Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection

Mariana Rodrigues Makiuchi
Tokyo Institute of Technology
Tokyo, Japan
mariana@ks.c.titech.ac.jp

Tifani Warnita
Tokyo Institute of Technology
Tokyo, Japan
tifani@ks.c.titech.ac.jp

Kuniaki Uto
Tokyo Institute of Technology
Tokyo, Japan
uto@c.titech.ac.jp

Koichi Shinoda
Tokyo Institute of Technology
Tokyo, Japan
shinoda@c.titech.ac.jp

ABSTRACT

Depression is a common, but serious mental disorder that affects people all over the world. Besides providing an easier way of diagnosing the disorder, a computer-aided automatic depression assessment system is demanded in order to reduce subjective bias in the diagnosis. We propose a multimodal fusion of speech and linguistic representation for depression detection. We train our model to infer the Patient Health Questionnaire (PHQ) score of subjects from AVEC 2019 DDS Challenge database, the E-DAIC corpus. For the speech modality, we use deep spectrum features extracted from a pretrained VGG-16 network and employ a Gated Convolutional Neural Network (GCNN) followed by a LSTM layer. For the textual embeddings, we extract BERT textual features and employ a Convolutional Neural Network (CNN) followed by a LSTM layer. We achieved a CCC score equivalent to 0.497 and 0.608 on the E-DAIC corpus development set using the unimodal speech and linguistic models respectively. We further combine the two modalities using a feature fusion approach in which we apply the last representation of each single modality model to a fully-connected layer in order to estimate the PHQ score. With this multimodal approach, it was possible to achieve the CCC score of 0.696 on the development set and 0.403 on the testing set of the E-DAIC corpus, which shows an absolute improvement of 0.283 points from the challenge baseline.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Health care information systems**.

KEYWORDS

depression detection, affective computing, deep learning, multimodal systems, BERT, Gated CNN, CNN

AVEC '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6913-8/19/10…$15.00
https://doi.org/10.1145/3347320.3357694

ACM Reference Format:
Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In *9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347320.3357694

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 [27], 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 [5]. Furthermore, depression can also be a sign that someone is suffering from a neurocognitive disorder, such as dementia [4]. Therefore, the development of new methods and tools to support a fast and precise depression diagnosis is undoubtedly necessary.

In this regard, several studies [2, 14, 19–21, 41] proposed computer-aided solutions for seeking an automatic and objective depression detection method. 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 [12].

Even though the automatic depression detection has been widely investigated from different perspectives, it is still considered a challenge. In fact, the 2019 edition of the Audio/Visual Emotion Challenge and Workshop (AVEC 2019) [13] proposes the Detecting Depression with AI Sub-Challenge (DDS). This sub-challenge aims the automatic depression severity assessment of US Army veterans from audiovisual recordings of their interaction, in a clinical interview setting, with a virtual agent, which can be controlled by a human as a Wizard-of-Oz or an artificial intelligence (AI). Thus, besides the automatic depression severity evaluation, the DDS also seeks the comprehension of how the absence of a human controlling the virtual agent impacts on this automatic evaluation.

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 DDS test partition.

The remainder of this paper is organized as follows. In Section 2, we summarize relevant related works. In Sections 3 and 4, we respectively introduce the dataset used for the AVEC 2019 DDS and the evaluation metric employed in this challenge. Sections 5 and 6 are respectively dedicated to the description of the proposed methodology and to the presentation of the conducted experiments and the corresponding results. Finally, Section 7 concludes this paper and discusses future improvements to the methodology presented in Section 5.

2 RELATED WORKS

This section presents a summary of the current state-of-the-art with respect to topics related to the method of automatic depression detection conducted in this work.

2.1 Audio Representation

Audio data can be described by features of diverse nature, such as spectrograms, power, mel-frequency cepstral coefficients (MFCCs) and deep spectrum representations. A robust speech representation can be used to solve various paralinguistic tasks, which are related to events beyond linguistic [37]. Handcrafted representations of speech [34–36, 38] have been widely used to solve paralinguistic problems such as emotion, health state and personality traits recognition.

Recently, deep neural networks have been employed to extract discriminative features from speech [3, 6]. Compared to statistical functions designed from handcrafted feature sets, deep neural networks are able to learn a more robust data representation due to their data generalization ability. Although many advances in deep learning have occurred, spectrogram images are widely chosen as the input to audio representation models as opposed to raw speech input, because learning from raw data is still considered a challenging problem.

