Indian Engineering Journal
ISSN: 2984-9047
Vol 4 Issue 2 (2026)

Enhanced Hierarchical Attention Fusion Network for Multimodal Conversational Emotion Recognition using IEMOCAP and MELD Datasets

B. Karthikeyani[1], T. Parimalam[2], D. Yuvaraj[3]

[1]Research Scholar, Nandha Arts and Science College (Autonomous) Erode. keyani.sb@gmail.com

[2]Associate Professor & Head, Department of Computer Science, Nandha Arts and Science College (Autonomous)

Erode. pari.phd12@gmail.com

[3]Assistant Professor, Department of Computer Science and Design, Kongu Engineering College,

Erode, yuvrajbee@gmail.com

Abstract:

Emotion recognition is still facing lot challenges to handle complex and contextual emotions of human in real time as well as in the field of healthcare too. Lot of unimodal and advanced machine learning techniques were used to handle those data. From the unimodal to multimodal approach which comprises of handling different contexts like facial expressions, speech and text related information using advanced techniques namely convolutional neural networks (CNN), bidirectional long short-term memory (Bi-LSTM), and Transformer architectures. Still the accuracy is challenging one with the existing datasets like CMU-MOSI and CMU-MOSEI and achieved nearly 80% accuracy with the help of fusion techniques with some limitations in balancing factor with different modality, high complex during computation and less adoptability with real time data. To propose a work with enhanced fusion techniques for multimodal emotion recognition with the IEMOCAP and MELD dataset since it is having large number of data than the CMU based dataset. Effective facial based features are handled by the CNN; speech variations are captured and analyzed efficiently with the help of Bi-LSTM and for text analysis transformers are used with the help of specialized encoders. The classification of emotional cues and multimodal contexts achieves better result than the existing one nearly 8-10% with the help of enhanced fusion techniques.

Keywords: Multimodal Emotion Recognition, Conversational Emotion Recognition, Hierarchical Attention, Attention Fusion Network, IEMOCAP, MELD, Deep Learning, CNN, Bi-LSTM, Transformer, Multimodal Fusion, Facial Expression Recognition, Speech Emotion Recognition, Text Emotion Analysis.

1. Introduction

In human-computer interaction the term Emotion recognition plays a vital role in monitoring, computing processes and signal analysis. In general, the facial emotions are not only taken for consideration also voice based as well as context-based dialogue also taken into the consideration with the multimodal support. Though unimodal features often fail to predict the exact output of a person, the multimodal supports with the help of text, speech and facial reactions cumulatively generates the output with better accuracy. Recent emotional systems are mainly works based on handcrafted features and traditional machine learning methods which yields the result with limited performance in real world scenarios.

Some advanced methods in deep learning methods and fusion techniques are used in the improved emotion recognition. With the help of datasets CMU-MOSI and CMU-MOSEI with the limited number of data already the accuracy has reached up to 80% while using fusion techniques. So, the high volume of dataset has been chosen namely IEMOCAP and MELD to handle more amount of data with the help of advanced fusion techniques. The IEMOCAP dataset contains different data like speech, facial expressions, and motion-based cues-based data for analysing the exact one. Same as, the MELD dataset contains data like textual, audio, and visual information thorough every utterance. To identify the conversational context some advanced methods like DialogueRNN was introduced to track the speaker state. Still lot of challenges like imbalance in modality, representation of difference noises and complexity due to high computation remains in real time scenarios.

From the above all constraints to propose a work called **Enhanced Hierarchical Attention Fusion Network (E-HAF)** for multimodal conversational emotion recognition using the IEMOCAP and MELD datasets. A framework is used to handle three different feature extraction modules in an efficient way. First module is used to handle the spatial related facial feature extraction. Second module generally used under the network mechanisms for temporal speech feature extraction and the third module is used to generate the semantic representation of all text related feature extraction with the help of transformer-based encoders. Not like the tradition fusion based techniques to handle all the modalities, the E-HAF mechanism is represented with individual attention to each modality to perform better context aware fusion. To avoid the noisy modality, this enhanced fusion technique plays a dominant role while capturing the different emotions. The final result of the enhanced fusion techniques the framework achieves the better accuracy up to 88-90%. Finally, the enhanced techniques provide more robust and efficient solution for multimodal emotion recognition.

2. Related Work

The conversion of unimodal classification to multimodal context learning, the emotion recognition is dramatically changed from the existing to proposed advanced methods under deep learning and transformer encoders. The existing system were used with only in unimodal based information like face reactions, speech context and text information. Always informal conversation is expressed through combination of words, facial expressions and text generated by the humans. Now the recent surveys are focused on handling combined features like visual, speech and text related information with the help of advanced deep learning and attention based fusion techniques.

