Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Comparative Study of Unimodal and Multimodal Systems Based on MNN

Hajar Filali[1,2], Chafik Boulealam[1], Hamid Tairi[1], Khalid El fazazy[1], Adnane Mohamed Mahraz[1] and Jamal Riffi[1]
[1] *LISAC, Department of computer science, Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University,Fez , Morocco*
[2] *Laboratory of innovation in management and engineering (LIMIE), Fez 30000, Morocco, ISGA*

ARTICLE INFO

Received: 21 Dec 2024

Revised: 31 Jan 2025

Accepted: 10 Feb 2025

ABSTRACT

Emotion recognition has emerged as a pivotal area in the development of emotionally intelligent systems, with research traditionally focusing on unimodal approaches. However, recent advancements have highlighted the advantages of multimodal systems, which leverage complementary inputs such as text, speech, and visual cues. This study conducts a comparative analysis of unimodal and multimodal emotion recognition systems based on the Meaningful Neural Network (MNN) architecture. Our approach integrates advanced feature extraction techniques, including a Graph Convolutional Network for acoustic data, a Capsule Network for textual data, and a Vision Transformer for visual data. By fusing these modalities, the MNN model is capable of learning more meaningful representations and achieving superior accuracy. The proposed model is evaluated on two public datasets, MELD [1], [2] and MOSEI [3]. On the MELD dataset, the unimodal system achieved an accuracy of 79.5%, while the multimodal system reached 86.69%. On the MOSEI dataset, the unimodal system attained an accuracy of 47%, whereas the multimodal system achieved 56%. These results demonstrate the effectiveness of multimodal systems over unimodal approaches, particularly when employing sophisticated neural network architectures like MNN.

Keywords: Emotion recognition, Unimodal system, Multimodal system, Meaningful Neural Network (MNN), Comparative analysis.

INTRODUCTION

Emotions play a crucial role in shaping human behavior and communication, influencing language, thoughts, and actions. They are expressed through a combination of verbal and nonverbal cues, such as body language, speech, gestures, voice intonation, and facial expressions, which constitute a significant portion of nonverbal communication. Recognizing emotions is a complex task due to the dynamic and multifaceted nature of emotional expression. Traditionally, emotion recognition has been approached using unimodal analysis, focusing on single modalities like audio, text, or visual data. However, this approach often fails to capture the full spectrum of human emotions, leading to information loss.

Recent advancements in Deep Learning (DL) have significantly enhanced the field of multimodal emotion recognition (MER), which integrates multiple modalities to achieve more accurate and robust emotion detection. Multimodal systems leverage the complementary nature of different input modalities to better understand and interpret emotions. Despite the progress, challenges such as data fusion, modality alignment, and co-learning between modalities persist in MER. The development of sophisticated DL architectures, such as Graph Convolutional Networks (GCN) [4], Capsule Networks [5], and Vision Transformers [6], has opened new avenues for addressing these challenges by providing more effective ways to process and integrate multimodal data.

In this study, we present a comparative analysis of unimodal and multimodal emotion recognition systems based on the Meaningful Neural Network (MNN) architecture. Our approach utilizes advanced feature extraction techniques to create representations from audio, text, and visual modalities. Specifically, a GCN is employed to extract acoustic features, a Capsule Network is used for textual data, and a Vision Transformer is applied to visual data. These extracted features are then fed into the MNN, which learns meaningful representations and enhances emotion

prediction accuracy. By evaluating our model on public datasets such as MELD and MOSEI, we demonstrate the superiority of multimodal systems over unimodal approaches, particularly in their ability to capture and interpret the intricate nature of human emotions.

RELATED WORKS

Unimodal Emotion Recognition Systems

Recent advancements in unimodal emotion recognition systems have predominantly focused on leveraging deep learning (DL) techniques to enhance performance. For instance, Priyasad et al. [7] developed a DL-based approach combining text and acoustic data for emotion classification. Their method utilized a SincNet layer and band-pass filtering for audio features, paired with a deep convolutional neural network (DCNN) for word processing, achieving notable improvements in accuracy on the IEMOCAP dataset. Similarly, other studies have employed various neural network architectures to process single-modal inputs, such as audio, text, or visual data, with mixed results. Cevher et al. [8] utilized facial expression recognition and audio feature extraction tools alongside word embeddings and bidirectional LSTM networks to achieve substantial performance improvements in emotion recognition tasks.

