Fusing information from multiple sources for Multimodal Emotion Recognition

Aditya Bobde
G H Raisoni College of Engineering
Dept. of Artificial Intelligence
Nagpur, India
bobde_aditya.ai@ghrce.raisoni.net

Shivam Tawari
G H Raisoni College of Engineering
Dept. of Artificial Intelligence
Nagpur, India
tawari_shivam.ai@ghrce.raisoni.net

Vishal Narnaware
G H Raisoni College of Engineering
Dept. of Artificial Intelligence
Nagpur, India
narnaware_vishal.ai@ghrce.raisoni.net

Adarsh Suryawanshi
G H Raisoni College of Engineering
Dept. of Artificial Intelligence
Nagpur, India
suryawanshi_adarsh.ai@ghrce.raisoni.net

Achamma Thomas
G H Raisoni College of Engineering
Dept. of Artificial Intelligence
Nagpur, India
achamma.thomas@raisoni.net

***Abstract*—Emotion recognition is critical in affective computing, with applications ranging from healthcare to entertainment. Multimodal Machine Learning has emerged as a promising approach to improve the accuracy and robustness of emotion recognition systems by combining information from multiple modalities, such as audio, visual, and textual data. In this paper, we present our implementation of Multimodal Machine Learning for emotion recognition using state-of-the-art models for feature extraction of modalities, namely VideoMAE for video, MPNet for text, and CLAP for audio. This research discusses and reviews cross-modal interactions via different fusion techniques, such as Concatenation, TensorFusion, and Low-Rank Multimodal Fusion for Multimodal Emotion Recognition. We conduct experiments on the Multimodal EmotionLines Dataset (MELD) and propose a new framework leveraging Multimodal Machine Learning to understand verbal, vocal, and visual modalities and classify given content into one of seven emotions - Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear.**

***Index Terms*—Multimodal Machine Learning, Multimodal Emotion Recognition, Fusion Techniques, TensorFusion, Low-Rank Multimodal Fusion, VideoMAE, MPNet, CLAP**

I. Introduction

Emotion recognition has become a crucial area of research in the field of affective computing, as it plays a vital role in various applications, including healthcare, education, entertainment, and human-computer interaction. Due to the complexity and subjectivity of emotions, as well as the variability and noise in the data, it is difficult to accurately identify human emotions from several modalities, including facial expressions, voice, physiological signals, and textual content. Therefore, researchers have been exploring various approaches to improve the accuracy and robustness of emotion recognition systems, and one such approach is Multimodal Machine Learning (MML).

MML is an emerging research field that combines information from different modalities to improve the performance of machine learning models. In the context of emotion recognition, MML involves integrating data from multiple modalities, such as visual, audio, and textual data, to enhance the accuracy of emotion recognition models. The rationale behind this approach is that different modalities provide different cues about human emotions, and integrating them can lead to more accurate and robust emotion recognition.

In recent years, MML has shown promising results in emotion recognition, and researchers have been exploring various approaches for feature extraction, fusion, and classification in MML [1], [2]. Feature extraction involves extracting relevant features from each modality, such as facial landmarks, speech features, and textual features. Feature fusion involves combining the extracted features from different modalities to obtain a unified data representation. Classification involves using machine learning algorithms to classify emotions based on the unified representation.

In this paper, we propose a novel architecture for Multimodal Emotion Recognition. Our model relies on 3 modalities, Audio, Vision, and Text (A+V+T), to identify the emotion in the given content. The audio features are extracted using Contrastive Language-Audio Pretraining (CLAP) [3]. The Video features are extracted using Video Masked Auto Encoder (VideoMAE) [4], while the text is extracted using MPNet [5]. These features are combined or fused with fusion techniques such as Concatenation, TensorFusion [6], and Low-Rank Multimodal Fusion [7]. The final stage involves passing the joint representation through a classifier to identify the emotion in the given content.

This work aims to provide an overview of the state-of-the-art in MML for emotion recognition and identify the challenges and opportunities in this field. Specifically, we aim to address the following: different modalities used in MML for emotion recognition and their contribution to improving the accuracy of emotion recognition systems, and a comparison of existing approaches for feature extraction, fusion, and classification in MML for emotion recognition.

