MULTIMODAL TRANSFORMER FUSION FOR CONTINUOUS EMOTION RECOGNITION

Jian Huang[1,3]*,Jianhua Tao*[1,2,3]*,Bin Liu*[1]*,Zheng Lian*[1,3]*,Mingyue Niu*[1,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{jian.huang, jhtao, liubin, zheng.lian, mingyue.niu}@nlpr.ia.ac.cn

ABSTRACT

Multimodal fusion increases the performance of emotion recognition because of the complementarity of different modalities. Compared with decision level and feature level fusion, model level fusion makes better use of the advantages of deep neural networks. In this work, we utilize the Transformer model to fuse audio-visual modalities on the model level. Specifically, the multi-head attention produces multimodal emotional intermediate representations from common semantic feature space after encoding audio and visual modalities. Meanwhile, it also can learn long-term temporal dependencies with self-attention mechanism effectively. The experiments, on the AVEC 2017 database, shows the superiority of model level fusion than other fusion strategies. Moreover, we combine the Transformer model and LSTM to further improve the performance, which achieves better results than other methods.

Index Terms— Continuous emotion recognition, model level fusion, Transformer, multi-head attention

1. INTRODUCTION

Emotional intelligence enables the human-machine interaction more harmoniously. The emotions can be described by continuous space through attribute dimensions such as arousal and valence, which use numerical values to indicate emotional type and degree [1].

Human expresses emotional state related information through multimodal ways. One modality can be a semantic complementary for another modality in expressing similar emotions. Busso et al. [2] showed the fusion of audio and visual modalities improved the performance and robustness of emotion recognition systems measurably.

There are mainly three strategies in the efforts of multimodal emotion fusion, namely decision level fusion, feature level fusion and model level fusion [3]. On the decision level fusion, multiple modalities are modeled independently, then these single modal recognition results are combined to obtain final predictions [4], which it ignores the interactions between different modalities. Traditional feature level fusion directly feeds the concatenated features into a classifier or uses shallow-layered fusion models [5], but it has the difficulty to learn mutual relationships among different modalities. Another alternative strategy of feature level fusion is multimodal representation learning. The main approach is to learn joint representations from shared hidden layer connected with multiple modalities inputs. The models are usually based on deep learning frameworks, like deep autoencoder and DNN [6]. Kim et al. [7] proposed four Deep Belief Networks (DBNs) architectures to capture complex non-linear multimodal feature correlations for emotion recognition.

Compared with feature level and decision level fusion, model level fusion learns multimodal interactions inside the models and makes better advantages of deep neural networks. Prior researches included Hidden Markov Model (HMM) [8], kernel models [9] and neural networks [10][11][12]. The attention mechanisms were proposed to learn the alignment between audio-visual [10] and audio-text streams [11]. Chen et al. [12] proposed temporal fusion model to dynamically pay attention to relevant modality features through time, which made the improvements over traditional fusion strategies.

Long Short-Term Memory Networks (LSTM) are popular emotion recognition models due to their ability of learning emotional dynamic temporality with the recurrence structure. Recently, a more effective no-recurrence Transformer model was proposed [13]. It models long-term crucial temporal dependencies on the longer span of time which is more suitable to model emotional temporal process. Tsai et al. [14] utilized Transformer to analyze human multimodal language and exhibited the best performance. In this paper, we draw lessons from their works to learn semantic-level correlations across audio-visual modalities with the Transformer model and achieve model level fusion for continuous emotion recognition.

The rest of paper is organized as below: section 2 briefly introduces the proposed methods. Section 3 presents the database and feature sets. Section 4 describes the experimental results and analysis. Section 5 concludes this paper.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

2. PROPOSED METHODS

In this work, we perform Transform fusion for continuous emotion recognition. The Transformer model learns emotional long-term temporal dependencies with self-attention mechanism. What's more, we achieve model level audio-visual modalities fusion with the multi-head attention module.

2.1. Multimodal Transformer fusion

We utilize the Transformer model to achieve model level fusion, illustrated in Fig. 1. The whole model has no encoder-decoder structure, and consists of three parts: audio, visual and multimodal fusion module. Audio and visual modules, directly attending to low-level features, are responsible to learn respective emotional long-term temporal dependencies with self-attention mechanism. Then, multimodal fusion module attends to interactions between the outputs of two single modules and latently adapts streams from audio modality to visual modality via the attention. The motivation is to transform high level output representations of audio and visual modules into a common semantic feature space, then produce effective multimodal feature representations. On this basis, we directly predict the emotional value behind the representations to accomplish model level fusion.

