Multimodal Transformer for Automatic Depression Estimation System

Dang-Khanh Nguyen[1], Guee-Sang Lee[1], Soo-Hyung Kim[1], Hyung-Jeong Yang[1,4], Seung-Won Kim[1], Min Jhon[2], and Joo-Wan Kim[3]

[1] Department of AI Convergence, Chonnam National University, Gwangju, Republic of Korea
[2] Department of Psychiatry, Chonnam National University Hwasun Hospital, Gwangju, Republic of Korea
[3] Department of Psychiatry, Chonnam National University Hospital, Gwangju, Republic of Korea
[4] Corresponding Author: hjyang@jnu.ac.kr

Abstract. People are under increasing pressure as the pace of work and life quickens, increasing their chances of developing depression. Due to its severe consequence, a specific effort is spent to diagnose this mental disorder as soon as possible. Machine learning and deep learning are expected to support clinicians in detecting depressed subjects via visual, text, and audio data. In this paper, we exploit the useful information from multiple modalities by utilizing transformer-based fusion to handle depression diagnosis. We conduct the experiment on a depression dataset, D-Vlog, in order to examine our deep-learning model. The promising results create a foundation for the further development of the Automatic Depression Estimation System.

Keywords: Depression Recognition · Transformer fusion · Multimodal fusion

1 Introduction

World Health Organization predicts depression can be the most widespread mental disorder by 2030 [1]. It affects seriously the quality of life and probably leads to some physical or mental illness. There is a certain probability that the severely depressed subject can commit suicide [2]. Consequently, detecting depression is a crucial topic that requires thorough knowledge and experience. However, it is also subjective and time-consuming [3].

To support clinicians in detecting depressed subjects, an automatic depression estimation (ADE) system is used to collect audiovisual data from the target and diagnose the level of depression. This discriminative system is expected to improve the speed and accuracy of professionals in early detecting depression. Recently, there has been a noticeable amount of research developing machine learning models to exploit the audiovisual clues in the ADE systems.

In this research, we would like to leverage the power of deep learning, particularly, multimodal transformer, in order to diagnose depression via subjects'

information, such as acoustic clues, visual data, and interview transcripts. Moreover, we contribute one more architecture in the taxonomy of [6], named hierarchical cross-attention to concatenation. This idea is expected to fully exploit the information of multimodal to accomplish higher performance compared to the traditional approaches. To evaluate our method, we use a recent depression detection benchmark, D-Vlog, Depression V-log from the Youtube platform.

2 Related works

There is a significant effort on developing the transformer [12] as a convolution-free machine learning model to achieve various tasks. Vision Transformer [13] is devised to handle image classification tasks and other computer vision problems while Audio Spectrogram Transformer [14] works with audio signal input. On the other hand, Tsai [7] uses a transformer as a mechanism to fuse the embeddings from multiple modalities. Shvetsova [8] combines multimodal transformers with contrastive loss to attain an impressive result in the video retrieval task.

Yoon et. al. [4] collect an audiovisual dataset for depression classification and introduce a baseline model using multiple levels of transformer encoder layers. Initially, they apply self-attention for each modality followed by cross-attention transformer layers between audio and visual features. Afterward, these features are concatenated in temporal dimension and fed into multimodal transformer layers. A global average pooling is used to get the representation of the whole sample and a fully-connected layer is employed to generate the final prediction.

EATD-Corpus [5] is a Chinese depression database recording interviews with some depressed and non-depressed subjects. The audio and transcript are collected and an SDS score [17] of each subject is recorded. The authors also provide a recurrent neural network baseline model exploiting the sequence information of each modality. ELMo [15] extracts sentence embeddings from raw transcript while NetVLAD [16] is used to obtain acoustic features from Mel spectrograms. A simple concatenation followed by a fully-connected layer is applied to fuse the attributes of text and audio modalities.

3 Transformer-based Fusion Methods

Our exploration of multimodal fusion is inspired by the taxonomy of Xu in [6]. Based on the research, we implement three versions of the transformer-based networks: early concatenation, hierarchical attention, and cross-attention to concatenation. Additionally, we contribute a combined interaction, which is hierarchical cross-attention to concatenation. The illustration of these architectures is shown in Fig. 1.