The rest of the paper is organized as follows. Section

II summarizes the previous research work in Unimodal and Multimodal Emotion Recognition. Section III defines feature encoders for the different modalities used in MML for emotion recognition - Verbal, Vocal, and Visual. The approaches for fusing the modalities are discussed in Section IV. Methodology is described in Section V. Section VI introduces the dataset and implementation details of the proposed architecture. The results and findings are discussed in Section VII. Finally, in section VIII, we conclude the paper and provide directions for future research.

II. Related Work

Emotion Recognition of a particular content can be done in two ways - using one of the modalities (Unimodal), e.g. text or audio, or by taking and combining information from multiple modalities (Multimodal).

A. Unimodal

In recent years, text-based approaches for emotion recognition have gained attention, primarily due to the emergence of the Transformer model [8]. Li et al. [9], [10] have explored the use of BERT [11] and Transformers in ERC by encoding sentences and dialogs, respectively. Jiangnan et al. [12] extended this approach by designing three types of masks to capture different dependencies in a conversation. These masks learn the conventional context, Inter- and Intra-Speaker dependency.

In order to model interactions between interlocutors, Ghosal et al. [13] incorporated commonsense components including mental states, events, and causal relationships. Authors in [14] and [15] have also employed graph neural networks (GNNs) to encode inter-utterance and inter-speaker relationships in ERC. By adding speaker names to utterances and putting separation tokens between them, Kim et al. [16] suggested a straightforward method for modeling contextual information.

For the purpose of creating contextualized utterance representations, Wang et al. [17] employed LSTM-based encoders to capture inter- and intra-speaker dependencies. A Directed Acyclic Graph (DAG)-based ERC model was presented by Shen et al. in [18] that combines the advantages of recurrence-based and graph-based neural networks. A transformer-based approach was put up by Zhu et al. [19] to predict emotion labels by fusing common sense and current events. The EmotionFlow model was created by Song et al. [20] to encode user utterances by concatenating them with another query and using a random field to gather sequential information at the emotional level.

B. Multimodal

Co-attention layers play an important role in fusing the modalities [22]. [21] proposes HCAM, a Hierarchical Cross Attention Model for Multi-modal Emotion Recognition that leverages co-attention layers and are trained hierarchically. Previous research on incorporating contextual information from previous utterances has established the baseline for analyzing dyadic conversations. In some studies [23], [24], previous utterances from both parties are used along with contextual information to predict the emotional state of a given utterance. Majumder et al. [25] extend this work by modelling the uni-modal contextual information separately and then fusing tri-modal features hierarchically to obtain a more comprehensive feature representation of the utterance. DialogueRNN [26] treats the contextual information of each speaker and the global state as distinct entities and uses emotional context from both sources to make accurate predictions. Zhang et al. [27] introduce the ConGCN model, which uses Graph Convolution Networks to model Speaker-Utterance and Utterance-Utterance relationships concurrently in a single network by processing both audio and text utterance features. Multi-modal emotional behaviours are examined by Mao et al. [28] from both intra- and inter-modal aspects. Multi-head attention-based fusion [8] is used by the CMU-Mosei approaches, such as those of Loshchilov et al. [29] and Tsai et al. [30], to recognize emotions in a multi-modal way.

Most of the previous studies have not considered facial features that play a significant role in determining the emotional context of a conversation. These studies use frames as a whole entity without extracting essential parts, such as the face. Furthermore, most of these studies lack an active fusion strategy other than simple concatenation to exploit the abundance of information present in visual and acoustic data.

III. Modality Encoders

Video Masked AutoEncoder, Contrastive Language-Audio Pretraining, and Masked and Permuted Net are used in this research to encode Video, Audio, and Text respectively.

A. Video Masked AutoEncoder

Video Masked AutoEncoder or VideoMAE [4] is a self-supervised video pre-training method that utilizes masked autoencoders [31] to efficiently learn effective video representations without needing large-scale supervised datasets.

Masked autoencoders [31] are a type of neural network trained to reconstruct its input data after masking some parts. The masked areas can be randomly selected or follow a specific pattern, and the goal is for the network to learn how to fill in these missing pieces accurately. This technique has been used in self-supervised learning tasks, where the model learns from unlabeled data without explicit supervision signals.