Furthermore, we also achieve decision level fusion and feature level fusion based on the Transformer network as shown in Fig. 2. They are based on self-attention mechanism to learn emotional temporal dynamic information. The difference is where to fuse the representations. Decision level fusion firstly encodes audio and visual modalities with individual modules independently, then concatenates the representations for final predictions. Feature level fusion firstly concatenates the representations, then inputs them to single module for emotion modeling. Note the beginning of the model introduces a linear layer to transform original audio and visual inputs to emotional feature space.

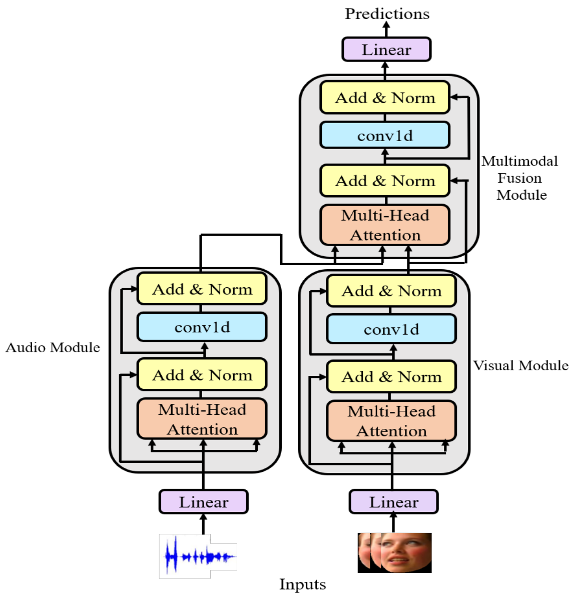
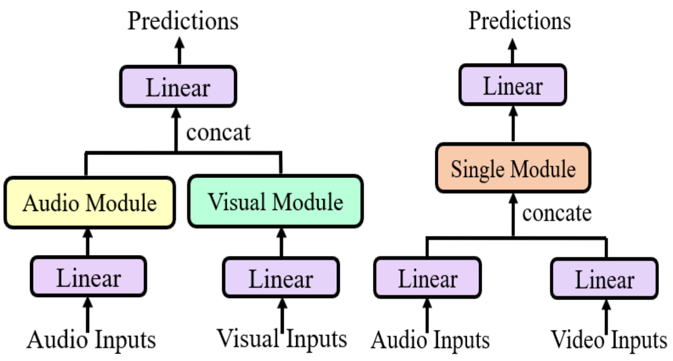


Fig. 1: The overview of the proposed model. Audio and visual modules learn emotional long-term temporal dependencies that attends to low-level features. Multimodal fusion module attends to interactions across audio-visual modalities and latently adapts streams from audio modality to visual modality via the attention.



(a) Decision level fusion (b) Feature level fusion

Fig. 2: Two traditional multimodal emotion fusion based on the Transformer network.

2.2. Multi-head attention

Actually, audio module, visual module and multimodal fusion module of section 2.1 are multi-head attention. It extends conventional attention mechanism to have h multiple heads, which allows each head to have a different role on attending the encoder outputs. Specifically, the multi-head attention calculates h times Scaled Dot-Product Attention independently, then concatenates their outputs to fed into another linear projection.

Scaled Dot-Product Attention has three inputs: queries, keys of dimension d_k and values of dimension d_v. One query's output is computed as a weighted sum of the values, which is computed by a designed function of the query with corresponding key.

$$\text{Attention}(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$$

For single module above, Q, K and V are same inputs. For multimodal fusion module, Q is encoded visual features and K, V are encoded audio features in Fig. 1. Besides, every multi-head attention module is followed with one conv1d layer to focus on short temporal context, which is temporal convolutional neural network [15].

3. DATABASE AND FEATURE SETS

3.1. Database

We use Audio/Visual Emotion Challenge and Workshop (AVEC 2017) database [16] to show the benefits of our proposed methods. The database collects spontaneous and naturalistic human-human interactions in the wild consisting of audio and visual modalities. The recordings are annotated time-continuously in terms of the emotional dimensions including arousal and valence. All emotional dimensions are annotated every 100ms and scaled into [-1, +1]. There are 64 German subjects in the dataset and are divided into the training set with 36 subjects, development set with 14 subjects and test set with 16 subjects. We focus on the estimation of arousal and valence in this work.

3.2. Features set

The audio and visual modalities are utilized for continuous emotion recognition. We adopt the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [17] as audio features. The window segment-level acoustic features are computed over overlapping four seconds resulting in 88 dimensional features, extracted by openSMILE [18]. The geometric features are regarded as visual features including facial landmarks locations, Facial Action Units (AUs), head pose features and eye gaze features. Then, we apply principal component analysis (PCA) to reduce the dimensionality and 0.95 variance is kept.