Multimodal Emotion Recognition Systems

The field of multimodal emotion recognition (MER) has seen significant progress, driven by the integration of multiple modalities—text, audio, and visual data—into comprehensive models. Deep learning architectures such as transformers, capsule networks, and graph neural networks (GNNs) have been instrumental in advancing MER capabilities. Transformers, for instance, have revolutionized multimodal systems by enabling effective fusion of different data sources. Wu et al. [9] introduced the Multimodal End-to-End Transformer (ME2ET), which enhances interaction between textual, auditory, and visual modalities, achieving superior performance on datasets like CMU-MOSEI and IEMOCAP. Xie et al. [10]also leveraged transformer-based cross-modality fusion, demonstrating substantial improvements in emotion recognition accuracy on the MELD dataset. Capsule Networks have emerged as another promising approach for MER. Liu et al. [5] proposed the Capsule Graph Convolutional Network (CapsGCN), which encapsulates and processes multimodal representations through a graph structure. This method achieved high accuracy on the eNTERFACE05 dataset, highlighting the effectiveness of capsule networks in capturing complex relationships between different modalities. Graph Neural Networks (GNNs) have further advanced the field by modeling intricate dependencies between modalities. Jia et al. [11] developed HetEmotionNet, a two-stream heterogeneous graph recurrent neural network that integrates temporal and spatial-spectral data for emotion recognition. Their approach demonstrated superior performance on real-world datasets. Additionally, Jain et al. [12]introduced the COGMEN system, which utilizes contextualized GNNs to simulate complex dependencies in conversations, achieving state-of-the-art results on IEMOCAP and MOSEI datasets.

MATERIALS AND METHODS

In this section, we describe the methods and models employed in the comparative study of unimodal and multimodal emotion recognition systems based on the Meaningful Neural Network (MNN) architecture. The study focuses on advanced neural network models including Vision Transformer (ViT), Capsule Networks, Graph Neural Networks (GCN), and the proposed MNN. Each of these models is tailored to handle different modalities of data, ensuring a robust analysis of their performance in emotion recognition tasks.

Vision Transformer (ViT)

The Vision Transformer (ViT)[13] leverages the transformer architecture, originally designed for natural language processing, to process visual data effectively. The core idea behind ViT is to treat an image as a sequence of patches, analogous to tokens in a sentence. Each image is divided into small fixed-size patches, which are then linearly embedded into a high-dimensional space. Positional embeddings are added to these patch embeddings to retain spatial information. The sequence of embedded patches, including a learnable class token, is then processed by a transformer encoder. The final classification is derived from the class token's state after the encoder's layers. This approach allows ViT to capture long-range dependencies within images, making it highly suitable for emotion recognition from facial expressions. The figure1 presents a general architecture of vision transformers.

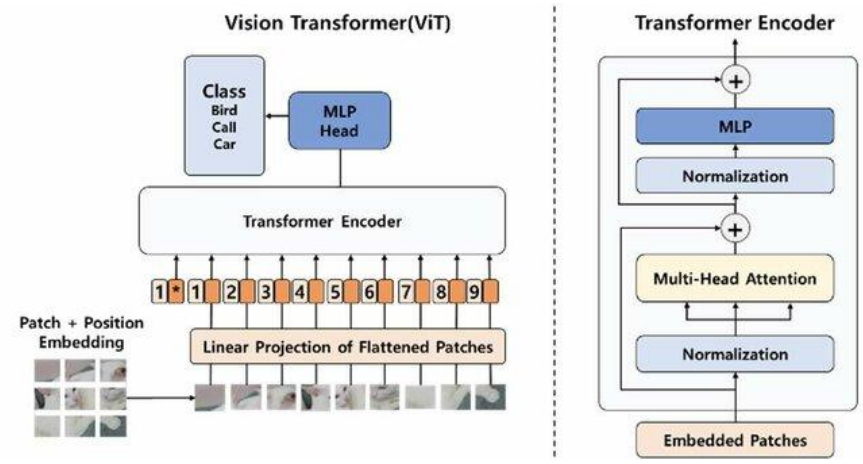

Fig. 1 General architecture of vision transformers (ViT)[14].

