in speech is less exploited. This information is also crucial to SER because in some cases, the emotion of an utterance can be determined by the linguistic semantics. For example, "It's really a bad day!" indicates that the speaker is in a sad mood.

However, fusing the acoustic and textual information is not trivial. There are generally two kinds of interactions that need to be considered when fusing multimodal information, namely, the intra-modal interactions and the inter-modal interactions [10]. The intra-modal interactions refer to the fine-grained feature interactions within a single modality. Such as the frame-frame relationships in acoustic features and the word-word relationships in textual features. By modeling the intra-modal interactions, we can capture modality-specific patterns for emotion prediction. Since saying a sentence in different tones may deliver completely different emotions, it's necessary to model the frame-word relationships between audio and text. These are the so-called inter-modal interactions. The inter-modal interactions are either synchronous (for example, an emphasis on a specific word) or asynchronous (for example, laughter after speaking something funny).

Recently, several works have explored to fuse the acoustic and textual information for SER. Generally, these works can be categorized into three types. The first type builds independent models for each modality and combines their outputs for final emotion classification [11, 12, 13, 14]. Different architectures can be adopted for each modality to best suit different inputs. For example, Yoon et al. [11] employ two Long Short-Term Memory (LSTM) networks to encode audio and text. Tripathi et al. [13] apply 1D-CNN for word embeddings and 2D-CNN for spectral features. Although the intra-modal interactions can be captured, the inter-modal interactions are not explored. The second type utilizes the aligned audio and text as inputs [15]. The aligned features are first fused and then fed into a temporal model for sequential learning. Thus, the inter-modal interactions can be captured in the whole process. Nevertheless, the cost is to provide alignment information. To overcome this issue, the third type utilizes the attention mechanism to infer the latent cross-modal relationships between audio and text. Yoon et al. [16] propose a novel multi-hop mechanism to iteratively select and aggregate information from one modality by conditioning on the other modality. Xu et al. [17] utilize the attention mechanism to learn the latently aligned speech frames for each word. However, none of them explicitly model both intra- and inter-modal interactions of audio and text.

To address the above issues, we propose a novel Multimodal Cross- and Self-Attention Network (MCSAN) in this paper. MCSAN is mainly composed of a cross-attention module and two self-attention modules. The cross-attention module utilizes the cross-attention mechanism to propagate information between audio and text, while the self-attention modules employ the self-attention mechanism to propagate information within each modality. Thanks to these modules, MCSAN can explicitly model both inter- and intra-model interactions of audio and text. To verify MCSAN's effectiveness, we conduct experiments on two datasets. The results show that it outperforms state-of-the-art methods. We also perform ablation studies to justify the design choice of our model.

2. THE PROPOSED MODEL

As shown in Fig.1, MCSAN first uses an audio encoder and a text encoder to encode acoustic and textual features respectively. Then the encoded feature sequences are fed into the cross- and self-attention modules to learn the inter- and intra-modal interactions of audio and text. Finally, the outputs from these modules are concatenated and sent into a fully connected classifier for emotion prediction. Details are introduced as follows.

2.1. Audio Encoder

Suppose that the input acoustic feature sequence of an utterance is represented as $\mathbf{X}_a = \{\mathbf{x}_a^1, \mathbf{x}_a^2, ..., \mathbf{x}_a^{T'_a}\}^T \in \mathbb{R}^{T'_a \times d_a}$ (T'_a is the number of acoustic frames, d_a is the feature dimension). We adopt the architecture of CNN with LSTM as the audio encoder. Specifically, two 1D temporal convolutional layers are used to capture the local patterns. Since T'_a is typically large, each convolutional layer is followed by a max-pooling layer to reduce the temporal resolution and facilitate subsequent learning. Then a bidirectional LSTM (BiLSTM) layer is employed to capture the temporal dependencies within the sequence. The forward and backward hidden states of the BiLSTM layer are averaged to obtain the encoded acoustic features. The overall process can be summarized as follows:

