Regular Paper

Speech-linguistic Multimodal Representation for Depression Severity Assessment

Mariana Rodrigues Makiuchi[a)] Tifani Warnita[1,b)] Kuniaki Uto[1,c)] Koichi Shinoda[1,d)]

Abstract: Depression is a common, but serious mental disorder that affects people all over the world. Therefore, an automatic depression assessment system is demanded to make the diagnosis of this disorder more popular. We propose a multimodal fusion of speech and linguistic representations for depression detection. We train our model to infer the Patient Health Questionnaire (PHQ) score of subjects from the E-DAIC corpus. For the speech modality, we apply VGG-16 extracted features to a Gated Convolutional Neural Network (GCNN) and a LSTM layer. For the linguistic representation, we extract BERT features from transcriptions and input them to a CNN and a LSTM layer. We evaluated the feature fusion model with the Concordance Correlation Coefficient (CCC), achieving a score of 0.696 on the development set and 0.403 on the testing set. The inclusion of visual features is also discussed. The results of this work were submitted to the Audio/Visual Emotion Challenge and Workshop (AVEC 2019).

Keywords: depression detection, affective computing, deep learning, multimodal systems, BERT, gated CNN, CNN

1. Introduction

Depression is one of the most common mental disorders in the United States (US). In fact, according to the data collected from the 2017 National Survey on Drug Use and Health (NSDUH) [1], an estimate of 7.1% of all adults in the US had at least one major depressive episode. Although considered quite common all over the world and among a wide range of ages [2], this disorder cannot be neglected since it can cause severe and negative impacts. The abilities of a person in performing daily activities can be degraded and depression can result in undesirable effects in their thoughts, feelings and actions [3]. Therefore, the development of new methods and tools to support a fast and precise depression diagnosis is undoubtedly necessary.

In this regard, several studies [4], [5], [6] proposed computer-aided methods for an automatic and objective depression detection. This is important to reduce subjective biases, to popularize the diagnosis of this condition and to aid the diagnosis in complex situations, such as the ones presented by some elderly people [7].

Even though the automatic depression detection has been widely investigated from different perspectives, it is still considered a challenge [8].

In this paper, we present a multimodal approach for automatic depression detection that combines highly representative speech and textual features acquired with gated convolutional and convolutional neural network based models. Moreover, the proposed architectures used for the extraction of these features employ a Long Short-Term Memory (LSTM) layer in order to characterize the data's temporal behaviour. Our proposed multimodal model achieves the best result of 0.403 evaluated with the Concordance Correlation Coefficient (CCC) in the E-DAIC test partition.

The results presented in this paper [9] were submitted to the Audio/Visual Emotion Challenge and Workshop (AVEC) 2019 to compete on the Detecting Depression with AI Sub-Challenge.

2. Related Works

The main topics related to the method of automatic depression detection conducted in this work are:

- **Natural Language Processing (NLP)**: The recent expansion of the distributional vectors approach promoted by the Bidirectional Encoder Representations from Transformers (BERT)[10] made this model achieve the state-of-the art on eleven NLP tasks. Therefore, due to their powerful language representation, pre-trained BERT models were chosen to extract textual features from interview transcripts in this work.
- **Multimodality**: Promising results were acquired in numerous tasks, such as speech separation [11] and sound source localization, audio-visual action recognition and on/off screen audio source separation [12].
- **Depression detection**: There have been several works to develop an automatic depression detection based on visual features extracted from the body movement [5], [13], [14] as well as works that fused these visual features with audio information [4]. However, although these multimodal works consider audio and visual features for the depression assessment, the work presented in this paper differs from previous works, since we consider the semantic content of the patient's speech. In addition, in this work, the audio features were extracted with a VGG-16 [15] architecture and are not represented in a bag-of-audio approach as in [4].

