

Hybrid network feature extraction for depression assessment from speech

Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Zhao, Ziping, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn Schuller. 2020. "Hybrid network feature extraction for depression assessment from speech." In Proceedings: Interspeech 2020, 25-29 October 2020, Shanghai, edited by Helen Meng, Bo Xu, and Thomas Zheng, Online-Ressource, 4956-60. ISCA Archive. <https://doi.org/10.21437/interspeech.2020-2396>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under the following conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publizieren>





Hybrid Network Feature Extraction for Depression Assessment from Speech

Ziping Zhao^{1*}, Qifei Li^{1,2*}, Nicholas Cummins^{3,4}, Bin Liu²
Haishuai Wang^{5,1}, Jianhua Tao², Björn W. Schuller^{1,3,6}

¹College of Computer and Information Engineering, Tianjin Normal University, China

²National Laboratory of Pattern Recognition, CASIA, Beijing, China

³Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

⁴Department of Biostatistics and Health Informatics, IoPPN, King's College London, London, UK

⁵Department of Computer Science and Engineering, Fairfield University

⁶GLAM – Group on Language, Audio, & Music, Imperial College London, UK

zhaoziping@tjnu.edu.cn

Abstract

A fast-growing area of mental health research is the search for speech-based objective markers for conditions such as depression. One vital challenge in the development of speech-based depression severity assessment systems is the extraction of depression-relevant features from speech signals. In order to deliver more comprehensive feature representation, we herein explore the benefits of a hybrid network that encodes depression-related characteristics in speech for the task of depression severity assessment. The proposed network leverages self-attention networks (SAN) trained on low-level acoustic features and deep convolutional neural networks (DCNN) trained on 3D Log-Mel spectrograms. The feature representations learnt in the SAN and DCNN are concatenated and average pooling is exploited to aggregate complementary segment-level features. Finally, support vector regression is applied to predict a speaker's Beck Depression Inventory-II score. Experiments based on a subset of the Audio-Visual Depressive Language Corpus, as used in the 2013 and 2014 Audio/Visual Emotion Challenges, demonstrate the effectiveness of our proposed hybrid approach.

Index Terms: depression detection, DCNN, self-attention, complementary features

1. Introduction

Major depressive disorders (MDDs) are highly prevalent in modern society [1]. Early interventions, aimed at predicting the onset of MDD, are an essential means of helping to reduce this burden. As an attempt to aid in MDD diagnosis, the problem of automatically detecting and monitoring depression through speech signal analysis has recently attracted considerable attention [2]. This work focuses on the use of the audio domain in order to estimate the clinical depression score of an individual, given an (often lengthy) audio recording of their voice.

A considerable challenge currently being faced by researchers in the field of depression analysis is that of how best to extract discriminative, robust, and depression-salient features from the acoustic content of a speech signal. As in most areas of intelligent signal analysis, deep neural networks have become the predominant approach to automated depression analysis for discriminative representation learning [3, 4]. Many recent works in this field have leveraged either recurrent neural networks (RNNs) or deep convolutional neural networks (DCNNs) as feature extractors, with varying degrees of success [5, 6, 7, 8, 9].

*Equal Contribution

There has been increasing interest in incorporating both RNNs and DCNNs into a single architecture in order to capture both long-term and local dependencies [10, 11]. In [7], for example, a hybrid network combining CNN and long short-term memory (LSTM) RNN, was employed for speech-based depression severity assessment; results demonstrated that this set-up is able to outperform a CNN-only system. Similarly, in [12], a system comprising a gated convolutional neural network (GCNN) followed by an LSTM layer demonstrates the effectiveness of such hybrid methods for depression recognition.

However, a major drawback of including RNNs in such scenarios is that they are difficult to parallelise and not time-efficient [13]. Furthermore, even with the addition of memory cell structures such as LSTM, RNNs struggle to capture long-range dependencies [14]. This issue becomes particularly pronounced in depression analysis, where files can be upwards of 15 minutes in length [2, 15].

Self-attention networks (SAN) [16], which utilise the attention mechanism as the basic building block, can help in capturing long-term contextual dependencies. SANs have demonstrated their ability to capture contextual dependencies in several natural language processing (NLP) tasks [16, 17, 10, 18] and, more recently, have produced state-of-the-art SER results [19]. Despite having certain advantages over RNNs, such as faster training and inference times due to parallelisable computation and a lower number of trainable parameters, the suitability of SAN for identifying depression through speech analysis remains understudied.