$$\mathbf{X}^a = ConvBlock(ConvBlock(\mathbf{X}^a)) \tag{1}$$

$$\overleftarrow{\mathbf{h}}_a^t, \overrightarrow{\mathbf{h}}_a^t = BiLSTM(\mathbf{x}_a^t, \overleftarrow{\mathbf{h}}_a^{t+1}, \overrightarrow{\mathbf{h}}_a^{t-1}), t = 1, 2, ..., T_a \tag{2}$$

$$\mathbf{h}_a^t = \frac{1}{2}(\overleftarrow{\mathbf{h}}_a^t + \overrightarrow{\mathbf{h}}_a^t), t = 1, 2, ..., T_a \tag{3}$$

where $ConvBlock(\cdot) = MaxPool(Conv1D(\cdot))$, T_a is the number of acoustic frames after the second pooling layer. We denote $\mathbf{H}_a = \{\mathbf{h}_a^1, \mathbf{h}_a^2, ..., \mathbf{h}_a^{T_a}\}^T \in \mathbb{R}^{T_a \times d}$ (d is the unified feature dimension) as the encoded acoustic feature sequence.

2.2. Text Encoder

Suppose that the input textual feature sequence of an utterance is represented as $\mathbf{X}_l = \{\mathbf{x}_l^1, \mathbf{x}_l^2, ..., \mathbf{x}_l^{T_l}\}^T \in \mathbb{R}^{T_l \times d_l}$ (T_l is the number of words, d_l is the feature dimension). Considering that T_l is usually small, we only use a bidirectional LSTM layer to encode the word-level textual features. The encoded textual feature sequence $\mathbf{H}_l = \{\mathbf{h}_l^1, \mathbf{h}_l^2, ..., \mathbf{h}_l^{T_l}\}^T \in \mathbb{R}^{T_l \times d}$ can be obtained as follows:

$$\overleftarrow{\mathbf{h}}_a^t, \overrightarrow{\mathbf{h}}_a^t = BiLSTM(\mathbf{x}_a^t, \overleftarrow{\mathbf{h}}_a^{t+1}, \overrightarrow{\mathbf{h}}_a^{t-1}), t = 1, 2, ..., T_a \tag{4}$$

$$\mathbf{h}_l^t = \frac{1}{2}(\overleftarrow{\mathbf{h}}_l^t + \overrightarrow{\mathbf{h}}_l^t), t = 1, 2, ..., T_l \tag{5}$$

2.3. Cross-Attention Module

The cross-attention module aims to capture the inter-modal interactions between each pair of acoustic frames and textual words. The module is composed of a position embedding layer (for simplicity, we do not depict it in Fig.1) and N stacked cross-attention layers and feed-forward layers. The position embedding layer is used to inject temporal information into the feature sequence [18]. The main insight of the module is to utilize the cross-attention mechanism to learn the associations between two modalities and then propagate information from one modality to the other modality according to the learned associations. In the following part, we introduce the cross-attention mechanism in detail.

To learn the associations between audio and text, we first need to transform each feature sequence into three terms, which are the query, key, and value, using linear projections:

$$\mathbf{Q}_a, \mathbf{Q}_l = \mathbf{W}_a^Q \mathbf{H}_a, \mathbf{W}_l^Q \mathbf{H}_l \tag{6}$$

$$\mathbf{K}_a, \mathbf{K}_l = \mathbf{W}_a^K \mathbf{H}_a, \mathbf{W}_l^K \mathbf{H}_l \tag{7}$$

$$\mathbf{V}_a, \mathbf{V}_l = \mathbf{W}_a^V \mathbf{H}_a, \mathbf{W}_l^V \mathbf{H}_l \tag{8}$$

where $\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m \in \mathbb{R}^{T_m \times d}$ are the query, key, and value of the feature sequence of modality m, $m \in \{a, l\}$. $\mathbf{W}_m^Q$, $\mathbf{W}_m^K$, $\mathbf{W}_m^V \in \mathbb{R}^{d \times d}$ are the corresponding projection matrices.

