Deep Multimodal Emotion Recognition using Modality Aware Attention Network for Unifying Representations in Neural Models

Sungpil Woo*
Autonomous IoT
ETRI / KAIST
Daejeon, South Korea
woosungpil@etri.re.kr

Muhammad Zubair*
Autonomous IoT
ETRI
Daejeon, South Korea
zubair5608@etri.re.kr

Sunhwan Lim
Autonomous IoT
ETRI
Daejeon, South Korea
shlim@etri.re.kr

Daeyoung Kim
School of Computing
KAIST
Daejeon, South Korea
kimd@kaist.ac.kr

Abstract

This paper introduces a multi-modal emotion recognition system aimed at enhancing emotion recognition by integrating representations from physiological signals. To accomplish this goal, we introduce a modality aware attention network to extract emotion-specific features by influencing and aligning the representation spaces of various modalities into a unified entity. Through a series of experiments and visualizations conducted on the AMIGO dataset, we demonstrate the efficacy of our proposed methodology for emotion classification, highlighting its capability to provide comprehensive representations of physiological signals.

1 Introduction

Affective computing represents an emerging field dedicated to endowing intelligent systems with the ability to recognize and interpret human emotions[16, 13]. This endeavor seeks to instill computers with sophisticated affect recognition and interpretation capabilities akin to those found in humans, achieved through the utilization of robust computational models.

Nonetheless, the morphological attributes of physiological signals exhibit individualized variations influenced by an individual's prevailing physiological processes and mental state, rendering these signals subject to temporal fluctuations[12, 1]. Furthermore, the integration of multi-modal data introduces a layer of complexity, posing challenges for classification models. As a result, the adoption of efficient deep learning methodologies becomes imperative to effectively address these complexities.

This study delves into the realm of deep multi-modal representation learning with a specific focus on physiological signals. we propose a modality aware attention network meticulously crafted to extract emotion-specific features by influencing and aligning the representation spaces of these heterogeneous modalities into a cohesive and unified entity. We substantiate the efficacy of our proposed methodology for emotion classification through rigorous experimentation and visualizations conducted on the AMIGO dataset[1].

*equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Methodology

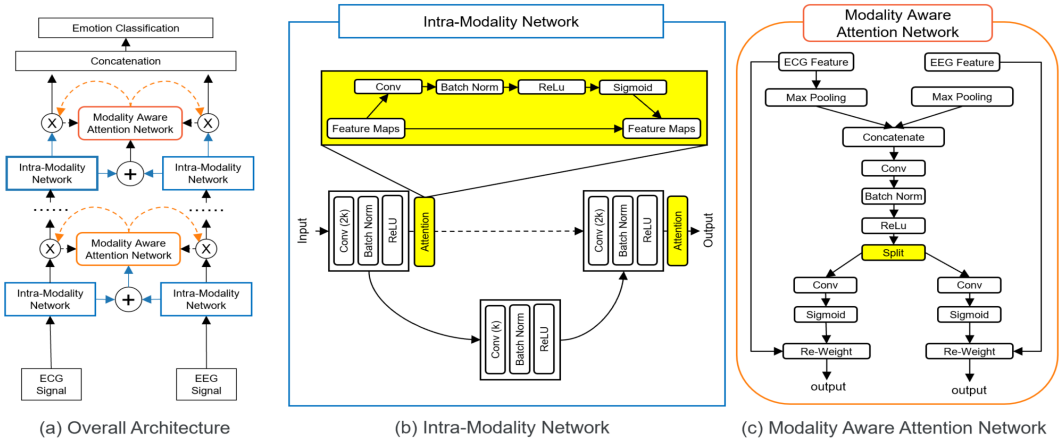

Figure 1: Overview of proposed multi-modal emotion recognition system

We introduced a deep multimodal emotion recognition system, as illustrated in Figure 1. The core objective of this proposed emotion recognition system is to acquire an effective deep representation of physiological signals through the extraction of emotion-specific information from diverse multimodal sources. Our approach involved the utilization of ECG and EEG data to discern and classify two distinct levels of arousal and valence.

2.1 Intra-Modality Network

The proposed Intra-Modality Network is responsible for extracting and generalizing intra-modality information by exploiting variations of ECG and EEG signals triggered by emotional stimuli. The proposed Intra-Modality Network is used in both branches to extract emotion-specific information at different levels. Each unit on the Intra-Modality Network follows a U-NET Architecture[9] with a squeeze and Excitation strategy[10]. For EEG branch, we employed 2D CNN followed by batch normalization and activation layers as an elementary unit of the U-NET architecture. Whereas for ECG branch we use 1D CNN with batch normalization and activation layer to compose U-NET architecture. To emphasize more on target-specific extraction of features, we adopted an attention mechanism as depicted in Figure 1-(b). The proposed architecture tackles the data shift issue corresponding to intra-subject variability triggered by mood and external stimuli.

