IS CROSS-ATTENTION PREFERABLE TO SELF-ATTENTION FOR MULTI-MODAL EMOTION RECOGNITION?

Vandana Rajan[1], *Alessio Brutti*[2], *Andrea Cavallaro*[1]

[1]Centre for Intelligent Sensing, Queen Mary University of London, UK
[2]Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

Humans express their emotions via facial expressions, voice intonation and word choices. To infer the nature of the underlying emotion, recognition models may use a single modality, such as vision, audio, and text, or a combination of modalities. Generally, models that fuse complementary information from multiple modalities outperform their uni-modal counterparts. However, a successful model that fuses modalities requires components that can effectively aggregate task-relevant information from each modality. As cross-modal attention is seen as an effective mechanism for multi-modal fusion, in this paper we quantify the gain that such a mechanism brings compared to the corresponding self-attention mechanism. To this end, we implement and compare a cross-attention and a self-attention model. In addition to attention, each model uses convolutional layers for local feature extraction and recurrent layers for global sequential modelling. We compare the models using different modality combinations for a 7-class emotion classification task using the IEMOCAP dataset. Experimental results indicate that albeit both models improve upon the state-of-the-art in terms of weighted and unweighted accuracy for tri- and bi-modal configurations, their performance is generally statistically comparable. The code to replicate the experiments is available at `https://github.com/smartcameras/SelfCrossAttn`

***Index Terms*—** Multi-modal, emotion recognition, attention

1. INTRODUCTION

Emotion recognition (ER) models use one or more modalities, such as audio (language and para-language), images (facial expressions and body gestures) and text (language) to infer the class of underlying emotion [1]. Multi-modal models are designed to effectively fuse relevant information from different modalities and generally outperform uni-modal models [2, 3]. ER models may use as input raw signals (speech or face images) [4, 5, 6] or handcrafted features [3, 7]. Commonly used speech features are low-level descriptors, such as formants, pitch, log energy, zero-crossing rate and Mel Frequency Cepstral Coefficients (MFCCs) [3, 7]. Facial expressions can be represented by fixed features based on entities that are always present on the face, such as eyes, mouth and eyebrows and/or transitory features based on temporary entities like wrinkles and bulges [8]. Tokenized words can be mapped into linguistic features using word embedding algorithms, such as word2vec [9] or GloVe [10].

ER models based on Deep Neural Networks (DNNs) may contain convolutional layers to extract local task-relevant components from the input and recurrent layers to facilitate the global sequential modelling [4, 5]. Attention mechanisms [11] integrated in DNN architectures encourage the ER model to focus on task-relevant time instants [3, 6]. The general purpose of attention mechanism is to provide varying levels of weights to different time-steps in a sequence. There are two types of attention mechanisms, namely self (or intra-modal) attention and cross (or inter-modal) attention. A self-attention mechanism computes the representation of a uni-modal sequence by relating different positions of the same sequence [6, 12]. Cross-modal attention mechanisms use one modality to estimate the relevance of each position in another modality [13]. For example, a self-attention mechanism between 2 recurrent layers can be used to emphasise task-relevant time-steps in an input speech signal [6], whereas an iterative multi-hop cross-attention mechanism may select and aggregate information from multi-modal features obtained with Gated Recurrent Unit (GRU) layers [3, 7]. Transformers [14], which contain a Multi-Head Attention (MHA) module, are also becoming popular in modelling uni-modal as well as multi-modal emotional data [15, 13, 16, 17]. Cross-modal transformers use cross-attention to calculate the relevance of each time-step in a target modality representation using a different source-modality [13, 17]. A serial [13, 17] or parallel [16] combination of cross and self-attention transformers aims to capture the cross-modal and intra-modal relationships for multi-modal fusion. Considering the interest in models combining self and cross attention-based transformer encoders [13, 16, 17], we conduct the first study comparing the two types of attention mechanisms (without the other transformer components). To understand the differences between the two types of attention mechanisms, we extensively compare a model based only on cross-attention and one based only on self-attention for bi- and tri-modal combinations. We compare the two models on the IEMOCAP [18] dataset for 7-class emotion classification and conclude that the cross-attention model does not outperform the self-attention model. Nevertheless, both models improve the state-of-the-art results on tri-modal as well as bi-modal emotion recognition tasks in terms of weighted and unweighted accuracy metrics.