In order to learn a task-specific representation from the two-dimensional speech description (i.e. spectrogram images), deep neural networks, such as convolutional neural networks (CNNs), have been often employed [18]. Due to their sparse weight connections, one of the main advantages of CNNs is the temporal information understanding without compromising the generalizability. Moreover, CNNs with a gating mechanism [8] show better results for the dementia detection task [42] due to their ability to overcome the vanishing gradient problem.

2.2 Natural Language Processing

Natural language processing (NLP) is the field of computer science concerned with the human-computer interaction through natural language. Recently, the introduction of distributional vectors (or word embeddings) [26, 31] as textual data representation in the NLP field made the high-accuracy solution of many challenging NLP tasks, such as the ones related to question answering [43, 44], sentiment analysis [33, 39] and natural language inference (NLI) [29], feasible.

Recent state-of-the-art works in the NLP field have been exploring techniques, such as model pre-training and bidirectional language representation, in order to grasp the semantic complexity of a language. [32], for example, proposes a bidirectional word representation from the combination of both forward and backward language models, thus achieving the state-of-the-art on six NLP tasks.

However, as opposed to [32], the recently proposed Bidirectional Encoder Representations from Transformers (BERT)[10] was designed to pre-train deep bidirectional representations from unlabelled text under a masked language model.

Pre-trained BERT models can be fine-tuned with a single additional output layer, thus achieving state-of-the-art results without drastic task-oriented architecture modifications. In fact, these models can achieve the state-of-the art on eleven natural language processing tasks. Therefore, due to their powerful language representation, pre-trained BERT models were chosen to extract textual features from interview transcripts in this work.

2.3 Multimodality

Recent works have shown promising results in numerous tasks due to the adoption of multimodal approaches. [11], for example, proposes a speaker-independent audio-visual model for speech separation that outperforms audio-only and audio-visual models on classic speech separation tasks. Moreover, [28] also considers the fusion of audio and visual signals to build a multisensory representation of videos, which can be satisfactorily applied to sound source localization, audio-visual action recognition and on/off screen audio source separation. The underlying hypothesis held by [11, 28] and other multimodal approaches is that information of different nature acquired from the same source can be extremely valuable to understand the problem's context and, thus, find its best solution.

2.4 Depression Detection

It is worth mentioning other attempts and studies conducted to develop automatic depression detection systems. [21], for instance, studies the contribution of upper body movement to the depression detection, while [19] analyses the influence of the whole body movement relatively to its parts on the same task. Both [21] and [19] represent the body movement in a bag-of-words approach and they both apply Space-Time Interest Points (STIP) to assist the feature extraction. The results of these works show that both the relative body and the upper body (head and shoulders) movements are significant for the depression detection and they also show the importance of the fusion of multiple features.

Similar to [19, 21], [2] analyses the influence of the head pose in the depression detection. According to the results shown in [2], the head pose holds effective cues in this disorder diagnosis, since their proposed model achieves an average accuracy of 71.2% on this task. [2] extracted head pose and movement features from videos using a 3D model projected on a 2D Active Appearance Model (AAM) and created a Gaussian Mixture Model (GMM) for each subject, which,

combined with the SVM classifier, composes the hybrid classifier used for the depression detection.

Moreover, [20] proposes the fusion of bags of audio and visual features for the depression diagnosis. These bags of features are then applied to a Support Vector Machine (SVM) classifier. The audio features include the fundamental frequency $f0$, loudness, intensity and mel-frequency cepstral coefficients (MFCC), while the visual features are related to the intra-facial muscle movements and the movements of the patient's upper body.

Although [20] considers audio features for the depression assessment, the work presented in this paper significantly differs from the one proposed in [20], since we consider the semantic content of the patient speech by extracting deep bidirectional textual features with a pre-trained BERT model. In addition, in the presented work, the audio features are not represented in a bag-of-audio approach, but as a set of deep spectrum features extracted with a VGG-16 [40] architecture [13].

3 E-DAIC CORPUS

The dataset adopted in the AVEC 2019 DDS is the Extended Distress Analysis Interview Corpus (E-DAIC) [9], an extension of the DAIC-WOZ corpus, which in turn is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) [16].

The DAIC corpus contains audiovisual recordings of patients interacting with an agent conducting a clinical interview designed to aid the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. In the E-DAIC corpus, this virtual agent can be a Wizard-of-Oz controlled by a human in another room or it can be fully automated, controlled by an artificial intelligence. The E-DAIC includes the transcript of the interactions automatically transcribed with Google Cloud's speech recognition service, the participants audio files, their facial features and each patient PTSD Checklist Civilian Version (PCL-C) [7] and Patient Health Questionnaire [23] depression module (PHQ-8) scores.