Following [18], we calculate dot products of the query and key of audio and text in a crossed way to estimate the associations between two modalities. Then the results are scaled and row-wisely normalized by the softmax function to get the attention weights. After that, we aggregate the value term of each feature sequence using the corresponding weights to obtain the propagated information between two modalities:

$$\Delta\mathbf{H}_{a \to l} = softmax(\mathbf{Q}_l \mathbf{K}_a^T / \sqrt{d}) \mathbf{V}_a \tag{9}$$

$$\Delta\mathbf{H}_{l \to a} = softmax(\mathbf{Q}_a \mathbf{K}_l^T / \sqrt{d}) \mathbf{V}_l \tag{10}$$

where $\Delta\mathbf{H}_{a \to l} \in \mathbf{H}^{T_l \times d}$, $\Delta\mathbf{H}_{l \to a} \in \mathbb{R}^{T_a \times d}$ represents the propagated information from audio to text and text to audio, respectively.

The process described above is single-head attention. In practice, we use multi-head attention, which can be done by doing single-head attention multiple times and then combining the results of each head. The details can be found in [18].

Finally, we update the features of one modality with the propagated information from the other modality.

$$\mathbf{H}_a = LayerNorm(\mathbf{H}_a + \Delta\mathbf{H}_{l \to a}) \tag{11}$$

$$\mathbf{H}_l = LayerNorm(\mathbf{H}_l + \Delta\mathbf{H}_{a \to l}) \tag{12}$$

To further increase the representation capacity, a fully connected feed-forward layer [18] is added behind the cross-attention layer:

$$\mathbf{H}_a = LayerNorm(\mathbf{H}_a + FeedForward(\mathbf{H}_a)) \tag{13}$$

$$\mathbf{H}_l = LayerNorm(\mathbf{H}_l + FeedForward(\mathbf{H}_l)) \tag{14}$$

We denote the outputs of the last stacked layer in the module as $\mathbf{H}_a^c$ and $\mathbf{H}_l^c$, respectively.

2.4. Self-Attention Module

Parallel to the cross-attention module, the self-attention module aims to capture the intra-modal interactions within audio and text. This module is similar to the cross-attention module except for the usage of the self-attention mechanism. The self-attention mechanism shares the same spirit with the cross-attention mechanism. The only difference is that the query, key, and value are from the same modality. Thus, the whole process for one stacked layer in the self-attention module can be summarized as follows:

$$\Delta\mathbf{H}_m = softmax(\mathbf{Q}_m \mathbf{K}_m^T / \sqrt{d}) \mathbf{V}_m \tag{15}$$

$$\mathbf{H}_m = LayerNorm(\mathbf{H}_m + \Delta\mathbf{H}_m) \tag{16}$$

$$\mathbf{H}_m = LayerNorm(\mathbf{H}_m + FeedForward(\mathbf{H}_m)) \tag{17}$$

where $\Delta\mathbf{H}_m \in \mathbb{R}^{T_m \times d}$ is the propagated information within modality m, $m \in \{a, l\}$. We denote the outputs of the last stacked layer in two self-attention modules in Fig.1 as $\mathbf{H}_a^s$ and $\mathbf{H}_l^s$, respectively.