2. CROSS AND SELF ATTENTION MODELS

Self and cross-attention models first process individual modalities using modality-specific encoders. The encoded features are then fed into self or cross Multi-Head Attention (MHA) [14] modules, respectively. A global representation of the utterance clip is generated as temporal average at the outputs of each attention module. The resulting features are then concatenated and their mean and standard deviation are obtained using a statistical pooling layer. The concatenation of mean and standard deviation vectors is then fed to fully connected layers. The emotional class predictions are obtained through a softmax operation. A detailed explanation is given as follows:

Let $X_a \in \mathbb{R}^{t_a \times d_a}$ be the audio features corresponding to an ut-

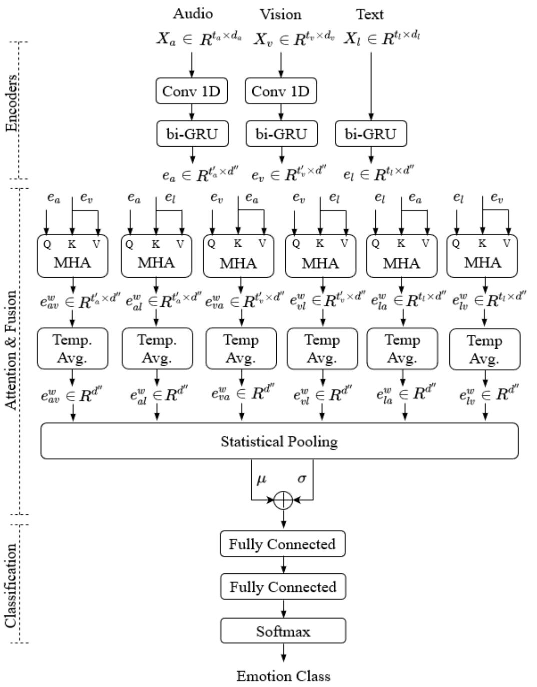

Figure 1: Architecture of the tri-modal cross-attention model. KEY - MHA: Multi-Head Attention; Temp.: temporal; Avg.: averaging; μ: mean; σ: standard deviation; $\bigoplus$: concatenation operation.

terance clip, where t_a is the sequence length and d_a is the feature dimension. The audio encoder consists of a 1D convolution layer followed by a bi-directional GRU. The convolution layer, which refines the input feature sequence by finding task-relevant patterns, operates as follows:

$$X'_a(t') = b(t') + \sum_{k=0}^{t_a-1} (W(t', k) * X_a(k)), \tag{1}$$

where $X'_a \in \mathbb{R}^{t'_a \times d'_a}$ is the output with length t'_a and dimension d'_a, $t' \in [0, t'_a - 1]$, $*$ is the convolution operator, W are the weights and b are the biases associated with the layer. Thus, the convolution layer modifies the sequence length as well as the feature dimension.

The bi-directional GRU layer models contextual inter-dependence of the features across time. For each element in the sequence, the bi-GRU layer computes the following functions:

$$\begin{cases} r_t = \sigma(W_{ir}X'_a(t) + b_{ir} + W_{hr}h_{t-1} + b_{hr}), \\ z_t = \sigma(W_{iz}X'_a(t) + b_{iz} + W_{hz}h_{t-1} + b_{hz}), \\ n_t = \phi_h(W_{in}X'_a(t) + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})), \\ h_t = (1 - z_t) \odot n_t + z_t \odot (h_{t-1}), \end{cases} \tag{2}$$

where h_t and h_{t-1} are the hidden states at times t and $t-1$, $X'_a(t)$ is the input at time t. r_t, z_t and n_t are the reset, update and new gates, W and b are the corresponding weights and biases, σ and ϕ_h are the sigmoid and hyperbolic tangent functions and $\odot$ is the Hadamard product. At the output of bi-GRU, the forward and backward hidden states for each time-step are concatenated and the refined audio features can be represented as $e_a \in \mathbb{R}^{t'_a \times d''}$, where d'' is twice the number of hidden neurons in the GRU.