A standard dataset is used in this work to develop the better feature extraction process. IEMOCAP dataset having data like face reactions, speech context and text information for multimodal emotional process. MELD extends this direction to multiparty conversations by providing text, audio, and visual information for each utterance, along with emotion and sentiment labels. These datasets support the evaluation of both context modelling and multimodal fusion in emotion recognition tasks [1], [2].

2.1 Existing System: Mechanisms used for Maximum Accuracy

A. Traditional Emotion Recognition Systems:

Actually, the feature based or decision-based emotions are commonly used in the Traditional multimodal emotion recognition systems. In general, text, audio and video combinations are mainly used to derive the feature-based decision using fusion techniques. Below the simple representation of fusion technique is given:

$$z_{base} = [h_T \| h_A \| h_V], \qquad \hat{y} = softmax(W_c z_{base} + b_c) \qquad (1)$$

Here, h_T, h_A, and h_V denote text, audio, and visual feature representations respectively, and $\|$ denotes feature concatenation.

B. DialogueRNN:

The global conversation-based contexts and specific states modeling are used to improve the conversational emotional recognition. The speaker state update can be measured with the help of tracking the Speaker state utterance which is understood with the current utterance with respect to the Previous behavior [3].

$$q_{s_i,i} = GRU_p\left(q_{s_i,i-1}, [u_i, g_i]\right) \qquad (2)$$

where $q_{s_i,i}$ is the state of speaker s_i at utterance i, u_i is the utterance representation, and g_i is the global conversational context.

The matching utterances are derived from the emotion related information through graph-based message passing with the help of adjacency matrix values [4]. A simplified graph convolution update can be written as:

$$H^{(l+1)} = \sigma\left(\sum_{r \in R} \tilde{A}_r H^{(l)} W_r^{(l)}\right) \qquad (3)$$

where $\tilde{A}_r$ is the normalized adjacency matrix for relation r, and $H^{(l)}$ is the node representation at layer l.

C. Transformer-based multimodal systems:

Fusion techniques improve the learning by multimodal transformers along with the cross-modal interactions which does not require any word-level alignment. Based on the dependencies between the modalities which helps to maintain a strongest mechanism represented as given below.

$$Attn(Q_m, K_n, V_n) = softmax\left(\frac{Q_m K_n^T}{\sqrt{d_k}}\right) V_n \qquad (4)$$

Here, Q, K, and V are query, key, and value matrices, and the notation represents attention from modality m to modality n.

The multimodal graph representation is mainly based on the utterance representation and modality-based information which is represented as:

$$H_{MM}^{(l+1)} = GCN\left(H_T, H_A, H_V, A_{context}, A_{modality}\right) \qquad (5)$$

The fused graph combines textual, acoustic, visual, contextual, and modality-level relations.

Based on the existing work, earlier HAF-based systems on CMU-MOSI and CMU-MOSEI achieved nearly 80% accuracy, while the proposed work aims to improve performance on IEMOCAP and MELD through enhanced attention-based fusion.

3. Proposed Work: Module-Wise Explanation and Working Mechanism

To propose a work which free from the above limitations namely Enhanced Hierarchical Attention Fusion Network (E-HAF) is handling with IEMOCAP and MELD datasets for multimodal conversational emotion recognition.

3.1 Visual Feature Extraction Module

From video frames the facial features are extracted with the help of advanced Convolutional Neural Network method. Some of the commonly used emotion cures namely happy, sad, angry, surprise and neutral. Based on the muscle movement in the face of the speaker the CNN extract the facial pattern in a hierarchical way vis local learning which includes Eye movement, mouth movement and common facial reactions. The extracted visual representation is defined as:

$$h_V = CNN(X_V; \theta_V) \qquad (6)$$

where X_V is the visual input and θ_V denotes the CNN parameters.

3.2 Speech Feature Extraction Module

The Bi-LSTM network process with the speech related modules by capturing the temporal speech dynamics. Based on few parameters like pitching, rhythm, intensity in voice and speaking rate. The acoustic representation is given as:

$$h_A = BiLSTM(X_A; \theta_A) \qquad (7)$$

where X_A is the acoustic input and θ_A denotes the BiLSTM parameters.