The method achieves impressive results on small datasets by using customized video tube masking with an extremely high ratio, which encourages more effective video representation extraction during the pre-training process. VideoMAE [4] also demonstrates that data quality is more important than data quantity for self-supervised video pre-training and that domain shift between pre-training and target datasets is crucial.

B. Contrastive Language-Audio Pretraining

Contrastive Language-Audio Pretraining (CLAP) [3] is a pipeline for developing an audio representation by combining audio data with natural language descriptions. This approach involves constructing a contrastive language-audio pretraining model that incorporates feature fusion mechanisms and keyword-to-caption augmentation to enable the model to process audio inputs of variable lengths and enhance performance.

Contrastive refers to a type of learning where the model learns by contrasting similar and dissimilar examples. In other words, it tries to learn representations that make similar examples more alike and dissimilar ones less alike. The model is trained using a contrastive learning approach that compares the embeddings of audio and text, and the loss function is:

$$L = \frac{1}{2N} \sum_{i=1}^{N} \left(\log \frac{\exp \left(E_i^a \cdot E_i^t / \tau \right)}{\sum_{j=1}^{N} \exp \left(E_i^a \cdot E_j^t / \tau \right)} + \log \frac{\exp \left(E_i^t \cdot E_i^a / \tau \right)}{\sum_{j=1}^{N} \exp \left(E_i^t \cdot E_j^a / \tau \right)} \right) \tag{1}$$

where,

$$E_i^a = MLP_{\text{audio}} \left(f_{\text{audio}} \left(X_i^a \right) \right) \tag{2}$$

$$E_i^t = MLP_{\text{text}} \left(f_{\text{text}} \left(X_i^t \right) \right) \tag{3}$$

where (X_i^a, X_i^t) is one of the i-indexed audio-text pairings. The audio encoder $f_{audio}(\cdot)$ and the text encoder $f_{text}(\cdot)$, both with projection layers, produce the audio embedding E_i^a and the text embedding E_i^t, respectively.

The model is evaluated across several tasks, including text-to-audio retrieval, zero-shot audio classification, and supervised audio classification, and has shown superior performance in text-to-audio retrieval and state-of-the-art performance in zero-shot audio classification [3].

CLAP [3] is related to Contrastive Language-Image Pre-training (CLIP) [32] in that they use contrastive learning as their main training objective. However, while CLIP focuses on images and text, CLAP focuses on audio data with natural language descriptions.

C. Masked and Permuted Net

Masked and Permuted Net (MPNet) [5] is a novel pre-training method for language understanding that combines the advantages of BERT and XLNet while addressing their limitations.

BERT [11] and XLNet [33] are pre-training models for natural language processing. BERT [11] uses masked language modeling (MLM) to predict missing words in a sentence, while XLNet [33] uses permuted language modeling (PLM) to capture the dependency among predicted tokens. MPNet [5] combines the advantages of both models by using permuted language modeling like XLNet but also takes auxiliary position information as input, like BERT, which reduces the position discrepancy between pre-training and fine-tuning.

Experimental results show that MPNet [5] outperforms previous state-of-the-art pre-trained models, such as BERT [11], XLNet [33], and RoBERTa [34], on a variety of downstream NLP tasks.

IV. Multimodal Fusion

Multi-modal data fusion combines data from multiple sources to improve the quality of the information. This can be done by fusing data from different modalities, such as images, text, and audio.

A. Concatenation

One of the simplest methods for fusing modalities is concatenation. The concatenation operation is a simple way to combine two or more vectors. Given two vectors, x, and y, the concatenation operation is defined as follows:

$$z = x \oplus y \tag{4}$$

Where z is the new vector that is formed by concatenating x and y.

Concatenation is a simple and effective method for fusing modalities. However, it can be computationally expensive to concatenate large vectors. And it can lead to information loss. This is because the features from each modality are combined into a single feature vector. This can result in the loss of important information specific to each modality.