The PHQ-8 and the PCL-C attempt to assess, respectively, the depression and the Post-Traumatic Stress Disorder (PTSD) severities. The PCL-C score ranges from 0 to 85, while the PHQ-8 score ranges from 0 to 24. The PHQ-8 score's cutpoints are defined at 5, 10, 15 and 20 for mild, moderate, moderately severe, and severe depression levels, respectively.

In the E-DAIC dataset, there are 275 subjects, who are US Army veterans. For the DDS, this dataset was divided into train, development and test partitions with 163, 56 and 56 subjects respectively. In the train and development sets, the interviewer can be either a human in a Wizard-of-Oz setting or an AI, while, in the test set, there are only interviews conducted by the AI.

4 EVALUATION METRIC

The performance metric adopted in the AVEC 2019 DDS is the Concordance Correlation Coefficient (CCC) [24], 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 the 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

This section presents the designed models for the AVEC 2019 Depression Detection Subchallenge (DDS). Sections 5.1 to 5.3 explore single modalities representations for feature extraction and for the depression detection task itself. Section 5.4 presents a model for feature level fusion of these single modalities for the depression severity assessment task. Finally, Section 5.5 briefly introduces the baseline model proposed by [13], to which our approach is compared in Section 6.

5.1 Audio Model

We use a deep spectrum representation extracted from a pretrained 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, which 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 CNN-based Model. Our first speech model is composed by stacked convolution blocks followed by fully-connected (dense) layers. Each convolution block is composed by a 1D convolution layer followed by batch normalization, the ReLU activation function and a max-pooling layer to halve the input size. We use nine convolution blocks, each with a different number N of convolution filters for their convolution layers. Thus, the number of filters of each of the nine convolution layers of our model is, from the input to the output, equal to $N = [128, 64, 64, 64, 64, 32, 32, 32, 32]$. The output from the convolution blocks are then flattened and input to a fully-connected layer with 256 hidden neurons. Finally, the output of this fully-connected layer is applied to a batch normalization, the ReLU activation function, a dropout layer and another fully-connected layer composed by one neuron, which outputs the PHQ-8 score as a single value. This model was trained with the mean-squared error loss function and the Adam optimizer. The full CNN model is depicted in Figure 1.

5.1.2 GCNN-LSTM-based model. Besides having the CNN-based model, we also trained a GCNN-based model composed by stacked gated convolution blocks followed by a LSTM layer and a fully-connected layer. This GCNN-LSTM model differs greatly from the CNN model, because a LSTM layer is added to the GCNN-LSTM model and the convolution blocks are replaced by gated blocks, which consist of two convolution layers followed by a gating mechanism and a max-pooling layer. For each gated block, the input to the max-pooling layer is defined as

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

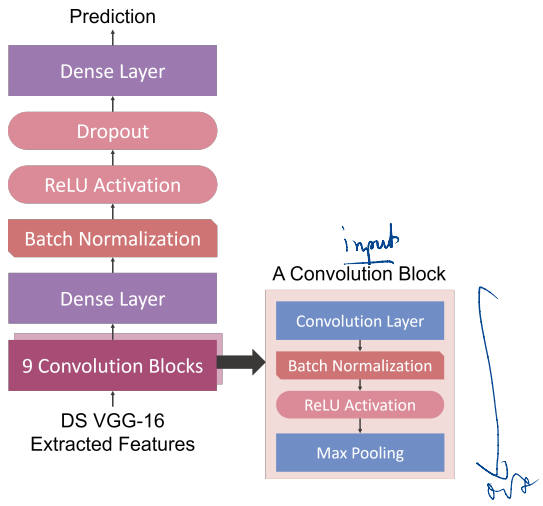

Figure 1: Our CNN model for speech-based depression assessment with nine convolution blocks.

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, 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 also trained with the mean-squared error loss function and the Adam optimizer [22]. A complete representation of our GCNN-LSTM model is shown in Figure 2.

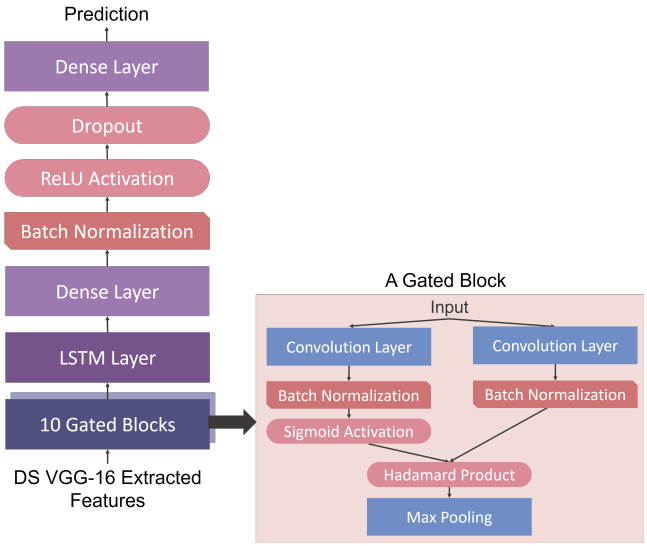

Figure 2: Our best GCNN-LSTM model for speech-based depression assessment with nine gated convolution blocks and a LSTM layer.