3.3 Text Feature Extraction Module

Transformer based encoders are used to understand the semantic and contextual meaning based on the utterance transcripts. Some of the emotional cues were direct like charging words from the speaker, negotiation, contextual speech. The textual representation can be written as:

$$h_T = Transformer(X_T; \theta_T) \qquad (8)$$

where X_T is the text input and θ_T denotes the Transformer encoder parameters.

4. Output and Comparative Analysis with Suggested Figures

4. 1. Dataset Description

To propose a work which is intended to find the accuracy between two parties and more than two parties based on two datasets namely IEMOCAP and MELD. Between two parties' conversation-based context, audio visual based and facial motion based were processed in the IEMOCAP dataset. More than two or Multi parties-based emotion supports were processed with the help off utterance level-based text, audio and video-based modalities in MELD dataset with the below figure 1.

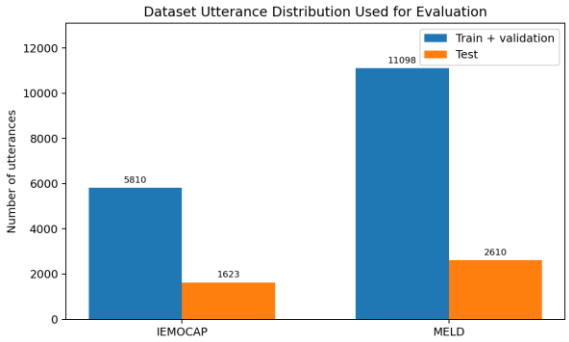

Figure 1. Dataset utterance distribution suggested for the output/comparison section.

4. 2. Experimental Setup

Different weight mechanisms are used in the Enhanced hierarchal with visual based, acoustic based and textual based cues. It can handle multi way to attain the better accuracy like suppose text is unclear but voice and audio is too strong then based on the higher attention the decision will be taken. Below the working mechanism of E-HAF with its output flow is given.

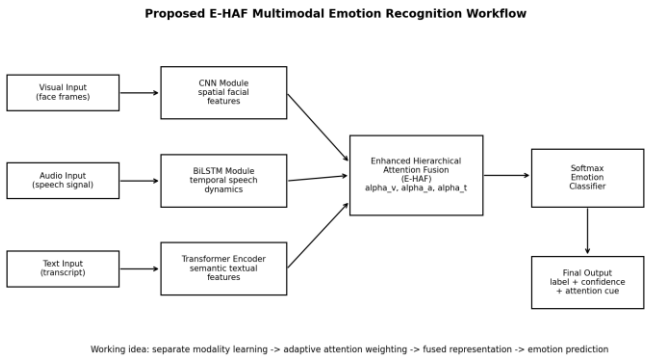

Figure 2. Proposed E-HAF architecture and output workflow.

4.3 Comparison with Existing Systems

To reduce the imbalance in modality and dominance in choosing suitable emotional cues proposed a E-HAF which handles hierarchical attention through CNN, BiLSTM and Transformer encoders [3]-[7]. Below the Figure 3 state the comparison between the existing HAF with E-HAF based on the accuracy.

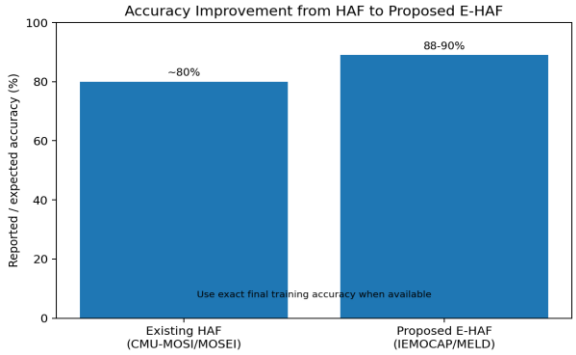

Figure 3. Existing HAF versus proposed E-HAF accuracy comparison based on the provided draft accuracy range.

5. Evaluation of Proposed E-Haf Performance

The evaluation takes placed in 4 states such as, the initial state calculates the accuracy from the utterance level compared with the true labels. In the next state, report generated based on class wise, precision, recall and F1-score wise to identify the complex emotion classes. In the third state, analysis has to be done based on the classification with complex pairs. At last, in final state, all the mismatched components were removed one by one and show the final prediction based on the hierarchical attention shows in the below Figure 4.