Early concatenation is a naive approach where the embeddings from two modalities are concatenated in the temporal dimension. A classification token is prepended then the sequence is fed into multiple transformer encoder layers to jointly learn the sequence information and the interaction between two modalities. After the multimodal transformer, the classification token is considered as

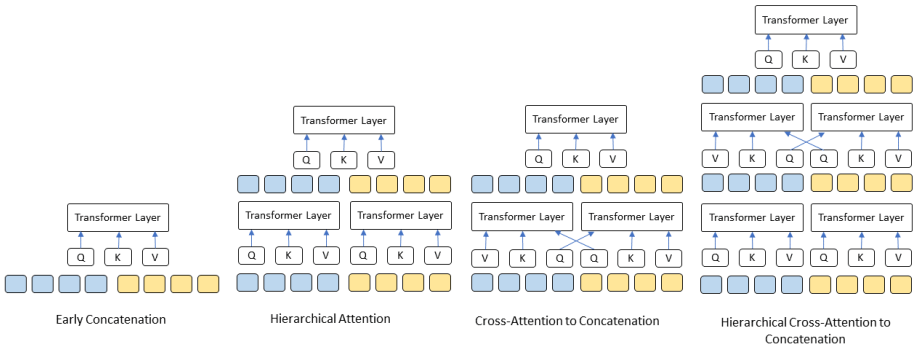

Fig. 1. Multimodal Transformers. "Q", "K", and "V" stand for query, key, and value embedding, respectively

the representation of the whole sample and a multilayer perceptron is used to generate the output for a specific task.

Hierarchical Attention and Cross-attention to Concatenation are two extended versions of the early concatenation, where self-attention and cross-attention are devised, respectively. In hierarchical attention, intra-modality information is learned by the attention mechanism via encoder layers. On the other hand, cross-attention exploits inter-modality knowledge. In both approaches, the multimodal transformer is utilized similar to the early concatenation concept.

The final approach, hierarchical cross-attention to concatenation, is our proposal for this paper. Firstly, we apply self-attention for acoustic and visual features separately. After learning the intra-modality knowledge, the model applies cross-attention where "query" comes from the source modality, "key" and "value" come from the target modality. After finishing exchanging inter-modality information, the features of the two modalities are concatenated in temporal dimension and fused by the multimodal transformer. Instead of using global average pooling, we use a classification token to acquire the representation of the video. A multilayer perceptron is adapted to generate the prediction logit.

4 Experiments and results

4.1 Dataset

Yoon et. al. introduce an audiovisual dataset for depression classification. The samples are collected from Youtube videos by searching for vlog-related keywords, such as: 'depression vlog', 'daily vlog', 'depression journey', etc. The videos are then extracted into acoustic and visual features. To obtain audio modality, OpenSmile [10] toolkit and eGeMAPS [11] are employed. The sound is re-sampled with a frequency of 1Hz and the final acoustic features comprised of spectral flux, loudness, MFCCs (Mel-frequency cepstral coefficients), etc. Re-

garding the visual modality, dlib [9] open-source software is utilized to extract facial landmarks.

The videos are re-sampled at a frequency of 1Hz for both modalities so the two modalities are aligned in terms of time. As a result, the dimension of visual and acoustic features are $t \times 136$ and $t \times 25$, respectively, where t is the length of the video in seconds. Moreover, the number of depressed and non-depress samples in D-Vlog dataset are 555 and 406, respectively Therefore, unlike the former depression databases, it does not suffer from the imbalance issue. The dataset is split into train, validation, and test set with a ratio of 7:1:2.

4.2 Experiment setting

Pytorch framework is used for our implementation. The embedding dimension and feed-forward dimension in transformers are both set to 256. The number of attention heads and encoder layers is equal to 8 and 4, respectively. During the training process, we chose a learning rate of $1e-5$ and a drop-out probability of 0.1. For each network configuration described in section 3, we trained the model in 10 epochs with a repetition of 10, and the average scores of 10 experiments were recorded.

For each sample in the dataset, we downsample the audio and visual features in the temporal dimension with the rate of 4 and different offsets. By this processing, we create 4 new shortened versions with a length of one-fourth compared to the original samples. The downsampling augmentation not only decreases the complexity of the training by shortening the sequence length but also creates more samples for training the machine learning models.