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, since the semantic content of speech can reveal a person's habits, their likes and dislikes, their opinions, their emotions and the quality of their personal relationships, which are elements that should be considered in a depression diagnosis.

Thus, in order to represent the semantic content of the E-DAIC corpus interviews by incorporating context information from both left and right directions (i.e. past and future within the interview), textual features were extracted from the automatically transcribed transcripts [9], which are part of the E-DAIC corpus, briefly introduced in Section 3. This textual feature extraction was performed by the pretrained BERT-large model [10], which has 24 layers (i.e. Transformer blocks), hidden size equal to 1024 and 16 self-attention heads, thus totalizing 340 million parameters.

The features were extracted from the last BERT layer and they were composed by a single feature array for each word token. Thus, for each subject, BERT-large features 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 input samples to the textual models (i.e. textual features extracted with BERT) 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 LSTM-based Model. In order to discern the features as a time series, a simple model composed of a 64 and a 32-dimensional Long Short-Term Memory (LSTM) layers both with a dropout rate equal to 0.1 was initially designed for the DDS. A diagram depicting this model structure is shown in Figure 3.

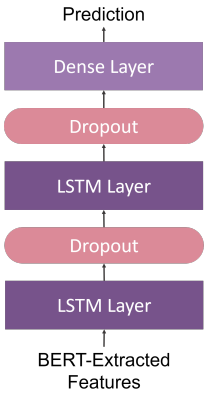

Figure 3: LSTM-based model for depression assessment with textual features.

The input features depicted in the diagram of Figure 3 are the features extracted from the BERT-large model for a single patient.

In Figure 3, the final dense layer (i.e. fully connected layer) is responsible for converting the representation acquired with the LSTM layers from the feature space to a single prediction, which corresponds to the PHQ score of the patient. Besides being used to predict the PHQ score directly, this model was also employed to extract highly representative textual features, which served as input to the fusion model proposed in this work. However, although this model, when trained to predict the PHQ score, could achieve a CCC value equal to 0.360 over the validation partition, its performance over the test set, 0.048 points evaluated with the same metric, was not satisfactory. Moreover, it is known that, although Recurrent Neural Network (RNN) models can represent temporal patterns, they require longer training time when compared to other models [30].

In order to address this issue, we trained another model for depression assessment using only text features extracted from BERT-large as the input. This model is presented in the following section.

5.2.2 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 the usage of only a pair of LSTM layers, as in the previous model depicted in Figure 3, would not drastically reduce the features matrices' dimensionality, making the prediction task challenging for that model's dense layer. Moreover, 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 [25]. In fact, CNN-based models have been giving notable results in numerous tasks over the past few years [15].

Thus, similar to the CNN-based model for audio features presented in Section 5.1.1, we define seven convolution blocks with different number of filters N for each convolution layer. These convolution blocks have the same structure as the blocks in the CNN-based model of Section 5.1.1 and their convolution 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 Figure 4.

In both models used to assess depression with only textual features, 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. However, the optimizer chosen for the LSTM-based model represented in Figure 3 is the stochastic gradient descent, while, in the CNN-LSTM model, depicted in Figure 4, is Adam [22] with parameters β_1 equal to 0.9 and β_2, to 0.999.

5.3 Visual Model

For the visual model, we utilize a deep visual representation extracted from a ResNet-50 model [17] as input. We choose the time dimension $T = 6000$ for the visual features and apply them to a GCNN model similar to the GCNN-LSTM model presented in Section 5.1 except the fact that, in the visual model, there are not recurrent layers. Thus, the visual model can be depicted as in Figure 5,

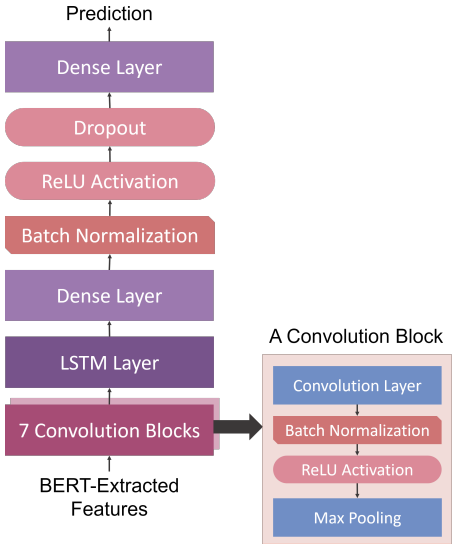

Figure 4: CNN and LSTM-based model for depression assessment with textual features.

in which, from the input to the output, the convolution filters of the gated blocks have size $N = [512, 256, 256, 128, 128, 64, 64, 32, 32, 16]$.