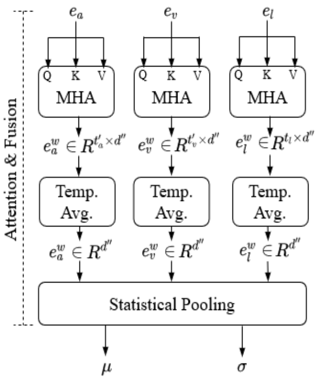

Figure 2: Attention and fusion module in the tri-modal self-attention model. The rest of the model is same as the tri-modal cross-attention model. KEY - MHA: Multi-Head Attention; Temp.: temporal; Avg.: averaging; μ: mean; σ: standard deviation.

Similar to audio, the vision encoder consists of one 1D convolution layer followed by a bi-GRU layer. If $X_v \in \mathbb{R}^{t_v \times d_v}$ represents the vision features corresponding to an utterance, then at the output of vision encoder, the features are refined to $e_v \in \mathbb{R}^{t'_v \times d''}$. For the text modality, the encoder consists of only one bi-GRU layer. The input and output of text encoder can be represented by $X_l \in \mathbb{R}^{t_l \times d_l}$ and $e_l \in \mathbb{R}^{t_l \times d''}$ respectively.

We use the MHA module [14] for self and cross-attention modelling. An MHA module consists of multiple such attention operations to capture richer interpretations of the sequence. Each MHA module requires 3 inputs, namely, Query (Q), Key (K) and Value (V), each of which is first projected H times to different sub-spaces using linear layers, where H refers to number of heads. Projections for each sub-space $h \in \{0, ..., H-1\}$ can be calculated as

$$Q_h = W_h^Q e_m, \tag{3}$$

$$K_h = W_h^K e_m, \tag{4}$$

$$V_h = W_h^V e_m, \tag{5}$$

where $m \in \{a, v, l\}$ denotes the modality. In each of these sub-spaces, scaled dot-product attention is performed on the projections. For a sub-space h, the attention operation is given as

$$Att_h(Q_h, K_h, V_h) = \texttt{Softmax}\left(\frac{Q_h {K_h}^T}{\sqrt{d_k}}\right) V_h, \tag{6}$$

where $Att_h(\cdot)$ and d_k refer to the attention operation in sub-space h and feature dimensionality, respectively. The outputs of all H attentions are concatenated and passed through a linear layer to obtain the final output of an MHA module.

In the cross-attention model, a source modality is given as K and V, whereas a target modality is fed as Q (see Fig. 1). The intuition behind such an approach is to discover cross-modal interactions by adapting the source modality to the target modality [13]. As an example, let us take the case of audio as target modality and vision as the source modality. The refined audio features $e_a \in \mathbb{R}^{t'_a \times d''}$ are transformed to Q using Eq. 3 and vision features $e_v \in \mathbb{R}^{t'_v \times d''}$ to K and V using Eq. (4)-(5). The cross-modal MHA module then maps the vision to the audio modality and outputs vision features adapted to audio $e^w_{av} \in \mathbb{R}^{t'_a \times d''}$. Note that the sequence length of the cross-attention weighted output is the same as the target modality audio.

Table 1: Results of a 7-class emotion classification task presented as mean ± standard deviation. AMH refers to AMH [3] for tri-modal models and to MHA [7] for bi-modal models. KEY - A: audio; V: vision; T: text; Self: self-attention model; Cross: cross-attention model. The best results in each row are in bold font. The symbol * refers to the only three results with statistically significant difference between the self and cross models.