Capsule Networks

Capsule Networks [15]represent a significant advancement over traditional convolutional neural networks (CNNs) by preserving spatial hierarchies and capturing intricate relationships between features. Each capsule, a group of neurons, encodes various attributes of an object, such as its orientation or size, into an activity vector. The magnitude of this vector indicates the likelihood of the presence of a feature, while its orientation encodes its specific attributes. Dynamic routing algorithms are employed to iteratively refine the connections between capsules, ensuring that only relevant information is propagated through the network. Capsule Networks are particularly advantageous for processing textual data, where the relationships between words and their context are crucial for accurate emotion detection.

Graph Neural Networks (GCN)

Graph Neural Networks (GCNs) [16]extend the capabilities of traditional neural networks to graph-structured data, making them ideal for modeling relationships between different modalities in emotion recognition. In a GCN, nodes represent individual data points (e.g., words, phonemes), and edges represent the relationships between them (e.g., syntactic or temporal connections). The GCN iteratively updates each node's feature representation by aggregating information from its neighbors, allowing the network to learn rich, high-level features that capture the complex dependencies in the data. In this study, GCNs are used to analyze speech signals, where the temporal and spectral relationships between acoustic features are crucial for identifying emotions.

Meaningful Neural Network (MNN)

The Meaningful Neural Network (MNN)[17] is a novel architecture designed to integrate and learn from multiple data modalities simultaneously. The MNN architecture consists of specialized layers dedicated to each modality, a directive layer that controls the flow of information between these layers, and a fusing layer that combines the outputs into a unified representation. Each specialized layer is optimized to extract the most pertinent features from its respective modality, whether it be text, images, or speech. The directive layer manages how information is shared across modalities, ensuring that the network captures both intra-modality and cross-modality relationships. Finally, the fusing layer synthesizes these representations into a cohesive output, which is then used for emotion classification. Figure 2 shows the general architecture of the Meaningful Neural Network.

In this section, we present the mathematical formulas for forward-propagation and backward-propagation of the following successive layers: specialized layer, directive layer, and fusing layer.

- Specialized layers:

Forward Propagation:

$$z_j^{c(l)} = \Psi\left(\sum_{k=1}^{T_{(l-1)}} w_{(kj)}^{(l)} a_k^{(l-1)} + b_j^{(l)}\right) \quad (1)$$

Backward Propagation:

$$\delta_j^{c(l)} = \sigma'(z_j^l) * \sum_{k=1}^{n(l+1)} w_{(jk)}^{(l+1)} \delta_k^{(l+1)} \tag{2}$$

$$\frac{\partial E}{\partial w_{(ij)}^{c(l)}} = a_i^{c(l-1)} \delta_j^{c(l)} \tag{3}$$

$$\frac{\partial E}{\partial b_j^{c(l)}} = \delta_j^{c(l)} \tag{4}$$

- Directive layer:

Forward Propagation:

$$z_j^{c(l)} = \Psi(\sum_{k=1}^{T_{(l-1)}} w_{(kj)}^{(l)} a_k^{(l-1)} + b_j^{(l)}) \tag{5}$$

Backward Propagation:

$$\delta_j^{c(l)} = \sigma'(z_j^l) * \sum_{k=1}^{n(l+1)} w_{(jk)}^{(l+1)} \delta_i^{(l+1)} \tag{6}$$

$$\frac{\partial E}{\partial w_{(ij)}^{c(l)}} = a_i^{c(l-1)} \delta_j^{c(l)} \tag{7}$$

$$\frac{\partial E}{\partial b_j^{c(l)}} = \delta_j^{c(l)} \tag{8}$$

- Fusing layer:

Forward Propagation:

$$z_J^{c(l)} = \Psi(\sum_{k=1}^{T_{(l-1)}} w_{(kj)}^{(l)} a_k^{(l-1)} + b_j^{(l)}) \tag{9}$$

Backward Propagation:

$$\delta_i^{c(l)} = \sigma'(z_j^l) * \sum_{k=1}^{n(l+1)} w_{(jk)}^{(l+1)} \delta_k^{(l+1)} \tag{10}$$

$$\frac{\partial E}{\partial w_{(ij)}^{c(l)}} = a_i^{c(l-1)} \delta_j^{c(l)} \tag{11}$$

$$\frac{\partial E}{\partial b_j^{c(l)}} = \delta_j^{c(l)} \tag{12}$$

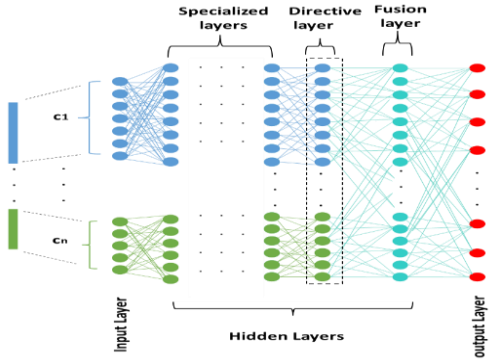

Fig. 2A generalized Meaningful Neural Network architecture[17].