Herein, we propose a hybrid network based on a SAN and a DCNN to extract complementary features from acoustic low-level descriptors (LLDs) and 3D log Mel spectrograms respectively. The core idea behind the SAN is to model the inner dependencies between elements with different positions in the learned feature sequence, which should enhance the learning of depression-salient information. To the best of the authors' knowledge, this is the first time that SAN has been deployed for the task of speech-based depression severity assessment.

Through the action of the DCNN, the hybrid network becomes able to retain high resolution of temporal structure in feature learning. Inspired by the positive results of 3D log Mel spectrum features in SER [20, 21], we employ log-Mel, deltas, and delta-deltas as 3D input to the CNN model. We utilise this 3D input as the delta; moreover, the delta-delta features are capable of effectively capturing the effects of depression in speech [22], while also being less susceptible to the impact of non-relevant acoustic factors.

Our two main contributions can be summarised as follows: i) We propose a hybrid network that combines self-attention networks with DCNN to produce discriminative representation for depression analysis from speech; ii) We conduct extensive experiments that demonstrate the effectiveness of this approach.

2. Proposed Method

Our proposed hybrid model consists of three components (Figure 1). The first is the *SAN model*, trained on low-level acoustic features; the second is the *DCNN model*, which we train to capture the spatial information from 3D log-Mel spectrograms; the third component is the *depression severity prediction module*, in which average pooling is applied to aggregate complementary features into utterance-level features, which are in turn utilised as the input of a support vector regressor (SVR) to predict the depression score (Figure 2). The remainder of this section covers two key aspects of our model, namely the SAN model (Section 2.1) and the generation of the 3D log-Mel spectrograms (Section 2.2).

2.1. Self-attention network

Self-attention is an attention technique based on an encoder-decoder structure that does not employ any form of recurrence. Instead, it uses weighted correlations between the elements of the input sequence [16]. In this paradigm, the encoder maps an input sequence into several attention matrices, while the decoder uses these matrices to generate a new output token. The *Transformer*, the model that uses *self-attention*, has been demonstrated to achieve state-of-the-art performance in several NLP tasks, with a computing cost that is one or two orders of magnitude (depending on the size of the model) lower than that of conventional RNNs [17, 10, 18]. Note that this section focuses only on the implementation of the encoder, as a decoder is not required in our proposed hybrid network.

Self-attention calculates queries, keys (properties of the input) and values (the output) for the frames in a given hidden sequence H through linear transformation of the input sequence X , as follows:

$$Q = W_q X; K = W_k X; V = W_v X, \quad (1)$$

where the matrices Q, K, V denote the set of queries, keys and values of an input/output sequence, while W_q, W_k, W_v represent the learnt linear operations. A scaled dot-product operation is performed on the query and key to obtain the similarity weights, which are then normalised by the softmax function. The attention matrix is calculated as follows:

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where d_k is a scaling factor, set as the dimensionality of K .

The Z is the attention matrix ($N \times d_k$), where N is the number of elements in the input sequence. In order to obtain one-dimensional feature representations $\{s_1, s_2, \dots, s_{d_k}\}$, we perform an additional operation in the last self-attention block: i. e., the *SA1* of Figure 1 and s_j can be calculated as follows:

$$s_j = \frac{\sum_{i=1}^N Z_{i,j}}{N}, j = 1, 2, \dots, d_k. \quad (3)$$

Compared to multi-head attention, the advantages of single-head self-attention are its lower memory usage and fewer hyper-parameters [23]. Accordingly, in this work, we apply two separated single-head self-attention blocks rather than a two-head self-attention model.

2.2. 3D Log-Mels Spectrogram Generation

In recent years, applying CNN to capture information in the spectrograms for depression detection purposes has achieved excellent performance [9, 24]. However, static spectrograms can contain personalised information about the speaker, which can negatively influence the performance of depression detection from speech [21, 22]. Inspired by the successful use of 3D log-Mel spectrograms in SER [20], our hybrid system also uses log-Mels, together with deltas and delta-deltas, as the input to the DCNN.