4. EXPERIMENTAL AND ANALYSIS

4.1. Experimental setup

All multi-head attention modules of Fig. 1 and Fig. 2 are composed of two 4-head attention layers. Besides, each of two attention layers has a residual connection and layer normalization. The number of hidden nodes of attention layer and output channels of conv1d layer are 64. Similar to the works [19], we employ data augmentation, temporal pooling and delay compensation strategies. We use adam optimization algorithm and dropout with the rate 0.5 (12 mini-batch size and 70 training epochs). The evaluation measure is the Concordance Correlation Coefficient (CCC) [20].

4.2. Multimodal Transformer fusion

Firstly, we perform unimodal continuous emotion recognition as the comparison. The single modality features are inputed to the multi-head attention module for continuous emotion modeling with self-attention mechanism. The experimental results in Table 1, show unimodal model can achieve effective performance. It indicates the Transformer model can be applied for emotion modeling successfully and verifies the strong strength and universality of the Transformer model. The self-attention structure learns whole dependencies from global information, and models emotional long-term dynamic information effectively. Besides, the visual features achieve better performance than audio features in arousal and valence, which is similar to the works [5][16].

Then, we utilize three different strategies to fuse audio-visual modalities. The order of audio and visual features as the inputs has an impact on the model level fusion [14]. Fig. 1 shows the final generated high-level representations are based on the visual features attending to audio features (audio on the left and video on the right), represented by "AV" in Table 1. In addition, we can exchange their orders (video on the left and audio on the right), represented by "VA" in Table 1.

The experimental results show that most of multimodal systems achieve better performance than unimodal systems. Multimodal fusion makes positive effects on performance improvement. Decision level fusion obtains better performance than feature level fusion, which is similar to the works [6]. Thus, decision level fusion realizes better complementarity of emotional information from late individual predictions, while the simple concatenation of feature level fusion can't learn mutual correlations exactly.

On the model level fusion, the performance of "AV" model is better than visual features, and "VA" model is better than audio features. The "AV" model achieves better performance than the "VA" model. It reveals the "AV" model captures the visual features as the principle part while the "VA" model captures the audio features. The results conform with that the performance of visual features is superior to audio features. The "AV" model achieves better performance than decision level fusion and feature level fusion, obtaining best CCC 0.629 in arousal and 0.593 in valence. The "AV" model enables visual modality for receiving information from audio modality and learns the correlation of audio-visual modalities to accomplish model level fusion. What' more, model level fusion models long-term temporal dependencies efficiently in the phase of multimodal fusion, which promotes the performance significantly. In general, the performance of arousal is better than valence. Perhaps, arousal needs short temporal context [21] which makes it easier to emotional modeling for the Transformer model.

Table 1: Performance of unimodal and multimodal fusion in arousal and valence from audio and visual features.

	Arousal	**Valence**
Audio	0.471	0.315
Visual	0.602	0.557
Decision level fusion	0.611	0.589
Feature level fusion	0.597	0.586
VA	0.494	0.371
AV	**0.629**	**0.593**

Table 3: CCC comparison of unimodal and multimodal fusion between our proposed models and other methods.

	Arousal			Valence		
	Audio	Visual	Multimodal	Audio	Visual	Multimodal
SVM [16]	0.344	0.466	0.525	0.351	0.400	0.507
LSTM [19]	0.497	0.601	0.615	**0.438**	**0.662**	0.671
Transformer	0.471	0.612	0.629	0.315	0.557	0.593
Transformer+LSTM	**0.519**	**0.623**	**0.654**	0.421	0.647	**0.708**

4.3. Multimodal Transformer fusion with LSTM

Multimodal fusion module generates a latent cross-modal representations that fuses audio-visual information. In this section, we explore the capability of high-level representations. Specifically, we add one LSTM layer before final linear layer in all models. The number of hidden nodes of the LSTM layer is 64 and other parameter settings are similar. The experimental results are shown in Table 2.

Compared Table 2 with Table 1, the performance of all models is improved in different degree. It indicate the Transformer model can generate expressive high-level emotional representations, which can be better inferred by LSTM model. Besides, the model combining Transformer and LSTM model can learn emotional temporal dependencies better to obtain a promising increase in performance. The "AV" model achieves best performance 0.654 in arousal and 0.708 in valence, which is 0.025 higher in arousal and 0.105 in valence than Table 1.