2.2 Modality Aware Attention Network

We design a modality-aware attention network for the fusion of emotion-specific information from inter-modality data. The proposed network assures the extraction of the most relevant and significant information from the multi-modal signals and significantly regulates the representation space of each modality associated with the target emotion. The modality-aware attention learns the weights for each modality based on its contribution to emotion classification.

The attention masks for ECG (A_{ecg}) and EEG (A_{eeg}) features are generated by exploiting the inter-modality relationship of the input feature maps from their respective modalities, F_{ecg} and F_{eeg}. A maximum pooling operation is performed on the temporal and spatial axis of ECG and EEG features maps, respectively. The extracted 1D feature maps are concatenated to estimate mutually imported and target-relevant feature scores. After the first convolutional layer, ECG and EEG feature maps are split to model the attention score for each modality separately. The generated attention masks are then used to scale their respective modality features yielding weighted feature maps that carry the most relevant and domain invariant information.

The proposed modality-aware attention network alleviates the sporadic representation of ECG and EEG in feature space by exploiting mutually important and target-specific features while taking the features of both modalities into consideration. As a result, the congregated representation of ECG and EEG features mutually boosts the classification performance of the emotion classifier.

3 Experimental Setting

3.1 Dataset

In our study, we assess the performance of our proposed method using the AMIGOS dataset. The AMIGOS dataset is a widely recognized resource in emotion recognition research due to its rich variety of data modalities. This dataset includes facial videos, synchronized physiological signals, and emotion annotations. Specifically, our research leverages Electrocardiogram (ECG)[4, 8, 14] and Electroencephalogram (EEG)[3, 6, 15, 17, 7, 5] signals for analysis.

3.2 ECG Pre-processing

To establish a consistent input data structure from the raw ECG signal, we begin by detecting the R peaks, followed by the computation of the heart rate variability (HRV) series. Prior to generating the HRV series, we normalize the ECG signal. To ensure uniform segment lengths within the HRV series, we apply the zero embedding technique, appending zeros to each sample while preserving consistent segment lengths across the dataset.

3.3 EEG Pre-processing

To establish an input data structure from the raw EEG signal, recent literature has highlighted the efficacy of Differential Entropy in emotion recognition using EEG[2], [11]. In this study, we computed the differential entropy for each frequency band. Initially, we decomposed the EEG signal from all channels into four specific frequency bands, namely theta, alpha, beta, and gamma. Subsequently, we calculated the differential entropy for each of these bands across all channels using the following formula:

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\pi\sigma^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} \tag{1}$$

$$exp \frac{(x-\mu)^2}{2\pi\sigma^2} = \frac{1}{2} \log 2\pi e\sigma^2 \tag{2}$$

4 Results and Discussion

This section offers a summary of the experimental results for our proposed method. Initially, we trained a multi-modal network using ECG and EEG signals in the absence of our proposed attention network. Subsequently, we integrated our proposed attention network into the model.

4.1 Multi-modal Emotion Classification Performance

The input configuration of the multi-modal network was deliberately engineered to guarantee that both ECG and EEG segments were synchronized with the same 10-second time frame. In the case of the ECG data, we extracted the Heart Rate Variability (HRV) series, incorporating zero padding. For the EEG data, we focused on extracting the differential entropy features within 10-second segments across various frequency bands. To assess the performance of our model, we adopted a rigorous 10-fold cross-validation methodology in our experimentation.

Table 1: Classification Performance

Classifier	Arousal(%)			Valence(%)		
	ACC	SEN & SPE	PPR	ACC	SEN & SPE	PPR
(a-1) Basic model	81.55%	80.28%	80.66%	85.85%	80.28%	85.51%
(b-1) Proposed model	**91.68%**	**90.91%**	**91.5%**	**93.69%**	**93.64%**	**93.67%**

Table 1 presents the classification performance of the two different setups of the multi-modal emotion recognition system. These results demonstrate that the proposed modality aware attention module outperform the base model and achieve better classification performance. The deep architecture assisted by the modality aware attention module played a key role in multi-modal learning.