	Weighted Accuracy				**Unweighted Accuracy**			
Modality	MDRE [19]	AMH [3, 7]	Cross	Self	MDRE [19]	AMH [3, 7]	Cross	Self
T	-	-	-	.474 ± .030	-	-	-	.535 ± .016
V	-	-	-	.454 ± .019	-	-	-	.513 ± .018
A	-	-	-	.365 ± .018	-	-	-	.452 ± .017
T+V	.524 ± .021	.526 ± .024	**.567 ± .022**	.563 ± .022	.579 ± .015	.580 ± .019	**.617 ± .015**	.614 ± .020
T+A	.418 ± .077	.491 ± .028	.501 ± .026	**.518 ± .031***	.498 ± .059	.543 ± .026	.562 ± .017	**.574 ± .018***
V+A	.376 ± .024	.371 ± .042	.481 ± .024	**.483 ± .026**	.477 ± .025	.471 ± .047	.566 ± .022	**.567 ± .026**
T+V+A	.490 ± .056	.547 ± .025	.578 ± .024	**.587 ± .022***	.564 ± .043	.617 ± .016	.636 ± .017	**.642 ± .019**

With 3 modalities, we have 6 combinations of source-target modalities and hence we use 6 MHA modules. In case of self-attention model, the input sequence corresponding to the same modality is used as Q, K and V (see Fig. 2). This helps to capture intra-modal interactions in each modality. For cross-attention model, statistical pooling is done across the concatenation of the temporal averages of 6 cross-modal sequences, whereas for the self-attention model, it is done across the concatenation of the temporal averages of the self-attended sequences of all the 3 modalities.

The classifier for both models is:

$$\hat{y} = \texttt{Softmax}(f_{\theta_2}(f_{\theta_1}([\mu \parallel \sigma]))), \tag{7}$$

where μ and σ are the mean and standard deviation obtained from the output of statistical pooling layer, $\|$ represents concatenation operation, f_{θ_1} and f_{θ_2} denote the 2 fully connected layers with parameters θ_1 and θ_2, respectively, and $\hat{y}$ denotes the one-hot vector of emotion prediction.

3. VALIDATION

In this section, we discuss the dataset and results of using cross- and self-attention models for 7-class bi-modal and tri-modal emotion recognition. We also discuss comparison with state-of-the-art methods and experiments with additional model configurations for both models.

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [18] dataset which contains approximately 12 hours of audio-visual dyadic emotional interactions in acted and spontaneous settings. The dataset, recorded with 5 male and 5 female speakers, includes the ground-truth text transcripts. The labelling of each utterance was determined by majority voting from 3 annotators. There is lack of consensus amongst researchers on the use of IEMOCAP dataset. Some use it for 4 class classification [13] by merging different classes (*happy* and *excited*, *angry* and *frustrated*), while others [3, 7, 19, 16] perform 7-class classification. We follow the latter. Since the creators of the dataset did not define a training-testing split, we use the same dataset partition and features as [3, 7, 19]. The final dataset contains 7,487 utterances in total (1,103 *angry*, 1,041 *excited*, 595 *happy*, 1,084 *sad*, 1,849 *frustrated*, 107 *surprise* and 1,708 *neutral*). Class sizes smaller than 100 utterances (*fear*, *disgust*, *other*) are eliminated [3]. We perform 5-fold cross validation to assess the model performance. Data in each fold are split into training, development, and testing sets (8:0.5:1.5). We train and evaluate the model 10 times (with 10 different random seeds) per fold, and the performance is assessed in terms of weighted accuracy (WA) and unweighted accuracy (UWA) metrics.

For the audio modality, 40D MFCC features (frame size is set to 25 ms at a rate of 10 ms with the Hamming window) are extracted and concatenated with their first and second order derivatives to obtain the final acoustic feature dimension of 120. Audio features are standardised by removing the mean and scaling to unit variance. For vision data, cropped face images of speakers are fed into a ResNet-101 [20] to obtain 2048D features at a frame rate of 3 Hz. For text modality, each word in an utterance is represented by a 300D GloVe [10] embedding. Note that the modalities are sampled at different rates and the maximum sequence length of audio, vision and text modalities is set to 1,000, 32 and 128 respectively.

The models are implemented using PyTorch [21]. The bi-modal and uni-modal versions of the tri-modal models are created by removing components corresponding to the unused modality/modalities. We use Adam [22] optimiser with a learning rate of 0.001. The learning rate is reduced by a factor 0.1 when the validation loss has stopped decreasing for 10 consecutive epochs. Training is stopped when UWA does not improve in the validation set for 10 consecutive epochs and the model with best validation UWA is used for testing. The batch size is 32 and all models are trained using the categorical cross-entropy loss.