First, we split raw speech signal into short frames with Hamming windows of 25ms and a 10ms shift. The power spectrum for each frame is then calculated and passed through the Mel-filter bank i to produce output p_i . A logarithmic operation is then conducted on p_i to obtain the log-mels spectrogram m_i . Finally, we calculate the m_i^d feature, which is the deltas of m_i via formula (4), while the value of N is set to 3. Similarly, the delta-deltas features m_i^{dd} are calculated by taking the derivative of the deltas, as shown in Equation (5).

$$m_i^d = \frac{\sum_{n=1}^N n(m_{i+n} - m_{i-n})}{2 \sum_{n=1}^N n^2} \quad (4)$$

$$m_i^{dd} = \frac{\sum_{n=1}^N n(m_{i+n}^d - m_{i-n}^d)}{2 \sum_{n=1}^N n^2} \quad (5)$$

After completing the above calculations, we obtain a three-dimensional feature representation $X \in R^{t \times f \times c}$ as the input of the DCNN model, where t denotes the length of frame, while f for the number of Mel-filter banks. In our work, f is set to 80, and c is 3, representing the static, deltas and delta-deltas log-mels spectrogram respectively.

3. Experiments and Results

To demonstrate the effectiveness of the proposed methods, we performed a set of experiments on the AVEC 2013 [25] and AVEC 2014 [26] depression databases reported next.

3.1. Experimental Corpus

The AVEC 2013 depression dataset includes 340 video recordings of 292 subjects performing human-computer interaction tasks while being recorded by audio-visual sensors. The average subject age is 31.5 years with a range of 18 to 63 years. The length of each recording varies from 20 to 50 minutes, with an average duration of 25 minutes per recording. The total duration of all recordings is 240 hours. The 16-bit audio was recorded at a sampling rate of 41KHz. More detailed information regarding the AVEC 2013 depression corpus is presented in Table 1.

We also use the AVEC 2014 depression database for our evaluation, which is a subset of the AVEC 2013 depression corpus. This subset comprises 150 videos of task-oriented depression data recorded in a human-computer interaction scenario. The total number of subjects is 84, ranging from 18 to 63 years in age.

The level of depression in the AVEC 2013 and AVEC 2014 depression datasets is labelled with a single value per recording using a standardised self-assessed subjective depression questionnaire, the Beck Depression Inventory-II (BDI-II) [27], within the range [0, 63]. As BDI-II value prediction is a regression task, the accuracy metric for the challenge is the *Root Mean Square Error* (RMSE) and *Mean Absolute Error* (MAE).

3.2. Features

In this paper, we use 3D log-mels spectrogram as the input of DCNN and LLDs for the self-attention network. The 3D feature

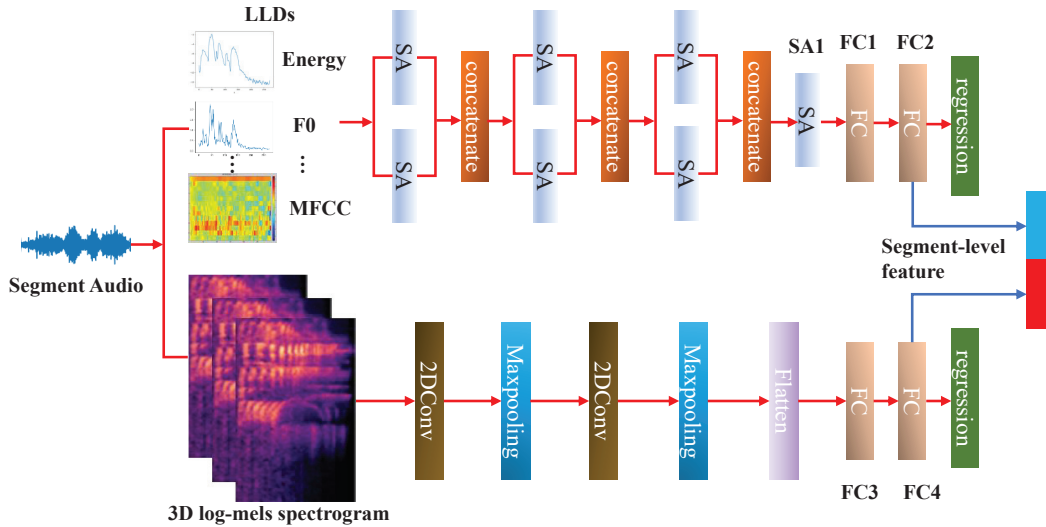


Figure 1: Frame of the proposed hybrid network for extracting segment-level complementary features. SA and FC provide the self-attention and the fully connected layer respectively.