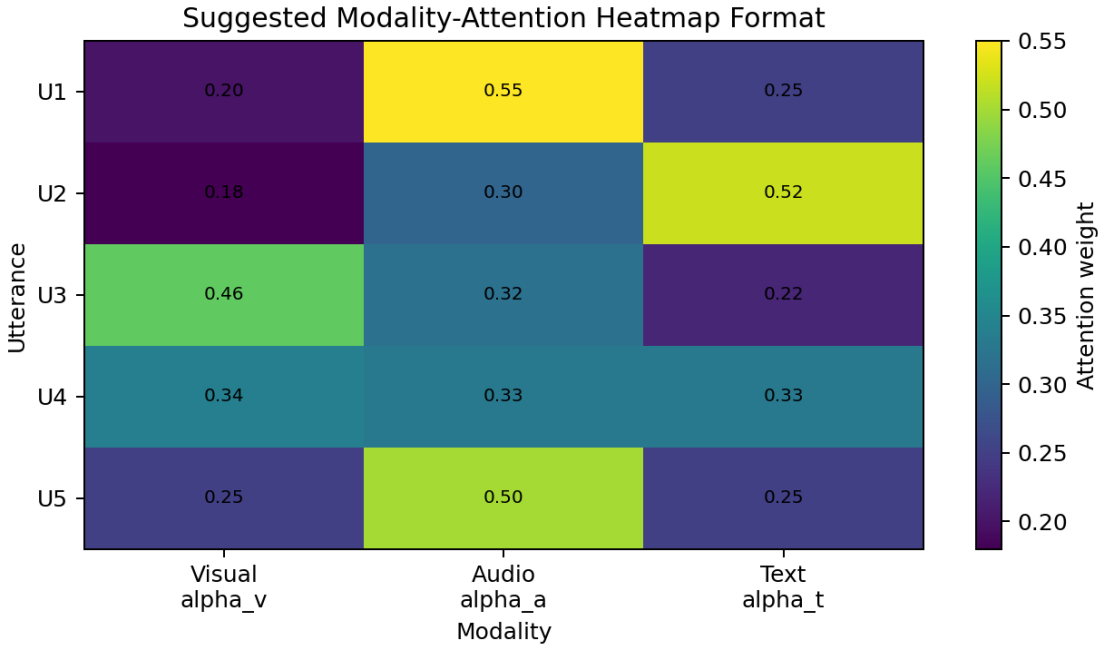

Figure 4. Suggested attention heatmap format for explaining the dominant modality per utterance. Replace template weights with real E-HAF attention scores.

6. Conclusion

The main aim of this research work is to addresses the major weakness in the unimodal emotion recognition systems as well as the integration of different modalities with the limited abilities to understand the verbal and non-verbal cues. The proposed work mainly comprises with three different methods namely CNN, Bi-LSTM and Transformer encoders which extracts features from the humans, speech conversations and textual information respectively. After all the process to identify the exact context based on the evaluation done by the Enhanced Hierarchical Attention Fusion mechanism (E-HAF) from different modalities at a time. From this the weak or

noise modality were reduced and deal with more emotional cues based on the utterance level. Based on the experimental results, it shows the accuracy between 88-90% over the existing HAF mechanisms when it deals with the IEMOCAP and MELD dataset. Also, this can be extended up to all affective computing-based applications, human-computer interactions and online education. It can further strengthen its accuracy, efficiency and real time applicability when it comes for all sort of applications.

References

[1] C. Busso, M. Bulut, C.-C. Lee, E. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, 2008, doi: 10.1007/s10579-008-9076-6.

[2] 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 Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 527-536, doi: 10.18653/v1/P19-1050.

[3] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," arXiv preprint arXiv:1606.06259, 2016.

[4] A. B. Zadeh, P. P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018, pp. 2236-2246, doi: 10.18653/v1/P18-1208.

[5] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada, 2017, pp. 873-883, doi: 10.18653/v1/P17-1081.

[6] 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 Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, LA, USA, 2018, pp. 2122-2132, doi: 10.18653/v1/N18-1193.

[7] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in Proc. AAAI Conference on Artificial Intelligence, vol. 33, no. 1, 2019, pp. 6818-6825, doi: 10.1609/aaai.v33i01.33016818.

[8] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 154-164, doi: 10.18653/v1/D19-1015.

[9] 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 Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 6558-6569, doi: 10.18653/v1/P19-1656.

[10] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in Proc. 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Online, 2021, pp. 5666-5675, doi: 10.18653/v1/2021.acl-long.440.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998, doi: 10.1109/5.726791.

[14] S. Kadu and B. Joshi, "A hierarchical attention-based fusion model for multimodal sentiment analysis of customer-generated review videos," Vietnam Journal of Computer Science, pp. 1-34, 2025, doi: 10.1142/S2196888825500228.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. International Conference on Learning Representations (ICLR), 2017.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.