4.3 Results

For D-Vlog benchmarks, the weighted average F1-score, recall, and precision are used for evaluation. Unsurprisingly, the hierarchical cross-attention to concatenation achieves the highest evaluation scores on the validation set of D-Vlog. Using self-attention or cross-attention separately is not an optimized setting, compared to early concatenation only. The detailed measures of each model are listed in Table 1.

Table 1. Evaluation scores of each model on D-Vlog validation set

Model	F1-score	Recall	Precision
Early Concatenation	63.92	64.71	64.58
Hierarchical Attention	62.32	63.14	63.44
Cross-Attention to Concatenation	61.36	62.55	62.87
Hierarchical Cross-Attention to Concatenation	**64.70**	**65.29**	**65.66**

After evaluating these transformer-based fusions on the validation set, we use the best model of each configuration and run it with the test split of D-Vlog.

Similarly, the hierarchical cross-attention to concatenation attains the highest F1 score and precision, compared to the baseline and other network settings. However, the baseline performs the best recall measure among the fusion methods. The weighted average F1 score, recall, and precision of these models are shown in Table 2.

Table 2. Evaluation scores of each model on D-Vlog test set

Model	F1-score	Recall	Precision
Depression Detector [4]	63.50	**65.57**	65.40
Early Concatenation	61.97	63.21	62.43
Hierarchical Attention	62.20	62.74	62.15
Cross-Attention to Concatenation	63.08	63.21	63.00
Hierarchical Cross-Attention to Concatenation	**63.86**	63.68	**65.83**

5 Conclusion

In this paper, we revised the multimodal fusion techniques that utilize the multi-head attention mechanism, especially, the multimodal transformer. By conducting the experiments on various network settings, we concluded that the combination of self-attention, cross-attention, and multimodal transformer can boost the performance of the model on a specific task, particularly, depression recognition. However, concatenating the modality feature in the temporal dimension can lengthen the sequence length. Consequently, the resources such as computational time and memory will increase dramatically due to the quadratic complexity of the pairwise attention with token sequence length. In the future, we will resolve this bottleneck to achieve better performance of the transformer-based fusion model.

Acknowledgements This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT). (NRF-2020R1A4A1019191).

References

1. Mathers, Colin D., and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030." PLoS medicine 3, no. 11 (2006): e442.
2. Kessler, Ronald C., Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R. Merikangas, A. John Rush, Ellen E. Walters, and Philip S. Wang. "The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R)." Jama 289, no. 23 (2003): 3095-3105.

3. Maj, Mario, Dan J. Stein, Gordon Parker, Mark Zimmerman, Giovanni A. Fava, Marc De Hert, Koen Demyttenaere, Roger S. McIntyre, Thomas Widiger, and Hans-Ulrich Wittchen. "The clinical characterization of the adult patient with depression aimed at personalization of management." World Psychiatry 19, no. 3 (2020): 269-293.
4. Yoon, Jeewoo, Chaewon Kang, Seungbae Kim, and Jinyoung Han. "D-Vlog: Multimodal Vlog Dataset for Depression Detection." Proceedings of the AAAI Conference on Artificial Intelligence (2022).
5. Shen, Ying, Huiyu Yang, and Lin Lin. "Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6247-6251. IEEE, 2022.
6. Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." arXiv preprint arXiv:2206.06488 (2022).
7. Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. "Multimodal transformer for unaligned multimodal language sequences." In Proceedings of the conference. Association for Computational Linguistics. Meeting, vol. 2019, p. 6558. NIH Public Access, 2019.
8. Shvetsova, Nina, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S. Feris, David Harwath, James Glass, and Hilde Kuehne. "Everything at Once-Multi-Modal Fusion Transformer for Video Retrieval." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20020-20029. 2022.
9. King, Davis E. "Dlib-ml: A machine learning toolkit." The Journal of Machine Learning Research 10 (2009): 1755-1758.
10. Eyben, Florian, 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, pp. 1459-1462. 2010.
11. Eyben, Florian, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE transactions on affective computing 7, no. 2 (2015): 190-202.
12. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).
13. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
14. Gong, Yuan, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer." arXiv preprint arXiv:2104.01778 (2021).
15. Neumann, ME Peters M., M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).
16. Arandjelovic, Relja, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. "NetVLAD: CNN architecture for weakly supervised place recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297-5307. 2016.
17. Zung, William WK. "A self-rating depression scale." Archives of general psychiatry 12, no. 1 (1965): 63-70.