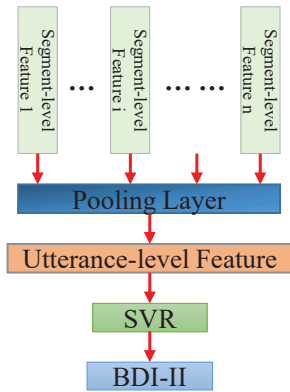


Figure 2: The process of aggregating features and detecting depression levels.

(the static, deltas and delta-deltas of the log-Mel spectrum from 40 filterbanks) is normalised by the global mean and the standard deviation and then fed into the DCNN model. Meanwhile, as a commonly used feature set for the task of speech emotion recognition, the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [28] is extracted using the openSMILE toolkit [29] with a window size of 25 ms and a 10 ms stride, then fed into the self-attention network.

3.3. Model Parameters

As shown in Fig. 1, there are two key components in our proposed hybrid network, namely the DCNN and the self-attention model. For the DCNN model, there are 32 kernels in the first CNN layer and 64 kernels in the second CNN layer. All of these kernel sizes are 3×3 and the padding mode is valid. After each CNN layer, there is a max pooling layer with a size of 2×2 . The number of neurons in the two fully connected layers is 1024. For the self-attention model, there are 512 neurons in each fully connected layer. Moreover, after every CNN layer and self-attention module, there is a batch normalization layer

Table 1: Summary of AVEC 2013 Depression Corpus.

Partition	Number ^a	Max ^b	Min ^c	Average ^d	Score Range
Train	50	27min20s	8min5s	14min20s	0-44
Dev.	50	23min55s	14min20s	14min20s	0-45
Test	50	23min57s	5min15s	15min57s	N/A

^a Number denotes the number of files.

^b Max denotes maximum length.

^c Min denotes minimum length.

^d Average denotes average length.

and *relu* activation function. The *FC1* and *FC3* are followed by a dropout layer in which the dropout rate is 0.5.

Note that these two components are trained separately and the objective functions are both RMSE. When the training of these two components is complete, the outputs of *FC2* and *FC4* are concatenated to create the segment-level complementary features.

Due to the smaller size of the available corpora, we also use an augmentation method that sets a window with a size of 500 frames and 50% overlap to divide the raw speech in the training set into segment-level speech in this work.

3.4. Results and Discussion

In this section, we present the experimental results of different pooling strategies, then compare the performance of our method with that of previous work.

3.4.1. Performance comparison of different pooling strategies

Firstly, a comparison of the results of the two pooling strategies (i. e. average-pooling and max-pooling) on the test set of AVEC 2013 and AVEC 2014 is presented in Table 2 and Table 3 in terms of RMSE and MAE. We can observe that our proposed hybrid network consistently achieves the best performance with average-pooling on both datasets. Moreover, the performances of SAN on both datasets are better than that of the Bidirectional LSTM RNN (BLSTM)-based approach, regardless of which kind of pooling strategy was used in our hybrid model. This re-

Table 2: Performance comparison of different pooling strategies on the test set of AVEC 2013.

Method	Average-pooling		Max-pooling	
	RMSE	MAE	RMSE	MAE
BLSTM	10.64	8.22	10.94	8.43
DCNN	10.30	7.83	10.90	8.76
BLSTM-DCNN	10.41	8.14	10.90	8.74
SAN	9.85	7.55	10.28	8.08
Proposed method	9.65	7.38	10.82	8.65

Table 3: Performance comparison of different pooling strategies on the test set of AVEC 2014.

Method	Average-pooling		Max-pooling	
	RMSE	MAE	RMSE	MAE
BLSTM	10.45	8.41	10.69	8.67
DCNN	10.28	8.45	10.73	8.91
BLSTM-DCNN	10.18	8.33	10.71	8.81
SAN	9.87	8.14	10.54	8.62
Proposed method	9.57	7.94	10.63	8.77

sult supports our hypothesis that SAN can be more helpful than the BLSTM RNN model in depression analysis. Furthermore, it has been demonstrated that the combination of DCNN and SAN features outperforms each feature taken separately when average-pooling was employed in our hybrid model. This observation demonstrates the effectiveness of our proposed model when applied to the task of speech-based depression analysis.