B. Tensor Fusion

The Tensor Fusion layer fuses the information from the different modalities by explicitly modeling the inter-modality dynamics. Inter-modality dynamics refer to the relationships between the features from different modalities. For example, in a video, the facial expressions of a person can be used to convey sentiment. The tensor fusion layer models inter-modality dynamics using a 3-fold Cartesian product from the modality embeddings. The 3-fold Cartesian product creates all possible combinations of features from the different modalities.

$$\mathbf{z}^m = \begin{bmatrix} \mathbf{z}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^a \\ 1 \end{bmatrix} \tag{5}$$

Where z^l is text embedding, z^v is video embedding, and z^a represents audio embedding.

Tensor Fusion is a mathematical operation that creates a new tensor by multiplying two tensors. It has no learnable parameters, meaning it is a fixed function. Although the output tensor is highly dimensional, it has been observed that Tensor Fusion does not overfit easily. This is because the output neurons of Tensor Fusion are easy to interpret and semantically meaningful. In other words, the manifold they lie on is not complex but just highly dimensional. This makes it easy for the subsequent network layers to decode the meaningful information.

C. Low-Rank Multimodal Fusion

Low-rank Multimodal Fusion (LMF) is a method for efficiently computing tensor-based multimodal representations with fewer parameters and computational complexity. LMF decomposes the weights into low-rank factors, reducing the number of parameters in the model and enabling linear scaling with the number of modalities. This method calculates by breaking down high-dimensional tensors into smaller matrices that can be multiplied together to reconstruct them.

The computation of the input tensor Z, which is formed by the unimodal representation, can be expressed as follows:

$$\mathcal{Z} = \bigotimes_{m=1}^{M} z_m, z_m \in \mathbb{R}^{d_m} \tag{6}$$

The tensor outer product over a set of vectors indexed by m, denoted by $\bigotimes_{m=1}^{M}$, is performed. Here, z_m represents the input representation with 1s appended to it.

The final fusion step is calculated as follows:

$$h = \left(\sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \right) \cdot \mathcal{Z} \tag{7}$$

Where,

$$\mathbf{w}_m^{(i)} = \left[w_{m,1}^{(i)}, w_{m,2}^{(i)}, \ldots, w_{m,d_h}^{(i)} \right] \tag{8}$$

and $\left\{ \left\{ w_{m,k}^{(i)} \right\}_{m=1}^{M} \right\}_{i=1}^{R}$ are the rank R decomposition factors.

The advantages of this approach are that it reduces computational complexity, requires fewer parameters, enables linear scaling with more modalities, and achieves competitive results on downstream tasks such as sentiment analysis or emotion recognition. However, it may not capture all possible interactions between different modalities due to its low-rank approximation.

V. Methodology

This research paper compares models trained on Concatenation, Tensor Fusion, and Low-Rank Multimodal Fusion for multimodal emotion recognition. The proposed model comprises several subnetworks, including SubNet for video and audio fusion and TextSubNet for text fusion. The MultimodalEmotionClassifier model takes input tensors for audio (audio_x), video (video_x), and text (text_x) data.

The SubNet class defines a subnetwork used for video and audio fusion. The TextSubNet class implements an LSTM-based subnetwork for text fusion. The MultimodalEmotionClassifier class incorporates the SubNet and TextSubNet subnetworks. It also includes post-fusion layers to combine the modalities further. These post-fusion layers consist of linear transformations and dropout operations. The input tensors are passed through the respective subnetworks during the forward pass. The outputs of the subnetworks are then passed through different fusion techniques. These fusion tensors are then processed through the post-fusion layers to obtain the final output.

The MultimodalEmotionClassifier model aims to learn a mapping between the input modalities and a scalar output value. The architecture and fusion strategy allows for capturing the relationships between different modalities, enhancing the overall understanding and performance.

The model is trained using appropriate loss functions and optimization algorithms to minimize the discrepancy between predicted and ground truth values. For training and optimizing the model, the following components are utilized:

1) Loss Function: The criterion used for the MultimodalEmotionClassifier model is the CrossEntropyLoss, which is suitable for multi-class classification tasks. It measures the dissimilarity between predicted and true probability distributions. By quantifying the average information content required to represent the true distribution with the predicted one, cross-entropy loss facilitates efficient model training and parameter optimization. It is calculated using equation 9.

$$\text{CrossEntropyLoss}(y, \hat{y}) = -\sum_{i} y_i \log(\hat{y}_i) \tag{9}$$

where y represents the true probability distribution, $\hat{y}$ denotes the predicted probability distribution, and i iterates over the classes or categories in the distribution.