The audio and vision encoders contain one 1D convolution layer each. The kernel size and stride length are both set to 1. The number of input and output channels for audio convolution layer are 1,000 and 500 respectively while for vision they are 32 and 25 respectively. The number of bi-GRU layers for all the 3 modalities is 1. The number of hidden neurons in each bi-GRU layer is 60. The number of attention heads in all MHA modules is 6 and a dropout rate of 0.1 is applied to reduce overfitting. The number of neurons in the first and second fully connected output layers are 60 (same as number of bi-GRU neurons) and 7 (number of output classes) respectively. All parameters were chosen based on the performance on validation set.

Table 1 shows the performance of the self and cross-attention models on 7-class uni-modal, bi-modal and tri-modal emotion recognition tasks. We report the mean and standard deviation obtained across 50 runs (5 folds × 10 repetitions) for each model. We also applied two-tailed t-test with the null hypothesis that the accuracy values of both self and cross-attention models have identical average (expected) values. Comparison of the uni-modal performances shows that the text outperforms the vision and audio modalities. This result is consistent with previous work [13, 19]. Since uni-modal performance evaluation is not possible with the cross-modal model, we report results with the uni-modal version of the self-attention

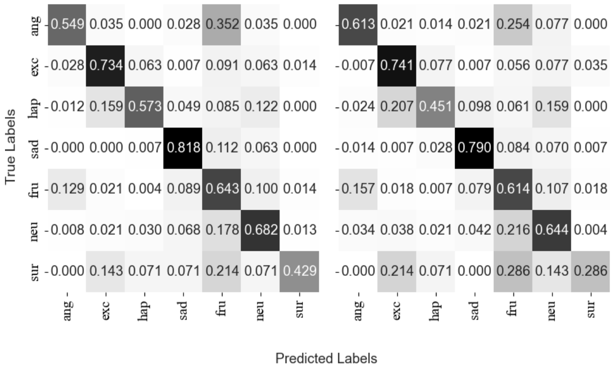

Figure 3: Confusion matrices of self (left) and cross-attention models (right) for tri-modal 7-class classification using a random fold. The emotions classes are abbreviated with their first 3 letters.

model. Among bi-modal models, the combination of vision and text modalities gives the best performance for both models. These results are consistent with previous work [7, 19]. Overall, both models provide comparable performances for bi- and tri-modal cases. Self-attention significantly outperforms cross-attention (P value $< .05$) only for T+A (text and audio) and the WA of T+V+A (text, video, and audio).

We compare with methods that use the same set of features and dataset partition. The tri-modal models are compared with AMH [3], the current state-of-the-art model, which uses a combination of uni-modal GRU layers and an iterative attention mechanism[1]. Note that the self-attention model exceeds the performance of AMH by 4.0 and 2.5 percentage points (pp) over mean in terms of WA and UWA, respectively. Similar figures for the cross-attention model are 3.1 pp and 1.9 pp. We also compare with MDRE [19], which uses recurrent layers to model uni-modal signals followed by aggregation and classification using fully connected layers. The better performance of the self and cross-attention models, as well as AMH, compared to MDRE can be attributed to the effectiveness of the attention mechanism. For bi-modal models, we compare with the bi-modal version of AMH called MHA [7] and MDRE. Again, both models outperform MHA and MDRE in all the 3 bi-modal cases. Note that we obtain bi-modal results by ablating the tri-modal models and not by fine-tuning for individual bi-modal cases. Also, AMH, MHA and MDRE use prosody features in addition to MFCC features for audio, whereas we use only MFCC features. The state-of-the-art result for text+audio case is obtained by [16] (0.560 WA and 0.612 UWA) which is significantly higher than the bi-modal T+A (text and audio) results. We hypothesize two reasons for this: (1) unlike [16], the bi-modal models are not fine-tuned for the bi-modal cases; (2) [16] uses transformer encoders that contain additional parameters that might help in learning more complex inter-modal relationships, whereas we use only the multi-head attention mechanism. Nevertheless, both models improve the state-of-the-art tri-modal results of AMH.