Different from Table 1, the performance of feature level fusion is slightly better than decision level fusion. Thus, the addition of LSTM layer can help feature level fusion to learn complicated mutual relationships across audio-visual modalities. Another difference is that the performance of valence is better than arousal, probably the combination model can model longer temporal contexts to improve the performance of valence.

We compare our results with the baselines [16] and the runner-up [19] of AVEC 2017, as shown in Table 3. The baselines utilized SVM and feature level fusion. The work [19] utilized LSTM model and explored decision level and feature level fusion strategies. However, there is no multimodal fusion results for the eGeMAPS and geometric features. For fair comparison, we conduct the experiments with two multimodal fusion strategies using these two features following their methods. The best performances are listed in Table 3.

Table 2: Performance of unimodal and multimodal fusion with LSTM layer in arousal and valence from audio and visual features.

	Arousal	Valence
Audio	0.519	0.421
Visual	0.613	0.647
Decision level fusion	0.637	0.652
Feature level fusion	0.632	0.665
VA	0.564	0.477
AV	**0.654**	**0.708**

Our methods achieve better performance than SVM both in unimodal and multimodal systems. The results demonstrate the potential benefits of the Transformer model for continuous emotion recognition. The "Transformer+LSTM" model obtains best performance of single modality in arousal, which indicates our proposed models are more conducive to arousal prediction. The "Transformer+LSTM" model achieves best performance of multimodal fusion, which is 0.039 higher in arousal and 0.027 higher in valence than LSTM. The Transformer model achieves long span modeling from global information and LSTM focuses temporal context information. As a result, their combination further promotes the performance effectively and integrates audio-visual modalities more closely on the model level fusion.

5. CONCLUSIONS

This work explores multimodal emotion fusion across audio-visual modalities with the Transformer network for continuous emotion recognition. The Transformer model is utilized to learn long-term temporal dependencies with self-attention mechanism. The results show the potential benefits of the Transformer model to obtain more promising performance for continuous emotion recognition. The multi-head attention module is utilized to consider the interactions between audio-visual modalities on the model level fusion, which obtains better performance than decision level and feature level fusion. Further, we combine the Transformer network and LSTM layer to explore the capability of high-level representations. The "Transformer+LSTM" model achieves better performance than SVM and LSTM model in multimodal fusion systems. Our proposed model can integrate audio-visual modalities information efficiently on the model level fusion. In the future, we will extend multimodal Transformer fusion to other modalities like textual modality to improve the performance.

6. ACKNOWLEDGEMENTS

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049)。

REFERENCES

[1] Hatice Gunes and Maja Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[2] Carlos Busso, Zhigang Deng, and Serdar et al. Yildirim, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[3] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA transactions on signal and information processing*, vol. 3, 2014.

[4] Mengyi Liu, Ruiping Wang, and Shaoxin et al. Li, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on multimodal interaction*. ACM, 2014, pp. 494–501.

[5] Jian Huang, Ya Li, and Jianhua et al. Tao, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 11–18.

[6] Cheng Wang, Haojin Yang, and Christoph Meinel, "A deep semantic framework for multimodal representation learning," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9255–9276, 2016.

[7] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 3687–3691.

[8] Zhihong Zeng, Jilin Tu, and Brian et al. Pianfetti, "Audio-visual affect recognition through multi-stream fused hmm for hci," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 2, pp. 967–972.

[9] JunKai Chen, Zenghai Chen, and Zheru et al. Chi, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 508–513.

[10] Linlin Chao, Jianhua Tao, and Minghao et al. Yang, "Audio visual emotion recognition with temporal alignment and perception attention," *arXiv preprint arXiv:1603.08321*, 2016.

[11] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[12] Shizhe Chen and Qin Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 571–575.

[13] Ashish Vaswani, Noam Shazeer, and Niki et al. Parmar, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[14] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *arXiv preprint arXiv:1906.00295*, 2019.

[15] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang, "Phoneme recognition using time-delay neural networks," *Backpropagation: Theory, Architectures and Applications*, pp. 35–61, 1995.

[16] Fabien Ringeval, Björn Schuller, and Michel et al. Valstar, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[17] Florian Eyben, Klaus R Scherer, and Björn W et al. Schuller, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[18] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] Jian Huang, Ya Li, and Jianhua et al. Tao, "Multimodal continuous emotion recognition with data augmentation using recurrent neural networks," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 57–64.

[20] I Lawrence and Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[21] Michel Valstar, Jonathan Gratch, and Björn et al. Schuller, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*. ACM, 2016, pp. 3–10.