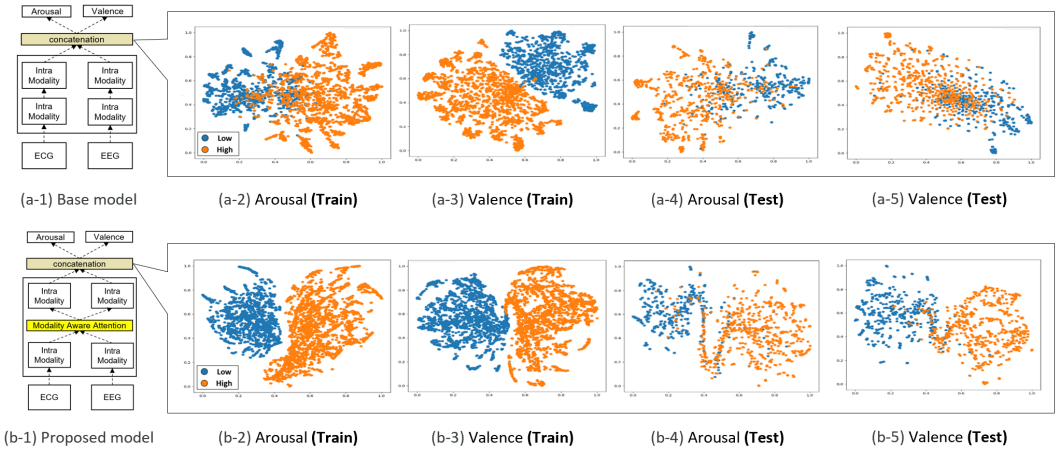

Figure 2: Model Architectures and Visualization Result

4.2 Visualization of Representation space

Figures 2 depict feature visualizations of arousal and valence for both base and proposed model, respectively. The visualizations demonstrate the substantial influence of the modality aware attention network on the feature space of the arousal and valence models.

Fig. 2-(a-4) and Fig. 2-(a-5) illustrate the overlapping feature distribution of two arousal and valence levels observed in the test set using the base model. In contrast, Fig. 2-(b-4) and Fig. 2-(b-5) offers a clear representation of margin enhancement and variance reduction, highlighting the effectiveness of the proposed modality aware attention network for making synergy of each modality.

5 Conclusion

The objective of this research is to develop a multi-modal emotion recognition system capable of efficiently learning representations from multi-modal signals. This goal is accomplished by consolidating the representation space and fostering positive synergy through the use of a modality aware attention network.

From our perspective, both the classification performance and the visualization illustrations strongly validate the soundness and efficacy of the proposed modality aware attention network. These results furnish compelling evidence supporting the implications of this attention network, encompassing the reduction of variance within class clusters and the establishment of a discrimination margin between classes. Collectively, these factors make a substantial contribution to the remarkable outcomes achieved by unifying representations through the provision of feedback via attention to each modality's representation space.

Future research efforts should focus on tuning the proposed model architecture, integrating additional modalities, such as facial video, and design the loss function to further enhance multi-modal classification performance.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2020-0-00048, Development of 5G-IoT Trustworthy AI-Data Commons Framework, 50%) and (2022-0-01032, Development of Collective Collaboration Intelligence Framework for Internet of Autonomous Things, 50%)

References

[1] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 2018.

[2] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, and H. Wang. An efficient lstm network for emotion recognition from multichannel eeg signals. *IEEE Transactions on Affective Computing*, 2020.

[3] B. H. Kim and S. Jo. Deep physiological affect network for the recognition of human emotions. *IEEE Transactions on Affective Computing*, 11(2):230–243, 2018.

[4] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083, 2008.

[5] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[6] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu. Emotion recognition from multi-channel eeg data through convolutional recurrent neural network. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 352–359. IEEE, 2016.

[7] W. Lin, C. Li, and S. Sun. Deep convolutional neural network for emotion recognition using eeg and peripheral physiological signal. In *International Conference on Image and Graphics*, pages 385–394. Springer, 2017.

[8] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, 6(4):385–394, 2015.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[10] A. G. Roy, N. Navab, and C. Wachinger. Recalibrating fully convolutional networks with spatial and channel 'squeeze excitation' blocks, 2018.

[11] T. Song, W. Zheng, P. Song, and Z. Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[12] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.

[13] L. Sun, Q. Li, S. Fu, and P. Li. Speech emotion recognition based on genetic algorithm–decision tree fusion of deep and acoustic features. *ETRI Journal*, 44(3):462–475, 2022.

[14] G. Valenza, A. Lanata, and E. P. Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE transactions on affective computing*, 3(2):237–249, 2011.

[15] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140:93–110, 2017.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[17] W.-L. Zheng, H.-T. Guo, and B.-L. Lu. Revealing critical channels and frequency bands for emotion recognition from eeg with deep belief network. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 154–157. IEEE, 2015.