2) Optimization Algorithm: The Adam optimizer is a popular and effective adaptive optimization algorithm that efficiently updates the model's parameters based on the gradients calculated during backpropagation. It combines the benefits of both adaptive learning rates and momentum methods. By maintaining adaptive learning rates for individual model parameters and incorporating momentum, Adam enhances convergence speed and stability during training. Its adaptive nature makes it well-suited for a wide range of neural network architectures and improves generalization performance.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t \tag{10}$$

where,

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{11}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{12}$$

and

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{13}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{14}$$

m_t and v_t represent the first and second moment estimates respectively. β_1 and β_2 are the exponential decay rates for the moment estimates. g_t refers to the gradient at time step t, and θ_t denotes the parameter values at time step t. η is the learning rate, and ϵ is a small constant to avoid division by zero.

VI. Implementation

Our proposed model is trained on Multimodal EmotionLines Dataset (MELD) [35]. The dataset was extracted from the popular TV show Friends, with over 13,000 utterances, and is an extension of EmotionLines Dataset [35], [36]. It has a diverse range of emotions - Anger, Disgust, Fear, Joy, Neutral, Sadness, and Surprise. Data collection involved capturing dialogues encompassing textual, audio, and visual modalities. MELD provides rich contextual information along with speaker annotations. Its comprehensive data extraction

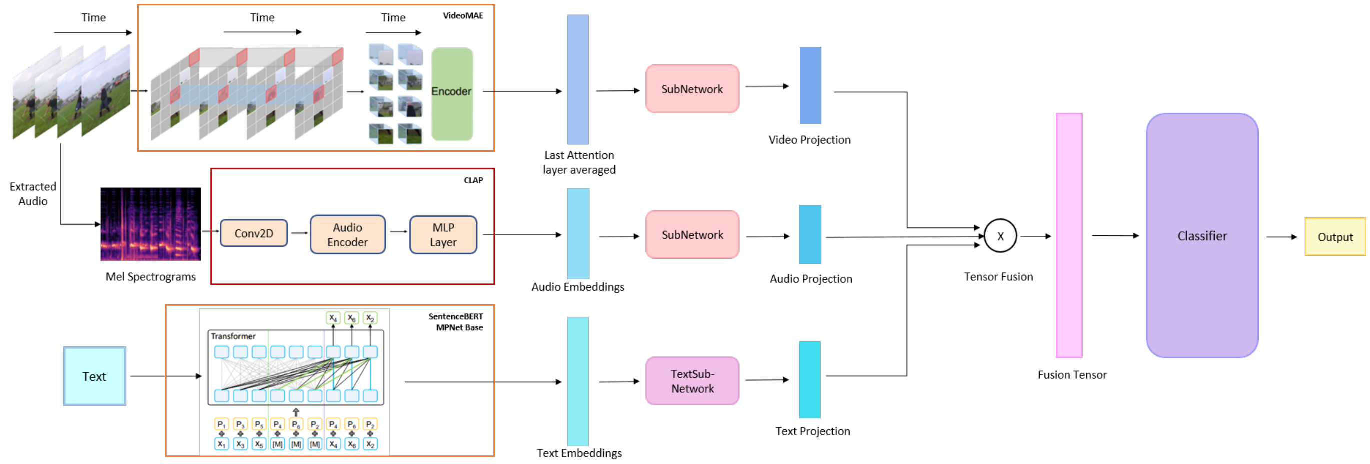

Fig. 1. Architecture of the proposed model.

methodology makes it an ideal choice for developing and evaluating novel approaches for multimodal fusion.

The training dataset has 9989 utterances, validation has 1109 utterances and the test has 2610 utterances. Number of utterances per emotion and split is shown in table I.