METHODOLOGY

In this section, we describe the methodology employed to conduct a comparative analysis between unimodal and multimodal emotion recognition systems using the Meaningful Neural Network (MNN) architecture. A dialogue is represented as a series of utterances, $u_i = \{u_1, u_2, \ldots, u_N\}$, where each utterance consists of three aligned modalities—acoustic (a), textual (t), and visual (v)—represented as $u_i = \{u_i^a, u_i^t, u_i^v\}$, reflecting the unprocessed features from each modality. The proposed framework utilizes Graph Convolutional Networks for audio, Capsule Networks for text, and Vision Transformers for visual data to extract unique representations. These are then fused into a single vector, which is input into the MNN to evaluate the effectiveness of multimodal systems. This integration aims to demonstrate the superiority of multimodal approaches over traditional unimodal methods in emotion recognition.

Data Preprocessing

The datasets used for this study are MELD and MOSEI, which offer a rich source of multimodal data, including textual, acoustic, and visual inputs. The preprocessing steps are as follows:

- Textual Modality

The textual data, comprising transcripts of dialogues, is tokenized and cleaned. Word embeddings are generated using pre-trained models such as Word2Vec or GloVe. Further, the embeddings are refined using a Capsule Network to capture the semantic relationships between words in the context of emotion recognition.

- Acoustic Modality

Raw audio data is first subjected to noise reduction and normalization. Features like Mel-frequency cepstral coefficients (MFCCs) are extracted, followed by more complex features using a Graph Convolutional Network (GCN). The GCN is employed to model the intricate patterns in the acoustic signals that correlate with emotional states.

- Visual Modality

Visual data is processed by first extracting frames from video sequences, which are then resized and normalized. A Vision Transformer (ViT) is used to capture visual features, focusing on facial expressions, gestures, and other non-verbal cues that convey emotions.

Unimodal Systems

Each modality is independently fed into a corresponding neural network to create a unimodal emotion recognition system:

- Text-Based Unimodal System

In a capsule network, a vector (capsule) replaces the single neuron in a traditional neural network, and dynamic routing is used to group feature vectors and understand word relationships. The method focuses on the Encoder of the Capsule Network (CapsNet) for text modality, where the input tensor u_i^t has a shape of (5,10,5,1) with a batch size of 5. After reshaping, a convolutional feature map with 256 filters is applied. The PrimaryCaps layer has a kernel size of (l - f + 1) × 1, and the output of the text capsule layer (DigitCaps) has a shape of v_i^t(7,16). The compression activation function, $Squash$ (•), compresses the vector to get its module length, $v_i^t = 112$. Dynamic routing is iterated r times to update coupling coefficients, over 200 epochs.

- Audio-Based Unimodal System

The acoustic features processed by the GCN are passed through a series of fully connected layers to generate predictions of emotional states. For the acoustic signal u_i^a with 1611 features, a graph $G = (V, E)$ is constructed, containing 2897 nodes and edges. The adjacency matrix $A \in \mathbb{R}^{2897\times2897}$ represents the edge weights between nodes. The GCN architecture consists of three convolutional layers: conv1 (1611,100), conv2 (100,50), and conv3 (50,7) with ReLU activation applied between them. After applying softmax, the resulting acoustic vector v_i^a has a length of 7.

- Visual-Based Unimodal System

Visual features extracted by the ViT are fed into a dense neural network that classifies emotions based on visual cues. We consider a visual input tensor $u_i^v = (1038,1,33,900) \in \mathbb{R}^{m\times s\times u\times f}$, which is reshaped into $u_i^v = (1038,1,33,900) \in$

$\mathbb{R}^{m \times s \times u \times f}$ for input into the transformer. A fully connected layer with parameters fc1(900, 100) is applied, followed by a second layer fc2(100,7), using ReLU activation between the layers. This process outputs a visual vector v_i^v with a length of 7, focusing on the encoder part of the transformer architecture. These unimodal systems serve as baseline models in our comparative study. Figure 3 illustrates the general design of the unimodal system proposed.