[1] Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan
[a)] mariana@ks.cs.titech.ac.jp
[b)] tifaniwarnita@gmail.com
[c)] uto@c.titech.ac.jp
[d)] shinoda@c.titech.ac.jp

3. E-DAIC Corpus

The Extended Distress Analysis Interview Corpus (E-DAIC) [16] is based on audiovisual recordings of patients being interviewed by a virtual agent, which can be a Wizard-of-Oz (Woz) controlled by a human in another room or it can be fully automated. The E-DAIC includes the automatically transcribed transcripts of the interactions, the participants' audio files, their facial features and each patient's Patient Health Questionnaire [17] depression module (PHQ-8) score.

In the E-DAIC dataset, there are 275 subjects, that are divided into train, development and test partitions with 163, 56 and 56 subjects respectively. In the train and development sets, the interviews can happen in either the Woz or the AI setting, while, in the test set, there are only interviews conducted by the AI.

4. Evaluation Metric

The performance metric adopted in this work is the Concordance Correlation Coefficient (CCC) [18], defined as

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{1}$$

in which ρ is the Pearson correlation coefficient between variables x and y, σ_x and σ_y represent the standard deviations of x and y and μ_x and μ_y, their respective means.

The CCC is computed to measure the correlation between the prediction and the gold standard and it varies from -1 to 1, in which 1, -1 and 0 respectively indicate that the two variables are identical, exactly opposite and uncorrelated.

5. Proposed Methodology

Sections 5.1 to 5.3 present the single modalities models, summarized in **Fig. 1**. Section 5.4 presents a model for feature level fusion of these modalities. Finally, Section 5.5 introduces the baseline [8], to which our approach is compared in Section 6.

5.1 Audio Model

We use a deep spectrum representation extracted from a pre-trained VGG-16 network using spectrogram images as input. For the audio of each subject, it results in the deep spectrum features $\boldsymbol{X_i} \in \mathbb{R}^{T\times F}$, in which T represents the time dimension, that varies according to the duration of the speech data, and F denotes the feature dimension. We add zero paddings to the input features so that all input samples have the same length as the longest speech data duration.

5.1.1 GCNN-LSTM-based model

We have trained the GCNN-based model depicted in Fig. 1(a). The gated blocks in this model are represented in Fig. 1(d) and, for each of these blocks, the input to the max-pooling layer is defined as

$$\boldsymbol{Y} = \mathrm{conv}\,(\boldsymbol{X}, \boldsymbol{W}) \odot \mathrm{sigm}\,(\mathrm{conv}\,(\boldsymbol{X}, \boldsymbol{Z}))\,, \tag{2}$$

in which *conv* represents the convolution operation, *sigm* is the sigmoid activation function, $\odot$ is the Hadamard product between two tensors, X is the gated block's input and $\boldsymbol{W}$ and $\boldsymbol{Z}$ are the trainable parameters of the respective convolution layers.

For each gated block of the GCNN model in Fig. 1(a), the convolution filters are, from the input to the output, defined as $N = [512, 256, 256, 128, 128, 64, 64, 32, 32, 16]$. These ten gated blocks are followed by a 32-dimensional LSTM layer and a fully-connected layer with 512 hidden neurons. The GCNN-LSTM model is trained with the mean-squared error loss function and the Adam optimizer [19].

5.2 Textual Model

In this work, it was hypothesized that linguistic features would provide valuable information regarded to the subject's mental health condition. Thus, in order to represent the semantic content of the E-DAIC corpus interviews, textual features were extracted from the automatically transcribed transcripts [16] with the pre-trained BERT-large model [10].

The features were extracted from the last BERT layer and, for each subject, they can be represented as a matrix of size $K\times 1024$, in which K is the number of word tokens in the subject's transcript. In order to guarantee that all the samples input to the textual models would have the same size, a zero padding was conducted over the $K \times 1024$ feature matrices so that K would be always equal to the number of word tokens found in the longest transcript.

Although the BERT model represents textual data by analysing embeddings in a bidirectional manner, in this work, we hypothesize that there are remaining temporal correspondences at the last BERT layer since a feature array of size 1024 is generated for each word token.