Considered as a whole, the performance of average-pooling is obviously better than that of max-pooling. However, the experimental results of our proposed model are slightly lower than those of SAN when max-pooling was used to aggregate the segment-level complementary features. The main reason for this is that the maximum value of the segment-level features are unable to fully and explicitly represent the segment-level information. Thus, the average-pooling strategy, which performs well at modeling the relationship between segment-level features and utterance-level features, is more suitable for the task of speech-based depression detection.

3.4.2. Performance comparison with other models

The effectiveness of our hybrid framework can be highlighted through comparison with other key results obtained on the AVEC 2013 and AVEC 2014 depression corpora in the literature (Table 4 and Table 5). It can be observed that the best MAE (7.38) and RMSE (9.65) on the AVEC 2013 test set, as well as the best MAE (7.94) and RMSE (9.57) on the AVEC 2014 test set, were achieved by our proposed hybrid network.

Furthermore, we observed that, for the test sets of both the AVEC 2013 and AVEC 2014 depression corpora, our proposed model outperforms the AVEC 2013 or AVEC 2014 audio baseline. Moreover, as shown in Table 2 and Table 3, although the performance of the BLSTM, DCNN, SAN and BLSTM-DCNN models is inferior to that of the proposed hybrid model, it still surpasses that of the AVEC 2013 or AVEC 2014 audio baseline and the multi-modal work proposed in [30, 31] in terms of RMSE and MAE. Meanwhile, our best results are superior even to those obtained by the AVEC 2014 Audio-Video baseline [26]

for RMSE.

Table 4: Performance comparison between the proposed model and other models on the test set of AVEC 2013.

Methods	RMSE	MAE
AVEC 2013 Audio Baseline [25]	14.12	10.35
PLS regression [30]*	10.96	8.72
DCNN [9]	10.00	8.20
CNN-LSTM-SVR [7]	9.79	7.48
Proposed method	9.65	7.38

* Indicates a multimodal system was utilised.

Table 5: Performance comparison between the proposed model and other models on the test set of AVEC 2014.

Methods	RMSE	MAE
AVEC 2014 Audio Baseline [26]	12.56	10.03
AVEC 2014 Audio-Video Baseline [26]	9.89	7.89
Fisher Vector Encoding [31]*	10.25	8.40
DCNN [9]	9.99	8.19
CNN-LSTM-SVR [7]	9.66	8.02
Proposed method	9.57	7.94

* Indicates a multimodal system was utilised.

Among all the works compared here, the study that most closely resembles our proposed model was presented by Niu et al. [7]. In their work, a hybrid model similar to our method was also proposed for the task of speech-based depression analysis. However, these authors only used the Mel Frequency Cepstrum Coefficient (MFCC) as the input to the hybrid network, which may have resulted in a lot of valuable information contained in the speech signal being missed.

4. Conclusions

In this work, we presented a novel hybrid network, which combines DCNN and self-attention networks, for the task of depression severity detection. This approach is highly suitable for speech-based depression detection, as it uses a hybrid framework capable of taking advantage of DCNN and SAN. Experimental results achieved on the AVEC 2013 and AVEC 2014 depression datasets verified the suitability of this approach and demonstrate that self-attention is a better building block compared to recurrence when conducting depression analysis from speech. Future work will explore the effectiveness of proposed hybrid network in other speech-related tasks.

5. Acknowledgements

The work presented in this paper was substantially supported by the Key Program of the National Natural Science Foundation of China (Grant No. 61831022), the National Natural Science Foundation of China (Grant No. 61702370, 61771472), the Key Program of the National Science Foundation of Tianjin (Grant No. 18JCZDJC36300), the technology plan of Tianjin (Grant No: 18ZXRHSY00100). This project also received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902, which receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