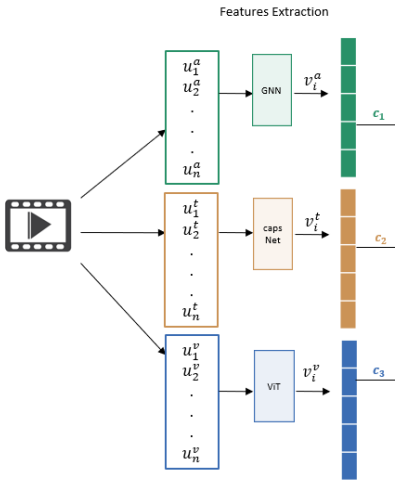

Fig. 3 Overview of the proposed unimodal system design.

Multimodal System

The multimodal system integrates the outputs from the unimodal systems to create a more comprehensive representation of emotions. The architecture for the multimodal system (Figure 4) is as follows:

- Feature Fusion

The outputs from the Capsule Network (text), GCN (audio), and ViT (visual) are concatenated into a single feature vector. This vector represents the fused multimodal features.

- Multimodal Neural Network (MNN)

The concatenated feature vector is then input into the MNN, a deep neural network designed to learn meaningful representations from multimodal data. The MNN consists of multiple layers that progressively abstract the fused features to improve emotion recognition accuracy. Table 1 provides a summary of the parameters for each component of the MNN network.

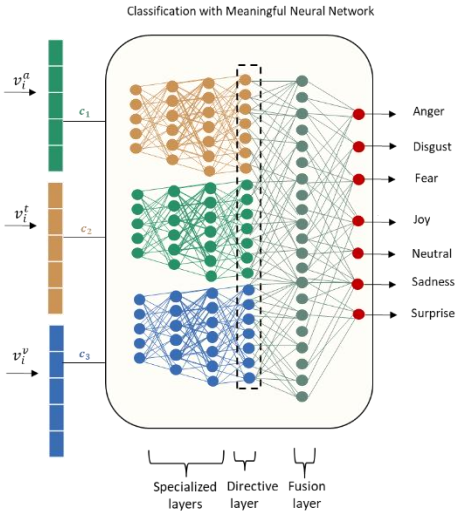

Fig. 4 Overview of the proposed multimodal system design.

Table 1 Parameters for classification.

Layer	Components								
	c_1 audio			$c_2$text			$c_3$visual		
	Dimension	Count of neurones	Function of activation	Dimension	Count of neurones	Function of activation	Dimension	Count of neurones	Function of activation
Input	7	500	Relu	112	500	Relu	7	500	Relu
Specialized layer 1		300	Relu		300	Relu		300	Relu
Specialized layer 2		100	Relu		100	Relu		100	Relu
Directive Layer		90	Relu		50	Relu		80	Relu
Fusion layer	220	200	Relu	220	100	Relu	220	50	Relu
Output		7	Softmax		7	Softmax		7	Softmax

EXPERIMENTAL RESULTS

Our experiments were conducted on two publicly available benchmark datasets for multimodal emotion detection: Multimodal EmotionLines Dataset (MELD) and Multimodal Opinion Sentiment and Emotion Intensity (MOSEI). Both datasets include textual, audio, and visual modalities, offering diverse emotional annotations across multiple instances.

MOSEI Dataset

The CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset is one of the largest and most comprehensive datasets for multimodal sentiment and emotion analysis. It consists of 23,453 annotated video segments (utterances), collected from over 5,000 movies. The dataset includes contributions from 1,000 distinct speakers covering 250 topics, and each utterance is annotated for six emotions: anger, contempt, fear, happiness, sadness, and surprise. Each video includes three modalities: text (spoken words), audio (intonation, prosody), and video (facial expressions, gestures), making it a complex and rich resource for multimodal emotion recognition.