5.2.1 CNN-LSTM-based Model

We opted for a model that combines CNN layers with one LSTM layer. This choice was founded on the observation that, since the BERT features are structured data in which it is possible to observe hierarchical patterns, it is natural to choose CNN layers to interpret the features extracted with BERT [20].

Thus we define seven convolution blocks, represented in Fig. 1(e), with different number of filters N for each convolution layer. These filters' size are, from the input to the output, $N = [128, 64, 64, 64, 64, 32, 32]$. The output of the last convolution block is then input to a 32-dimensional LSTM layer followed by a 256-dimensional fully-connected layer. The output of this fully-connected layer is then applied to a batch normalization, a ReLU activation function, a dropout layer with a dropout rate equal to 0.1 and a single dimensional fully connected layer, which outputs the prediction for the PHQ-8 score. A complete diagram of this model is shown in Fig. 1(b).

We consistently applied a batch size equal to 10, a learning rate equal to 10^{-3} and a loss function based on the CCC metric. The chosen optimizer is Adam [19] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$).

5.3 Visual Model

For the visual model, we utilize a deep visual representation extracted from a ResNet-50 model [21] as input. We choose the time dimension $T = 6000$ for the visual features and apply them to the model depicted in Fig. 1(c), in which, from the input to the output, the convolution filters of the gated blocks have size $N = [512, 256, 256, 128, 128, 64, 64, 32, 32, 16]$.

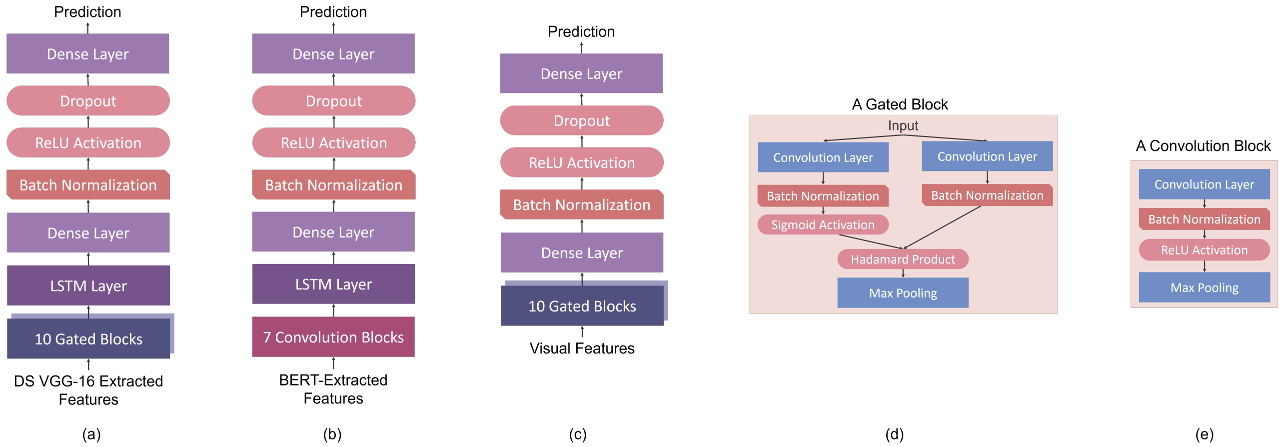

Fig. 1 Overview of the proposed models. (a) GCNN-LSTM model for speech-based depression assessment with nine gated convolution blocks and a LSTM layer. (b) CNN and LSTM-based model for depression assessment with textual features. (c) GCNN-based model for depression assessment with visual features. (d) A gated block. (e) A convolution block.

Similarly to the model presented in Section 5.1, we also use mean-squared error as the loss function and Adam as optimizer to train our network.

5.4 Fusion Model

We concatenate the embeddings obtained from the first dense layer of each modality in a single array, which is then input to the fusion network. This fusion model consists of a dense layer with 512 neurons, a batch normalization followed by a ReLU activation and a dropout layers and a final dense layer with a single neuron, that outputs the final PHQ score prediction. The model is trained to minimize the mean-squared error loss between the ground-truth PHQ score and the network prediction. Adam optimizer [19] is used.