Fig. 3 shows the confusion matrices for the self and cross-attention models. For both models we can observe that the classes *angry* and *frustrated* are more often confused with each other, and the class *happy* gets confused with *excited* (these 2 classes are inherently similar). The poor performance of both models on the class *surprise* can be attributed to the fact that this has the smallest sample size in the dataset. These observations are consistent with the previous literature [3].

Table 2: Weighted accuracy (WA) and Unweighted accuracy (UWA) for 7-class emotion classification using additional tri-modal model configurations. Self and Cross model results are also shown for comparison. KEY - SP: statistical pooling; Cross-noSP and Self-noSP: cross and self-attention models without SP; Cross+Self: combination model that concatenates mean and standard deviation vectors from self and cross-attention models.

Model	**WA**	**UWA**
Cross-noSP	.570 ± .021	.634 ± .015
Cross	.578 ± .024	.636 ± .012
Self-noSP	.584 ± .021	.638 ± .019
Self	**.587 ± .022**	**.642 ± .019**
Cross+Self	.585 ± .028	.642 ± .020

In addition to the two described model configurations, we also experimented with different variations of the tri-modal models. We removed the statistical pooling layer from both models to assess its significance. The outputs from all temporal averaging modules (see Fig. 1 & 2) were concatenated and passed to the classifier module. These models are shown as Cross-noSP and Self-noSP in Table 2. We can make two observations. Firstly, the self-attention model outperforms the cross-attention model (P value $< .05$ for WA) even after ablating statistical pooling. Secondly, the performance of both models decreases without the statistical pooling layer. We also assessed the performance of a combined model created by merging the self and cross-attention models (Cross+Self). The statistical pooling output from both models were concatenated and fed to a common classifier module. We can see that the performance is similar to that of the self-attention model. This might indicate that the cross-attention model does not contribute any additional, relevant information compared to that of the self-attention model.

4. CONCLUSION

Intrigued by the popularity of cross-attention mechanism in multi-modal fusion, we compared models based on self-attention and on cross-attention using the IEMOCAP dataset for tri-modal and bi-modal 7-class classification. Results show that there is no meaningful difference between the results of the two models. Thus, within the context of the dataset and architecture we used, we conclude that cross-attention does not outperform self-attention for multi-modal emotion recognition. Furthermore, both the self and the cross-attention models improve the state-of-the-art in the recognition task. Future work includes investigating the effectiveness of cross and self-attention models for other multi-modal tasks and modalities.

Acknowledgement. We thank the authors of [3, 7, 19] for providing the processed and partitioned IEMOCAP dataset. We acknowledge the use of the ESPRC funded Tier 2 facility, JADE.

5. REFERENCES

[1] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis, "Multimodal emotion recognition from expressive

[1] We use the revised results of AMH, MHA and MDRE from `https://github.com/david-yoon/attentive-modality-hopping-for-SER`. We note that the WA and UWA values were swapped by the authors and we rectify this error in Table 1.

faces, body gestures and speech," in *Proc. of the Int. Conf. on Artificial Intelligence Applications and Innovations*. Springer, 2007, pp. 375–388.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.

[3] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2020, pp. 3362–3366.

[4] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2016, pp. 5200–5204.

[5] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. of Selected Topics in Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017.

[6] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro, "Conflictnet: End-to-end learning for speech-based conflict intensity estimation," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1668–1672, 2019.

[7] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2019, pp. 2822–2826.

[8] Beat Fasel and Juergen Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[9] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," in *Int. Conf. on Learning Representations*, 2013.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[11] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Int. Conf. on Learning Representations*, 2015.

[12] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[13] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of the Conf. Association for Computational Linguistics*. NIH Public Access, 2019, p. 6558.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Adv. in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro, "Cross-modal knowledge transfer via inter-modal translation and alignment for affect recognition," *arXiv preprint arXiv:2108.00809*, 2021.

[16] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2021, pp. 4275–4279.

[17] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2020, pp. 3507–3511.

[18] 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. 335–359, 2008.

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

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NIPS-Workshop on Autodiff*, 2017.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations*, 2015.