6. References

- [1] World Health Organisation, “The European Mental Health Action Plan 2013–2020,” <http://www.euro.who.int>, 2015.
- [2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, July 2015.
- [3] N. Cummins, A. Baird, and B. W. Schuller, “The increasing impact of deep learning on speech analysis for health: Challenges and opportunities,” *Methods, Special Issue on on Translational Data Analytics and Health Informatics*, vol. 151, pp. 41–54, 2018, (IF: 3.998, 5-year IF: 3.936 (2017)).
- [4] N. Cummins, F. Matcham, J. Klapper, and B. Schuller, “Artificial intelligence to aid the early detection of mental illness,” in *Artificial Intelligence in Precision Health*, 1st ed., D. Barh, Ed. London, U.K.: Elsevier Academic Press, 2020, ch. 10, pp. 231–256.
- [5] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, “Hybrid depression classification and estimation from audio video and text information,” in *Proc. 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Mountain View, CA, USA, 2017, pp. 45–51.
- [6] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, “Multimodal measurement of depression using deep learning models,” in *Proc. 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Mountain View, California, USA, 2017, pp. 53–59.
- [7] M. Niu, J. Tao, B. Liu, and C. Fan, “Automatic depression level detection via lp-norm pooling,” in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 4559–4563.
- [8] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, “Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, April 2020.
- [9] L. He and C. Cao, “Automated depression analysis using convolutional neural networks from speech,” *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, July 2018.
- [10] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, “Disan: Directional self-attention network for rnn/cnn-free language understanding,” in *Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana, USA, 2018.
- [11] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, “Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition,” *IEEE Access*, vol. 7, pp. 97 515–97 525, July 2019.
- [12] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, “Multimodal fusion of bert-cnn and gated cnn representations for depression detection,” in *Proc. 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, Nice, France, 2019, pp. 55–63.
- [13] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones, “Character-level language modeling with deeper self-attention,” in *Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, Hawaii, USA, 2019, pp. 3159–3166.
- [14] T. Zhang, P. Zhao, Y. Liu, V. Sheng, J. Xu, D. Wang, G. Liu, and X. Zhou, “Feature-level deeper self-attention network for sequential recommendation,” in *Proc. 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, Macao, China, 2019, pp. 4320–4326.
- [15] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “AVEC 2017 – Real-life depression, and affect recognition workshop and challenge,” in *Proc. 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Mountain View, CA, 2017, pp. 3–9.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. 31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.
- [17] T. Scialom, B. Piwowarski, and J. Staiano, “Self-attention architectures for answer-agnostic neural question generation,” in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 6027–6032.
- [18] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, “Beyond rnns: Positional self-attention with co-attention for video question answering,” in *Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, Hawaii, USA, 2019, pp. 8658–8665.
- [19] L. Tarantino, P. N. Garner, and A. Lazaridis, “Self-attention for speech emotion recognition,” in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2578–2582.
- [20] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3d log-mel spectrograms with deep learning network,” *IEEE Access*, vol. 7, pp. 125 868–125 881, Aug. 2019.
- [21] M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, July 2018.
- [22] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, “Analysis of acoustic space variability in speech affected by depression,” *Speech Communication*, vol. 75, pp. 27–49, Dec. 2015.
- [23] S. Merity, “Single headed attention rnn: Stop thinking with your head,” *arXiv preprint arXiv:1911.11423*, 2019.
- [24] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. Messner *et al.*, “AVEC 2019 workshop and challenge: State-of-mind, depression with ai, and cross-cultural affect recognition,” in *Proc. 9th International Audio/Visual Emotion Challenge and Workshop (AVEC)*, Nice, France, 2019.
- [25] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proc. 3rd International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Barcelona, Spain, 2013, pp. 3–10.
- [26] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proc. 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Orlando, Florida, USA, 2014, pp. 3–10.
- [27] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, “Comparison of beck depression inventories-ia and-ii in psychiatric outpatients,” *Journal of Personality Assessment*, vol. 67, no. 3, pp. 588–597, June 1996.
- [28] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “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, Apr. 2016.
- [29] F. Eyben, F. Wengler, F. Gross, and B. Schuller, “Recent developments in openSMILE, the munich open-source multimedia feature extractor,” in *Proc. the 21st ACM International Conference on Multimedia (ACM MM)*, Barcelona, Spain, 2013, pp. 835–838.
- [30] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, “Depression recognition based on dynamic facial and vocal expression features using partial least square regression,” in *Proc. 3rd International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Barcelona, Spain, 2013, pp. 21–30.
- [31] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, “Depression estimation using audiovisual features and fisher vector encoding,” in *Proc. 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Orlando, Florida, USA 2014, pp. 87–91.