TABLE I
MELD Dataset Composition by Emotion

Emotion	Train	Validation	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

The train, validation and test data are passed through CLAP [3], VideoMAE [4], and MPNet [5] for audio, video and text, respectively.

The audio is extracted from the video with the help of moviepy library. CLAP takes audio of 64-dimensional Mel spectograms. The Fourier transform is computed on windows of 1024 samples. The lowest and the highest frequency of interest are 50 and 14000, meaning the short-time Fourier transform will not be calculated for values outside this range. The extracted Mel spectrograms are then passed through CLAPModel, which projects the audio into 512 dimensional vector.

The video is encoded using VideoMAE. All the frames of the videos for the model are resized to 224x224 and normalized with image mean: 0.485, 0.456, 0.406 and image std: 0.229, 0.224, 0.225 per channel. These are then passed through VideoMAE to extract video feature vectors of 768 dimension.

MPNet is used for encoding the text or utterances. All the words are lowercased and tokenized with the help of BertTokenizer. The max length of the utterance is 512 words. Tokenized words are then passed through MPNet to get embedding size of (512, 768). The embeddings are then normalized on the first dimension to get the final text embedding size of 768 dimensions.

All these encoded modalities are then passed into their respective Sub-Networks that converts the raw input vector into lower dimensional vectors. Audio and Video embeddings are processed with SubNet class that has 3 linear layers, and outputs 128-dimensional vector. On the other hand, Text embeddings are processed with TextSubNet, which has an LSTM layer followed by a linear layer. The output of this subnetwork is a 32-dimensional vector.

Finally, the fusion is computed using Tensor Fusion to get a dense representation combining information across all three modalities. The fusion is followed by 3 linear layers, the final layer outputs a vector of 7 dimensions indicating the probability of each emotion - Anger, Disgust, Fear, Joy, Neutral, Sadness, and Surprise. The whole architecture is shown in fig 1.

VII. Results and Discussion

Table II compares the result of different fusion techniques. Loss, accuracy and F1 score are used to compare the results of the model. Higher accuracy and F1 score are better, while the model is better with lower loss.

TABLE II
Metrics of Model on Test Set

Fusion	Loss	Accuracy	F1 Score
Concatenation	1.9164	47.6	54.5
Concat with Prj.	0.0011	57.2	60.7
Tensor Fusion	0.0008	59.1	62.3
LMF	0.0007	58.0	61.9

The baseline model, which implements simple concatenation of modality embeddings without any projection (subnetwork) and fusion performs the worst. This is in line with the reasoning that the model fails to capture and merge information from different modalities with simple concatenation. However, after adding the projection network, the model's performance jumps by over 10%. The main work of projection layers is to get intra-modality representations. This means that the feature vectors are denser and more rich in information. Tensor Fusion performs best, followed by Low-Rank Multimodal Fusion.

VIII. Conclusion

In this study, the objective was to compare different modality fusion methods and propose a novel framework for Multimodal Emotion Recognition. The task was to recognize emotion in one of the 7 classes - Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear from a multimodal video that has verbal, vocal and visual modalities. We compared the effect of fusion techniques such as Concatenation, Tensor Fusion, and Low-Rank Multimodal Fusion and presented a new architecture that takes audio, video, and text to recognize emotion. The embeddings are extracted using CLAP, VideoMAE, and MPNet for audio, video, and text respectively. The results in this paper suggest more work is required in Affective Computing for Multimodal Emotion Recognition. Future work could involve emotion recognition in conversation that takes previous utterances into consideration. Another research direction could involve finding better fusion methods. Explainable AI methods could be added to explain the prediction with natural language generation or a heatmap.

References

[1] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi, 'COGMEN: COntextualized GNN based multimodal emotion recognitioN', arXiv preprint arXiv:2205. 02455, 2022.

[2] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, 'M2FNet: multi-modal fusion network for emotion recognition in conversation', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4652–4661.

[3] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, 'Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation', in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[4] Z. Tong, Y. Song, J. Wang, and L. Wang, 'Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training', arXiv preprint arXiv:2203. 12602, 2022.

[5] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, 'Mpnet: Masked and permuted pre-training for language understanding', Advances in Neural Information Processing Systems, vol. 33, pp. 16857–16867, 2020.