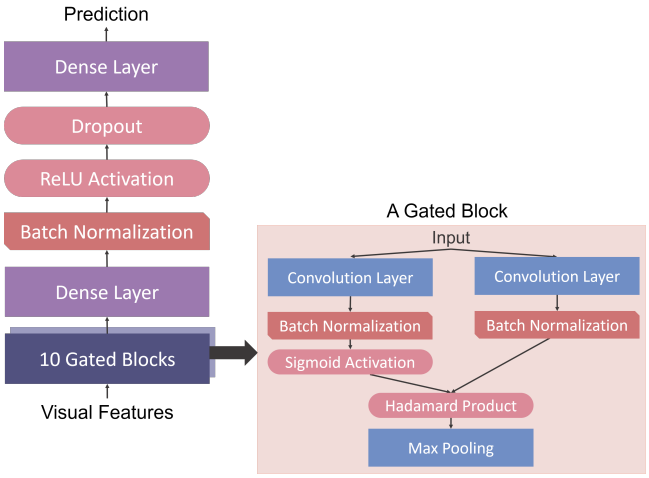

Figure 5: GCNN-based model for depression assessment with visual features.

Similarly to the models 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

In this work, we use the embeddings obtained from the first dense layer of each modality. The modality-specific representations are concatenated as one input vector and this resulting array is then input to the multimodal network. The fused feature array is trained

on a fully-connected layer with 512 hidden neurons. The multimodal network is trained to minimize the mean-squared error loss between the ground-truth PHQ score and the network prediction. Adam optimizer [22] is employed during the training. A diagram representing the fusion model is depicted in Figure 6.

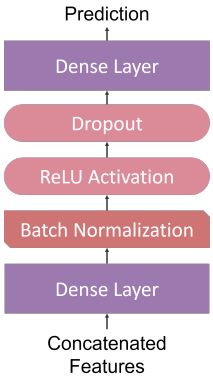

Figure 6: Fusion model. The input features are the concatenation of features acquired from the unimodal models presented in Sections 5.1 to 5.3.

5.5 Baseline Model

The results acquired during this work were compared to the ones obtained with the baseline model proposed by [13], which is briefly presented in this section.

The baseline model 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 a batch size of 15 is used consistently. All the available audio and visual features were input to this baseline model and the fusion of the various audiovisual representations was obtained by averaging their scores.

6 EXPERIMENTS AND RESULTS

In this section, the main results acquired with the models described in Section 5 will be presented and discussed. Moreover, experiments conducted in order to evaluate the number of CNN and GCNN blocks used in these models as well as the application of different visual features to the model presented in Section 5.3 are also exposed in this section.

6.1 Number of CNN/GCNN blocks [1]

In this section, the models presented in Sections 5.1.2 and 5.2.2 were trained with different amounts of GCNN and CNN blocks respectively in order to identify the ideal model configuration for the depression assessment evaluated with the CCC metric. Apart from the number of blocks, the other model hyperparameters were defined as presented in Sections 5.1.2 and 5.2.2. Each model configuration was trained five times and the average CCC and the maximum CCC on the development partition were acquired.

For the GCNN-LSTM audio model presented in Section 5.1.2, models with 1 to 10 gated blocks were evaluated with the CCC metric. For each model configuration, 5 models were trained and the maximum CCC as well as the average CCC for each configuration are reported in Figure 7.

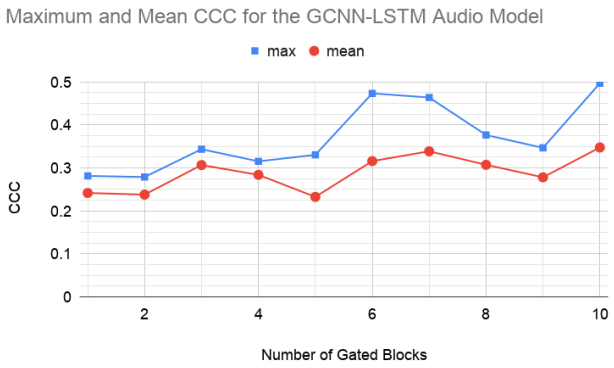

Figure 7: Maximum and mean CCC acquired with different number of gated blocks applied to the CGNN-LSTM audio model presented in Section 5.1.2. The maximum CCC value shown in the graph is equal to 0.497 and it was achieved with the GCNN-LSTM text model with 10 GCNN blocks.