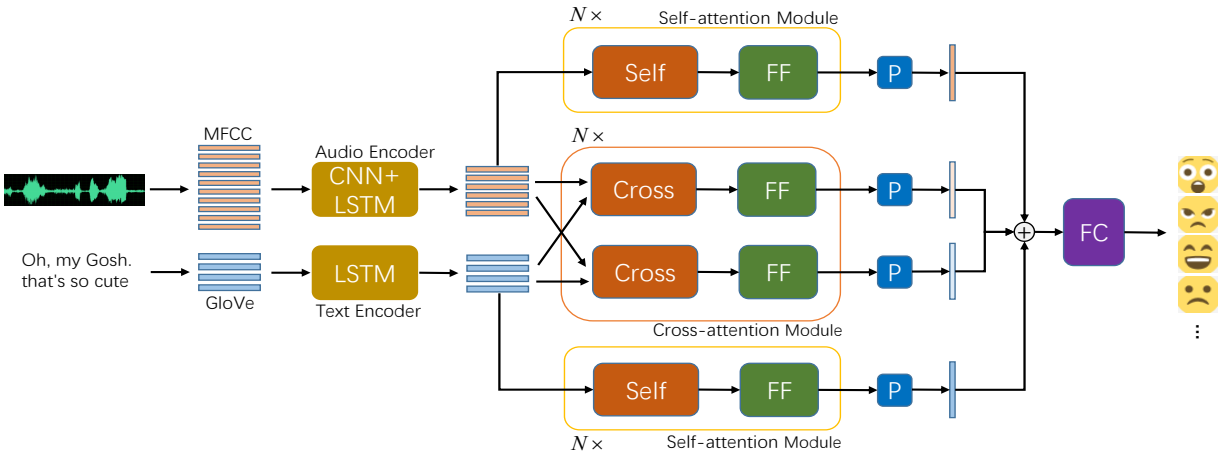

Fig. 1. The overall architecture of our proposed model. "Self": self-attention layer, "Cross": cross-attention layer, "FF": feed-forward layer, "P": global max-pooling layer. "FC": fully connected layer.

2.5. Classification

To perform final classification, we first summarize each output from the cross- and self-attention modules using a global max-pooling layer. Suppose that the summarized features for $\mathbf{H}_a^c$, $\mathbf{H}_l^c$, $\mathbf{H}_a^s$, $\mathbf{H}_l^s$ are $\mathbf{h}_a^c$, $\mathbf{h}_l^c$, $\mathbf{h}_a^s$, $\mathbf{h}_l^s \in \mathbb{R}^d$, respectively. Then we concatenate them to obtain the utterance-level representation. Finally, a fully-connected network and a softmax layer are followed to predict the underlying emotion. The cross-entropy loss is used to optimize the model. The above process is summarized as follows:

$$\mathbf{h} = Concat(\mathbf{h}_a^c, \mathbf{h}_l^c, \mathbf{h}_a^s, \mathbf{h}_l^s) \tag{18}$$

$$\hat{\mathbf{y}} = Softmax(f_\theta(\mathbf{h})) \tag{19}$$

$$\mathcal{L} = -\sum_i y_i log(\hat{y}_i) \tag{20}$$

where $\mathbf{y} = \{y_1, y_2, ..., y_n\}^T$ is the one-hot vector of the emotion label, $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}^T$ is the predicted probability distribution, n is the number of emotion categories, f_θ is the fully connected network with paramter θ.

3. EXPERIMENTS

3.1. Datasets

IEMOCAP [19] is the most commonly used dataset in SER. It contains about 12 hours of video recordings. To align with previous studies [20], we use 7,487 utterances from seven emotions: frustration, neutral, anger, sadness, excitement, happiness, surprise. Since there is no standard split for this dataset, we follow [20, 14] to perform 10-fold cross-validation, where 8:1:1 are used for training, validation and test, respectively. The weighted accuracy (WA, i.e., the overall accuracy) and unweighted accuracy (UA, i.e., the average accuracy over all emotion categories) is adopted as the evaluation metrics.

MELD [21] is a new multimodal dataset for emotion recognition in conversation. It consists of 13,708 utterances with seven emotions (i.e., anger, disgust, fear, joy, neutral, sadness, and surprise) of 1,433 dialogues from the classic TV-series Friends. The whole dataset is split into three parts: training (9,989), validation (1,109), and test (2,610). Following [21, 22], we report the weighted average F1 score on this dataset.