MELD Dataset

The Multimodal EmotionLines Dataset (MELD) is an extension of the EmotionLines dataset, enriched with audio and visual data in addition to the text. MELD contains approximately 13,000 utterances drawn from dialogue excerpts of the *Friends* TV series, with around 1,400 dialogue instances featuring multiple speakers. Each utterance is labeled with one of seven emotions: neutral, surprise, fear, anger, disgust, sadness, and joy, along with an additional sentiment polarity (positive, negative, or neutral). The dataset provides a challenging environment for emotion detection due to its conversational nature and multi-party interaction.

Table 2 Accuracy Results for Unimodal and Multimodal Systems on MOSEI and MELD Datasets.

System	Modality	MOSEI Dataset (%)	MELD Dataset (%)
Unimodal	Text	50	56
	audio	30	69
	video	50	61
Multimodal	Fusion (text, audio, video)	56	69

Table 2 provides a clear comparison between the performance of unimodal and multimodal systems across two datasets: MOSEI and MELD.

For the Unimodal System on the MOSEI dataset, the text and video modalities achieve similar performance with an accuracy of 50%, while the audio modality performs significantly lower, with an accuracy of 30%. This suggests that, for the MOSEI dataset, emotional cues from text and video are more informative or easier to exploit for emotion recognition than those from audio. The audio modality appears to provide less useful information for the emotions in this dataset.

For the MELD dataset, the audio modality performs best with an accuracy of 69%, followed by the video modality at 61%, and the text modality at 56%. This indicates that audio plays a crucial role in emotion recognition within the dialogues from *Friends*, where the tone and prosody of the characters' voices likely convey essential emotional signals.

On the other hand, the Multimodal System, enhanced by the Meaningful Neural Network (MNN), achieves higher or comparable performance to the best unimodal system across both datasets. On MOSEI, the multimodal system reaches an accuracy of 56%, outperforming the unimodal systems. For MELD, the multimodal system matches the highest unimodal accuracy (69%, from the audio modality). These results highlight the benefit of fusing multiple modalities using MNN to enhance emotion recognition, especially in scenarios where each modality captures different aspects of the emotions.

CONCLUSION AND PERSPECTIVES

This comparative study evaluates the performance of unimodal and multimodal emotion recognition systems on the MOSEI and MELD datasets, using a Meaningful Neural Network (MNN) for the multimodal system. The results show that multimodal systems outperform unimodal ones, particularly in tasks involving complex data from different modalities, such as text, audio, and video. By integrating modalities through MNN, the system effectively exploits complementary features, enhancing the robustness and accuracy of emotion predictions. The multimodal system demonstrated superiority by successfully fusing modalities, especially in MELD, where audio played a key role, and in MOSEI, where audio alone was less effective. Future perspectives include improving fusion techniques, handling missing modalities, extending to other datasets, optimizing for real-time applications, and developing interpretable models to make these systems more adaptive and applicable in various practical contexts.

REFRENCES

[1] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. 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*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. doi: 10.18653/v1/P19-1050.

[2] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, and L.-W. Ku, “Emotionlines: An emotion corpus of multi-party conversations,” *arXiv preprint arXiv:1802.08379*, 2018.

[3] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[4] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Fusion with Hierarchical Graphs for Multimodal Emotion Recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1288–1296.
[5] J. Liu *et al.*, "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6339–6343.
[6] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 985–1000, 2021.
[7] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3227–3231.
[8] D. Cevher, S. Zepf, and R. Klinger, "Towards multimodal emotion recognition in german speech events in cars using transfer learning," *arXiv preprint arXiv:1909.02764*, 2019.
[9] Y. Wu, P. Peng, Z. Zhang, Y. Zhao, and B. Qin, "An Efficient End-to-End Transformer with Progressive Tri-modal Attention for Multi-modal Emotion Recognition," *arXiv preprint arXiv:2209.09768*, 2022.
[10] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, 2021.
[11] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1047–1056.
[12] 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.
[13] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf Process Syst*, vol. 30, 2017.
[14] J.-H. Bang *et al.*, "CA-CMT: Coordinate Attention for Optimizing CMT Networks," *IEEE Access*, 2023.
[15] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Adv Neural Inf Process Syst*, vol. 30, 2017.
[16] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans Neural Netw*, vol. 20, no. 1, pp. 61–80, 2008.
[17] H. Filali, J. Riffi, I. Aboussaleh, A. M. Mahraz, and H. Tairi, "Meaningful Learning for Deep Facial Emotional Features," *Neural Process Lett*, vol. 54, no. 1, pp. 387–404, 2022.