From the graph depicted in Figure 7, it is possible to conclude that the mean CCC is relatively robust to the audio model configuration. However, the maximum CCC seems to have its higher values for models with 6 and 10 gated blocks. Although we have tested models with only 1 to 10 gated blocks, models with a larger amount of gated blocks should be further investigated since the graph in Figure 7 seems to show a trend for an increase in the CCC value.

The convolution filters' configuration for each tested model was defined in an ablation manner. Thus, the configuration defined in Section 5.1.2 for 10 gated blocks, $N = [512, 256, 256, 128, 128, 64, 64, 32, 32, 16]$, had its smaller filter removed one by one. Therefore, a 9 gated blocks configuration was defined as $N = [512, 256, 256, 128, 128, 64, 64, 32, 32]$ and a 4 gated blocks, as $N = [512, 256, 256, 128]$, for example.

For the text model presented in Section 5.2.2, models with 1 to 12 CNN blocks were trained and evaluated on the development partition. A graph with the average and the maximum CCC for each model configuration is depicted in Figure 8.

As it can be concluded from Figure 8, the model configuration with 8 blocks achieve the best maximum and average CCC values on the development partition. Moreover, the addition of blocks seems to improve the performance evaluated with the CCC metric from the model with 3 CNN blocks until the model with 8 blocks. However, the addition of extra CNN blocks to the 8 CNN blocks-LSTM model does not contribute to improve the model performance on the development partition.

[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.

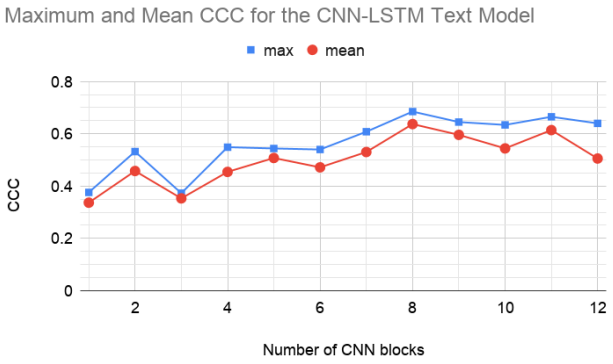

Figure 8: Maximum and mean CCC acquired with different number of CNN blocks applied to the CNN-LSTM text model presented in Section 5.2.2. The maximum CCC value shown in the graph is equal to 0.685 and it was achieved with the CNN-LSTM text model with 8 CNN blocks.

Table 1: Size N of convolution filters, from the input to the output, of each CNN-LSTM text model configuration

CNN blocks	Convolution Filters' Configuration
8	[128, 64, 64, 64, 64, 32, 32, 32]
9	[128, 64, 64, 64, 64, 32, 32, 32, 16]
10	[128, 64, 64, 64, 64, 32, 32, 32, 16, 16]
11	[128, 64, 64, 64, 64, 32, 32, 32, 16, 16, 8]
12	[128, 64, 64, 64, 64, 32, 32, 32, 16, 16, 8, 4]

The convolution filters' configuration for each CNN-LSTM text model was also performed in the same ablation manner as in the GCNN-LSTM audio model starting from the 7 blocks configuration exposed in Section 5.2.2. For models with more than 7 convolution blocks, the filter configuration is summarized in Table 1.

6.2 Different visual features [1]

The model presented in Section 5.3 was tested with all the visual features that are available in the database used for the AVEC 2019 DDS. Therefore, features extracted with VGG-16 and ResNet-50 architectures as well as Bag-of-Visual-Words (BoVW) and Facial Action Units (FAUs) were utilized. 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 is presented in Table 2.

As it can be observed from Table 2, the best model has 7 gated blocks and it uses VGG-extracted features as input. Moreover, it can be concluded that, for most of the models' configurations, a model that uses features extracted with deep models (VGG and ResNet) will have better results when compared to the same model using BoVW or FAUs approaches. This observation can be explained from the deep models' ability of extracting highly representative features and from the challenge of defining significant features in a handcrafted approach.

Table 2: Best CCC for different combinations of visual features and number of gated blocks applied to the visual model presented in Section 5.3. Cells filled with '-' indicate that it was not possible to acquire the corresponding results due to model limitations.