5.5 Baseline Model

The baseline model [8] consists of a 64-dimensional Gated Recurrent Unit (GRU) layer with a dropout rate of 0.2 followed by a 64-dimensional fully-connected layer that outputs a single value as the PHQ score. The loss function used during the train is a CCC-based loss function and the batch size is 15.

6. Experiments and Results

Sections 6.2 and 6.1 present experiments conducted with the models described in Section 5 and Section 6.3 exposes the main results for the depression assessment task.

6.1 Number of CNN/GCNN blocks *1

The models presented in Sections 5.1.1 and 5.2.1 were trained by varying only the amount of GCNN and CNN blocks. Each model configuration was trained five times and the average and the maximum CCC on the development partition were acquired.

For the audio model presented in Section 5.1.1, models with 1 to 10 gated blocks were evaluated. The convolution filters' configuration for each tested model was defined in an ablation manner. Thus, the configuration defined in Section 5.1.1 for 10 gated blocks had its smaller filters removed one by one. Therefore, a 4 gated blocks configuration was defined as $N = [512, 256, 256, 128]$, for example. The results showed that the CCC is maximum when the number of blocks is equal to 10.

For the text model presented in Section 5.2.1, models with 1 to 12 CNN blocks were evaluated on the development partition and it was found that the model configuration with 8 blocks achieves the best maximum and average CCC values.

The convolution filters' configuration for the text model with 12 convolution blocks was defined as [128, 64, 64, 64, 64, 32, 32, 32, 16, 16, 8, 4] and, for models with less blocks, the configuration was determined in the same ablation manner as in the GCNN-LSTM audio model starting from the 12 blocks definition.

6.2 Different visual features *1

The model presented in Section 5.3 was tested with all the available visual features in the database (i.e., features extracted with VGG-16 and ResNet-50 architectures, Bag-of-Visual-Words and Facial Action Units). As in Section 6.1, models with 1 to 10 gated blocks were evaluated and the best CCC for each combination of input features and model configuration was calculated.

The results showed that the best model has 7 gated blocks and it uses VGG-extracted features as input. However, the CCC value acquired with this model, $CCC = 0.373$, does not greatly differ from the one acquired with the 10 gated blocks model that uses ResNet-extracted features (CCC = 0.372).

6.3 Results

The results are summarized in **Table 1**. The Concordance Correlation Coefficient (CCC) and the Root Mean Square Error (RMSE) metrics were calculated for unimodal and multimodal models on both the development and the test partitions.

From Table 1, it can be seen that the best model in both development and test sets and in both CCC and RMSE metrics is the model that fuses audio and text features. Moreover, it is possible to conclude that, in every situation, the fusion of features

*1 The experiments presented in this Section were conducted after the AVEC 2019 DDS submission. Thus, the models presented here were not evaluated in the test partition, but only in the validation set.

Table 1 Results evaluated with CCC and RMSE metrics for the development (i.e. validation) and test sets for audio, text, visual and feature-level fusion models respectively presented in Sections 5.1.1, 5.2.1, 5.3 and 5.4.

Modality	**Model**	**CCC**		**RMSE**	
		Development	Test	Development	Test
-	Challenge baseline [8]	0.336	0.120	5.03	6.37
Audio (A)	GCNN-LSTM	0.497	-	5.70	-
Text (T)	CNN-LSTM (7 blocks)	0.608	-	4.51	-
	CNN-LSTM (8 blocks)	0.685	-	4.22	-
Visual (V)	GCNN	0.372	-	5.74	-
Fusion	A and T (7 blocks)	**0.696**	**0.403**	**3.86**	**6.11**
	A, T (7 blocks) and V	0.624	-	4.86	-

performed by multimodal models gives better results when compared to the unimodal models that generated these features. Thus, these results confirm the premise that multiple modalities provide a richer characterization of reality when compared to single modalities representations for the task of depression assessment.