3.2. Implementation Details

We extract 40-dimensional Mel-Frequency Cepstral Coefficients (MFCC) from speech signals. The window size and hop size are set to 25 ms and 10 ms, respectively. The max length of the MFCC feature sequence is set to 1000. We perform z-normalization before feeding them into the audio encoder. For textual features, we first apply word-tokenizer to the transcripts provided by the datasets. Then each word in an utterance is embedded into a 300-dimensional vector using the pre-trained GloVe model [23].

We implement our model within the PyTorch framework. The number of hidden neurons in the model is 128. The number of stacked layers in the cross- and self-attention module is 1. The number of heads is set to 8. The kernel size of the convolutional and max-pooling layer in the audio encoder is 3. To train the model, we use a Adam optimizer [24] with a learning rate of 0.001 on IEMOCAP and 0.0005 on MELD. The batch size is 256. We train the model at most 30 epochs on IEMOCAP and 20 epochs on MELD.

3.3. Baselines

On the IEMOCAP dataset, the following baselines are used for comparison:

1. MDRE [11] employs dual recurrent neural networks to encode audio and text and then combines the results of two modalities using a fully connected neural network for final emotion classification.
2. MHA [16] is based on MDRE, which additionally utilizes a novel multi-hop attention mechanism to automatically infer the correlation between audio and text.
3. Xu et al. [17] propose to use the attention mechanism to learn the latent alignment between audio and text.
4. CAN [14] aggregates the sequential information from aligned audio and text by using the attention weights of each modality in a normal and crossed way.

On the MELD dataset, we use two baselines for comparison:

1. cMKL [25] adopts CNN for feature extraction and uses multiple kernel learning to fuse multimodal features.
2. Liang et al. [22] employ two deep auto-encoders to learn latent representations of audio and text and concatenate them for classification.

3.4. Comparison to State-of-the-art Methods

We first evaluate our proposed model on the IEMOCAP dataset. The results [1] of 10-fold cross-validation on the dataset are presented in Table 1. From the table, we can observe that MCSAN outperforms above baseline models. Specifically, MCSAN improves the state-of-the-art CAN model by 3.3% absolute value in terms of WA. We should also mention that CAN needs the aligned audio and text as input. However, by virtue of the cross-attention mechanism, our model does not need alignment information. The improvement in terms of UA is even higher. MCSAN outperforms the state-of-the-art MHA model by 6.9%. We also present the performance of AMH [20] in Table 1. AMH is a tri-modal version of MHA by incorporating the visual information into MHA's framework. Although MCSAN only exploits the acoustic and textual information, it is comparable to AMH by achieving slightly worse performance in terms of WA but better performance in terms of UA. These results show the superiority of our proposed model.

Table 1. Model performance comparison on the IEMOCAP dataset. The results of 10-fold cross-validation are presented as $mean \pm std$. The result of Xu et al. is from [14]. "A": audio modality, "L": textual modality, "V": visual modality.

Model	Modality	WA	UA
MDRE [11]	A+L	0.498 ± 0.059	0.418 ± 0.077
MHA [16]	A+L	0.543 ± 0.026	0.491 ± 0.028
Xu et al. [17]	A+L	0.560 ± 0.028	0.450 ± 0.028
CAN [14]	A+L	0.579 ± 0.019	0.487 ± 0.017
AMH [20]	A+V+L	0.617 ± 0.016	0.547 ± 0.025
MCSAN (ours)	A+L	$\mathbf{0.612 \pm 0.012}$	$\mathbf{0.560 \pm 0.019}$

To further demonstrate the effectiveness of MCSAN, we then evaluate it on the MELD dataset. Table 2 presents the results on the test set of this dataset. We can notice that MCSAN outperforms the state-of-the-art by 3.1% absolute value in terms of weighted average F1 score. Moreover, our model even exceeds the corresponding semi-supervised model (denoted by "semi" in Table 2) which makes use of a large amount of unlabeled data.