Gated Blocks	CCC			
	FAUs	BoVW	VGG	ResNet
1	0.110	0.142	0.354	0.222
2	0.109	-0.012	0.365	0.200
3	0.111	0.195	0.365	0.123
4	0.107	0.041	0.283	0.174
5	**0.113**	**0.238**	0.152	0.104
6	0.105	0.070	0.257	0.325
7	0.100	-0.035	**0.373**	0.121
8	0.096	0.154	0.246	0.273
9	-	-0.034	0.218	0.311
10	-	0.185	0.277	**0.372**

6.3 Results

The results are summarized in Table 3. 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. The test set results are reported in Table 3 according to the information provided by the AVEC challenge organizers on four of our models. The CCC and RMSE results in Table 3 for the baseline model correspond to the higher values reported in [13] regardless of the model that provided these results. Thus, the value of 0.336 for the CCC score on the development set and 5.03 and 6.37 for the RMSE on the respective development and test partitions were obtained with the baseline fusion model. Moreover, the result reported as 0.120 for the CCC metric on the test partition was acquired with an unimodal model that takes visual features extracted with a ResNet-50 network as input.

In Table 3, the GCNN-LSTM audio model uses 10 gated blocks and, although this model configuration achieves the best results on the development set evaluated with the CCC metric compared to audio models with less gated blocks, as discussed in Section 6.1, models with more than 10 gated blocks should be further investigated, since they might give better results.

From Table 3, 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 features extracted from the GCNN-LSTM model, presented in Section 5.1.2 and text features acquired from the CNN-LSTM architecture, introduced in Section 5.2.2. Moreover, it is possible to conclude that, in every situation, the fusion of features 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

Table 3: 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. The audio models denominated by CNN and GCNN-LSTM (with 10 gated blocks) are respectively introduced in Sections 5.1.1 and 5.1.2 and they use features extracted with a VGG-16 architecture as their input. The text models indicated by LSTM and CNN-LSTM (7 CNN blocks-LSTM and 8 CNN blocks-LSTM) are respectively described in Sections 5.2.1 and 5.2.2 and their input is extracted with a BERT-large architecture. The visual model is presented in Section 5.3 and we report the results acquired with features extracted with a ResNet-50 architecture as discussed in Section 6.2. The fusion models presented in this table combine highly representative features extracted from the unimodal models in a feature-level fusion manner, as explained in Section 5.4.

Modality	Model	CCC		RMSE	
		Development	Test	Development	Test
-	Challenge baseline [13]	0.336	0.120	5.03	6.37
Audio	CNN	0.338	0.199	5.97	7.02
	GCNN-LSTM	0.497	-	5.70	-
Text	LSTM	0.360	0.048	4.97	6.88
	7 CNN blocks-LSTM	0.608	-	4.51	-
	8 CNN blocks-LSTM	0.685	-	4.22	-
Visual	GCNN	0.372	-	5.74	-
Fusion	CNN (audio) and LSTM (text)	0.452	0.213	5.08	6.42
	GCNN-LSTM (audio) and 7 CNN blocks-LSTM (text)	**0.696**	**0.403**	**3.86**	**6.11**
	GCNN-LSTM (audio), 7 CNN blocks-LSTM (text) and GCNN (visual)	0.624	-	4.86	-

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. Therefore, from the experiments conducted with visual features as the input, it was not possible to validate the significance of this type of features for the depression severity assessment.

Another conclusion that can be taken from the results presented in Table 3 is that, for all the models evaluated on the test partition, the CCC and the RMSE metrics were better when the model was evaluated on the development set compared to the test set. This fact indicates that the absence of a human conducting the interviewer as a virtual agent has a negative impact on the performance of the automatic depression diagnosis.

Finally, it can be observed that all the models presented in this paper outperforms the challenge baseline [13] when evaluated over the development partition with the CCC metric. Moreover, except for the unimodal text model based on LSTM layers, there is an improvement in all the reported CCC values for the test partition when compared to the same baseline. The best model presented in this paper, the multimodal fusion of audio features extracted from the GCNN-LSTM model and text features acquired from the CNN-LSTM architecture, outperforms the baseline in both CCC and RMSE metrics and over the development and the test partitions.

7 CONCLUSION

In this work, a multimodal approach for automatic depression detection was presented. First, models that individually consider text, audio and visual features were developed and tested. These unimodal models were then used as highly representative feature extractors and the resulting features were thus combined in a feature level fusion manner. The best results presented in this paper, CCC = 0.696 for the development set and CCC = 0.403 for the test set, were achieved with a multimodal model that combines text and audio features. This result indicates that the utilization of multiple modalities gives a richer representation of reality, from which an automatic depression severity assessment system could benefit.