However, the combination of audio, text and visual features gives worse results when compared to the fusion of audio and text features only. This discrepancy might be explained from the fact that we used only a portion of the visual features extracted with the ResNet-50 architecture since applying all features to the models would be computationally costly.

From the difference between the test and development partitions results presented in Table 1, it can be concluded that the absence of a human conducting the interviewer as a virtual agent degraded the performance of the automatic depression diagnosis.

Finally, it can be observed that each one of the models presented in this paper outperforms the challenge baseline [8] when evaluated over the development partition with the CCC metric.

7. Conclusion

Multimodal approaches using audio, visual and text features for automatic depression detection were presented. The best results acquired in this paper, CCC = 0.696 for the development set and CCC = 0.403 for the test set of the E-DAIC corpus, were achieved with a feature fusion model that combines text and audio representations. Thus, we conclude that multiple modalities give a richer representation of reality, from which an automatic depression severity assessment system could benefit.

Acknowledgments This work was supported by JST CREST JPMJCR19F5 and MEXT/JSPS KAKENHI 19H04133.

References

[1] Abuse, S., for Behavioral Health Statistics, M. H. S. A. and Quality: Results from the 2017 National Survey on Drug Use and Health: Detailed Tables (2018).

[2] Organization;, G. W. H.: Depression and Other Common Mental Disorders: Global Health Estimates (2017).

[3] Association, A. P.: What Is Depression? (2017).

[4] Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G. and Breakspear, M.: Multimodal assistive technologies for depression diagnosis and monitoring, *Journal on Multimodal User Interfaces*, Vol. 7, No. 3, pp. 217–228 (2013).

[5] Joshi, J., Goecke, R., Parker, G. and Breakspear, M.: Can body expressions contribute to automatic depression analysis?, *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, pp. 1–7 (2013).

[6] Sturim, D., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N. and McCree, A.: Automatic detection of depression in speech using gaussian mixture modeling with factor analysis, *Twelfth Annual Conference of the International Speech Communication Association* (2011).

[7] Evans, M. and Mottram, P.: Diagnosis of depression in elderly patients, *Advances in Psychiatric Treatment*, Vol. 6, No. 1, pp. 49–56 (online), DOI: 10.1192/apt.6.1.49 (2000).

[8] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Lui, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani and Maja Pantic: AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition, *Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC'19, co-located with the 27th ACM International Conference on Multimedia, MM 2019*, Nice, France, ACM (2019).

[9] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto and Koichi Shinoda: Multimodal Fusion of BERT-CNN and GCNN Representations for Depression Detection, *Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC'19, co-located with the 27th ACM International Conference on Multimedia, MM 2019*, Nice, France, ACM (2019).

[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[11] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T. and Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation, *arXiv preprint arXiv:1804.03619* (2018).

[12] Owens, A. and Efros, A. A.: Audio-visual scene analysis with self-supervised multisensory features, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648 (2018).

[13] Joshi, J., Dhall, A., Goecke, R. and Cohn, J. F.: Relative body parts movement for automatic depression analysis, *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, pp. 492–497 (2013).

[14] Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G. and Breakspear, M.: Head Pose and Movement Analysis as an Indicator of Depression, *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 283–288 (online), DOI: 10.1109/ACII.2013.53 (2013).

[15] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).

[16] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M. et al.: SimSensei Kiosk: A virtual human interviewer for healthcare decision support, *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1061–1068 (2014).

[17] Kroenke, K. and Spitzer, R. L.: The PHQ-9: a new depression diagnostic and severity measure, *Psychiatric annals*, Vol. 32, No. 9, pp. 509–515 (2002).

[18] Lawrence, I. and Lin, K.: A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, pp. 255–268 (1989).

[19] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[20] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D.: Backpropagation applied to handwritten zip code recognition, *Neural computation*, Vol. 1, No. 4, pp. 541–551 (1989).

[21] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).