[6] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, 'Tensor fusion network for multimodal sentiment analysis', arXiv preprint arXiv:1707. 07250, 2017.

[7] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, 'Efficient low-rank multimodal fusion with modality-specific factors', arXiv preprint arXiv:1806. 00064, 2018.

[8] A. Vaswani et al., 'Attention is all you need', Advances in neural information processing systems, vol. 30, 2017.

[9] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, 'Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations', in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4190–4200.

[10] J. Li, M. Zhang, D. Ji, and Y. Liu, 'Multi-task learning with auxiliary speaker identification for conversational emotion recognition', arXiv preprint arXiv:2003. 01478, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810. 04805, 2018.

[12] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, 'A hierarchical transformer with speaker modeling for emotion recognition in conversation', arXiv preprint arXiv:2012. 14781, 2020.

[13] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, 'Cosmic: Commonsense knowledge for emotion identification in conversations', arXiv preprint arXiv:2010. 02795, 2020.

[14] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, 'Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation', arXiv preprint arXiv:1908. 11540, 2019.

[15] D. Sheng, D. Wang, Y. Shen, H. Zheng, and H. Liu, 'Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition', in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4153–4163.

[16] T. Kim and P. Vossen, 'Emoberta: Speaker-aware emotion recognition in conversation with roberta', arXiv preprint arXiv:2108. 12009, 2021.

[17] Y. Wang, J. Zhang, J. Ma, S. Wang, and J. Xiao, 'Contextualized emotion recognition in conversation as sequence tagging', in Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue, 2020, pp. 186–195.

[18] W. Shen, S. Wu, Y. Yang, and X. Quan, 'Directed acyclic graph network for conversational emotion recognition', arXiv preprint arXiv:2105. 12907, 2021.

[19] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, 'Topic-driven and knowledge-aware transformer for dialogue emotion detection', arXiv preprint arXiv:2106. 01071, 2021.

[20] X. Song, L. Zang, R. Zhang, S. Hu, and L. Huang, 'Emotionflow: Capture the dialogue level emotion transitions', in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8542–8546.

[21] S. Dutta and S. Ganapathy, 'HCAM–Hierarchical Cross Attention Model for Multi-modal Emotion Recognition', arXiv preprint arXiv:2304. 06910, 2023.

[22] A. Bobde, S. Tawari, V. Narnaware, A. Suryawanshi, and P. Kshirsagar, 'Countering Dis-information by Multimodal Entailment via Joint Embedding', in 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), 2023, pp. 1060–1065.

[23] 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. Association for Computational Linguistics. North American Chapter. Meeting, 2018, vol. 2018, p. 2122.

[24] 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 (volume 1: Long papers), 2017, pp. 873–883.

[25] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, 'Multimodal sentiment analysis using hierarchical fusion with context modeling', Knowledge-based systems, vol. 161, pp. 124–133, 2018.

[26] 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, vol. 33, pp. 6818–6825.

[27] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, 'Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations', in IJCAI, 2019, pp. 5415–5421.

[28] Y. Mao et al., 'Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation', arXiv preprint arXiv:2010. 07637, 2020.

[29] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, 'A transformer-based joint-encoding for emotion recognition and sentiment analysis', arXiv preprint arXiv:2006. 15955, 2020.

[30] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, 'Multimodal transformer for unaligned multimodal language sequences', in Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019, vol. 2019, p. 6558.

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, 'Masked autoencoders are scalable vision learners', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[32] Y. Li et al., 'Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm', arXiv preprint arXiv:2110. 05208, 2021.

[33] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, ‘Xlnet: Generalized autoregressive pretraining for language understanding’, Advances in neural information processing systems, vol. 32, 2019.

[34] Y. Liu et al., ‘Roberta: A robustly optimized bert pretraining approach’, arXiv preprint arXiv:1907. 11692, 2019.

[35] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, ‘Meld: A multimodal multi-party dataset for emotion recognition in conversations’, arXiv preprint arXiv:1810. 02508, 2018.

[36] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku, and Others, ‘Emotionlines: An emotion corpus of multi-party conversations’, arXiv preprint arXiv:1802. 08379, 2018.