Moreover, the lower CCC and higher RMSE values for the test partition in comparison with the development set for all presented models reveal that the absence of a human conducting the virtual agent has a negative impact on the automatic depression assessment model accuracy.

Future research on more sophisticated fusion methods might improve the overall performance and the robustness of the multimodal model presented in this work. The possibility of improvement in the textual feature representation due to a more accurate speech transcript should be also further investigated. In addition, although our visual model gave suboptimal results compared to other unimodal models in this work, a better way of learning from visual features is another interesting and promising future direction to be explored and it could improve our model's accuracy since previous works have shown that visual information provides important cues for depression assessment. Finally, audio models with a larger amount of gated blocks will also be considered as a future work.

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

ACKNOWLEDGMENTS

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

REFERENCES

[1] Substance Abuse, Mental Health Services AdministrationCenter for Behavioral Health Statistics, and Quality. 2018. Results from the 2017 National Survey on Drug Use and Health: Detailed Tables.

[2] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. 2013. Head Pose and Movement Analysis as an Indicator of Depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 283–288. https://doi.org/10.1109/ACII.2013.53

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. *Proc. Interspeech 2017* (2017), 3512–3516.

[4] Alzheimer's Association et al. 2018. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* 14, 3 (2018), 367–429.

[5] American Psychiatric Association. 2017. What Is Depression? Retrieved June 5 2019 from https://www.psychiatry.org/patients-families/depression/what-is-depression

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 892–900.

[7] Edward B Blanchard, Jacqueline Jones-Alexander, Todd C Buckley, and Catherine A Forneris. 1996. Psychometric properties of the PTSD Checklist (PCL). *Behaviour research and therapy* 34, 8 (1996), 669–673.

[8] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 933–941.

[9] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619* (2018).

[12] Mavis Evans and Pat Mottram. 2000. Diagnosis of depression in elderly patients. *Advances in Psychiatric Treatment* 6, 1 (2000), 49–56. https://doi.org/10.1192/apt.6.1.49

[13] Fabien Ringeval and Björn Schuller and Michel Valstar and Nicholas Cummins and Roddy Cowie and Leili Tavabi and Maximilian Schmitt and Sina Alisamir and Shahin Amiriparian and Eva-Maria Messner and Siyang Song and Shuo Lui and Ziping Zhao and Adria Mallol-Ragolta and Zhao Ren, and Mohammad Soleymani, and Maja Pantic. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *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*, Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic (Eds.). ACM, Nice, France.

[14] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. 2013. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[16] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews.. In *LREC*. Citeseer, 3123–3128.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 801–804.

[19] Jyoti Joshi, Abhinav Dhall, Roland Goecke, and Jeffrey F Cohn. 2013. Relative body parts movement for automatic depression analysis. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 492–497.

[20] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 7, 3 (2013), 217–228.

[21] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can body expressions contribute to automatic depression analysis?. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–7.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals* 32, 9 (2002), 509–515.

[24] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[25] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[27] Geneva: World Health Organization;. 2017. Depression and Other Common Mental Disorders: Global Health Estimates.

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

[29] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933* (2016).

[30] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[32] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[33] Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. 2016. Multimodal emotion recognition for AVEC 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 75–82.

[34] B Schuller, A Batliner, S Steidl, F Schiel, and J Krajewski. 2011. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. INTERSPEECH 2011, Florence, Italy*.

[35] B Schuller, S Steidl, and A Batliner. 2009. The Interspeech 2009 Emotion Challenge. In *Proc. Interspeech 2009, Brighton, UK*. 312–315.

[36] B Schuller, S Steidl, A Batliner, F Burkhardt, L Devillers, C Müller, and S Narayanan. 2010. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2794–2797.

[37] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language - State-of-the-art and the challenge. *Computer Speech & Language* 27, 1 (2013), 4–39.

[38] B Schuller, S Steidl, A Batliner, E Nöth, A Vinciarelli, F Burkhardt, R van Son, F Weninger, F Eyben, T Bocklet, et al. 2012. The INTERSPEECH 2012 Speaker Trait Challenge. In *INTERSPEECH 2012, Portland, OR, USA*.

[39] Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 959–962.

[40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[41] Douglas Sturim, Pedro A Torres-Carrasquillo, Thomas F Quatieri, Nicolas Malyska, and Alan McCree. 2011. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Twelfth Annual Conference of the International Speech Communication Association*.

[42] Tifani Warnita, Nakamasa Inoue, and Koichi Shinoda. 2018. Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data. *Proc. Interspeech 2018* (2018), 1706–1710.

[43] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*. Springer, 451–466.

[44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked Attention Networks for Image Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.