Table 2. Model performance comparison on the MELD dataset. The result of cMKL is from [21].

Model	F1
cMKL [25]	0.555
Liang et al. [22]	0.561
Liang et al. [22] (semi)	0.571
MCSAN (ours)	**0.592**

[1] We adopt the revised results of MDRE, MHA, and AMH from the Github repository of AMH's author: https://github.com/david-yoon/attentive-modality-hopping-for-SER.

3.5. Ablation Study

In this section, we conduct several experiments on IECMOCAP to evaluate several key factors in our proposed model. Table 3 presents the results. First, we evaluate the effect of modality. From the table, we can observe a significant performance drop when only utilizing acoustic or textual information as input. This suggests that it's vital for SER systems to effectively fuse these two kinds of information. Second, we evaluate the effect of attention modules. When the self- or cross-attention module is removed, the model's performance decrease by 0.7%/1.6% in terms of WA and 1.0%/2.4% in terms of UA. This verifies the importance of these two modules and demonstrates that it's necessary to explicitly model both the inter- and intra-modal interactions. Besides, the model's performance is worse when the cross-attention module is removed, which indicates that modeling of inter-modal interactions is more critical than the modeling of intra-modal interactions. Third, we evaluate the effect of the model's architecture. Instead of putting the attention modules in parallel, we combine them in a sequential manner with different orders. However, neither "cross+self (seq)" nor "self+cross (seq)" is superior to the parallel architecture. Finally, we evaluate the effect of the model's capacity. We find that the model's performance goes down when we stack more layers in the self- and cross-attention modules. We believe that this might be caused by overfitting because the dataset may be too small to fully train large models.

Table 3. Ablation study on the IEMOCAP dataset.

Model	WA	UA
MCSAN	0.612 ± 0.012	0.560 ± 0.019
w/o audio	0.509 ± 0.010	0.469 ± 0.028
w/o text	0.491 ± 0.011	0.404 ± 0.010
w/o self	0.605 ± 0.013	0.550 ± 0.026
w/o cross	0.596 ± 0.012	0.536 ± 0.028
cross + self (seq)	0.606 ± 0.011	0.552 ± 0.025
self + cross (seq)	0.602 ± 0.012	0.554 ± 0.024
$N = 2$	0.601 ± 0.015	0.551 ± 0.025
$N = 3$	0.595 ± 0.012	0.541 ± 0.018

4. CONCLUSION

In this paper, we propose a novel Multimodal Cross- and Self-Attention Network (MCSAN) for speech emotion recognition. Thanks to the parallel cross- and self-attention modules, MCSAN can explicitly model the inter- and intra-modal interactions within/between audio and text. Experimental results on IEMOCAP and MELD demonstrate the effectiveness of MSCAN. In the future, we plan to extend our model to a tri-modal version by incorporating the visual information into our framework.

5. ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Plan of China (No.2018YFB1005003), 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] Björn W Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[3] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 21–26.

[4] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[5] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[6] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang, "Unsupervised representation learning with future observation prediction for speech emotion recognition," in *Proceedings of Interspeech*, 2019, pp. 3840–3844.

[7] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[8] Zixing Zhang, Bingwen Wu, and Björn Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6705–6709.

[9] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang, "Conversational emotion analysis via attention mechanisms," in *Proceedings of Interspeech*, 2019, pp. 1936–1940.

[10] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*. NIH Public Access, 2018, vol. 2018, p. 5642.

[11] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[12] Bagus Tris Atmaja, Kiyoaki Shirai, and Masato Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 519–523.

[13] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and Promod Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," 2019.

[14] Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung, "Multimodal speech emotion recognition using cross attention with aligned audio and text," *Proc. Interspeech 2020*, pp. 2717–2721, 2020.

[15] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2225–2235.

[16] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[17] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[20] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3362–3366.

[21] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[22] Jingjun Liang, Ruichen Li, and Qin Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, MM '20, p. 2852–2861, Association for Computing Machinery.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.