D-Vlog: Multimodal Vlog Dataset for Depression Detection

Jeewoo Yoon[1,3], **Chaewon Kang**[1], **Seungbae Kim**[2], **Jinyoung Han**[1,3,*]

[1] Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, Korea
[2] Department of Computer Science, University of California, Los Angeles, USA
[3] RaonData, Seoul, Korea
{yoonjeewoo, codnjs3}@g.skku.edu, sbkim@cs.ucla.edu, jinyounghan@skku.edu

Abstract

Detecting depression based on non-verbal behaviors has received great attention. However, most prior work on detecting depression mainly focused on detecting depressed individuals in laboratory settings, which are difficult to be generalized in practice. In addition, little attention has been paid to analyzing the non-verbal behaviors of depressed individuals in the wild. Therefore, in this paper, we present a multimodal depression dataset, D-Vlog, which consists of 961 vlogs (i.e., around 160 hours) collected from YouTube, which can be utilized in developing depression detection models based on the non-verbal behavior of individuals in real-world scenario. We develop a multimodal deep learning model that uses acoustic and visual features extracted from collected data to detect depression. Our proposed model employs the cross-attention mechanism to effectively capture the relationship across acoustic and visual features, and generates useful multimodal representations for depression detection. The extensive experimental results demonstrate that the proposed model significantly outperforms other baseline models. We believe our dataset and the proposed model are useful for analyzing and detecting depressed individuals based on non-verbal behavior.

Introduction

Major Depressive Disorder (MDD) is considered as one of the most common mental health disorders, and has become an important problem in society (Üstün et al. 2004). The World Health Organization (WHO) reported that more than 264 million people were diagnosed with clinical depression in 2020[1]. If the depression is left untreated, it can cause severe outcomes such as addiction, reckless behavior, and even suicide (Ghosh, Ekbal, and Bhattacharyya 2021). Hence, early detection of depression and appropriate clinical intervention accordingly is becoming increasingly important; with early detection and proper clinical intervention of depression, it can be regarded as a curable mental disorder (Lejuez, Hopko, and Hopko 2001).

In analyzing and detecting depression, non-verbal signals have been considered as important factors. Scholars have shown that depressed and non-depressed groups can be distinguished by non-verbal signals such as facial expressions, body movements, and vocal intensity (Edition et al. 2013; Sobin and Sackeim 1997). For example, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) describes that agitation (e.g., the inability to sit still, pacing, hand-wringing) or retardation (e.g., slowed speech and body movements, increased pauses before answering) can be observed from depressed patients.

With the recent advancements in computer vision and signal processing techniques, there have been great efforts in developing depression detection models using non-verbal (behavioral) signals (Senoussaoui et al. 2014; Yang et al. 2016, 2017a; Rodrigues Makiuchi et al. 2019). However, big challenges remain in automatic and accurate depression detection based on non-verbal signals. First, although many scholars have studied how to analyze and detect depressed individuals, mainly due to privacy issues, there are only a few publicly available datasets on depression (Gratch et al. 2014; Dibeklioğlu, Hammal, and Cohn 2017). Moreover, these datasets on depression have been developed in a laboratory setting through interviews, which may not capture the usual behavior of depressed individuals in the wild (Huang et al. 2020).

Therefore, we propose a new depression dataset that includes the non-verbal behavior of depressed individuals captured not in a laboratory but in daily lives. More specifically, we first collected vlogs of both depressed and non-depressed individuals using search keywords such as 'depression vlog', 'daily vlog', etc. We then manually annotated depression and non-depression vlogs based on the developed annotation rules, e.g., "Does the given video has a vlog format?", "Does the speaker currently suffer from depression?", etc., and extracted acoustic and visual features from them. Since the vlogs are voluntarily recorded and uploaded, we believe that our dataset contains more natural depression-related features that are distinct from non-depressed ones than the prior datasets collected in a laboratory setting. Additionally, we propose a multimodal deep learning model that only uses non-verbal features, acoustic and visual features, extracted from the given video in detecting depression. The proposed model utilizes the Transformer encoders to encode acoustic and visual sequences, and fuse them with a cross-attention mechanism for gener-

*Corresponding author.

[1]https://www.who.int/news-room/fact-sheets/detail/depression

ating an effective multimodal representation. Our extensive experiments demonstrate that the proposed model outperforms other baselines, which has implications on accurately and timely detecting depression at an early stage.

We highlight the main contributions of our work as follows:

- We make our dataset publicly available[2]. Considering privacy concerns, we only provide de-identified anonymized data. The dataset contains a total of 961 videos (i.e., around 160 hours) with 816 distinct speakers.
- To our best knowledge, this is the first attempt to apply cross-attention on multimodal depression detection. Our evaluation shows that the cross-attention layer is demonstrated as useful in representing relationships across modalities thereby achieving high performance in detecting depression based on multimodal data.
- The proposed model and dataset can be generalized and applied to other formats of video data (e.g., clinical interviews). The experimental result shows that the model trained with the proposed dataset also achieves a high depression detection performance for DAIC-WOZ, where videos are recorded in a clinical interview format.

Related Work

Datasets for Depression Analysis and Detection

As automatic depression detection has gained great attention, there have been efforts in developing datasets to analyze and detect depression (Yingthawornsuk et al. 2006; Maddage et al. 2009; Sharifa et al. 2012). However, due to privacy concerns, most of the datasets are only available for their own research and not released in public. There are only a few available datasets publicly open for research (Becker et al. 1994; Dibeklioğlu, Hammal, and Cohn 2017; Gratch et al. 2014). DAIC-WOZ (Gratch et al. 2014) is one of the widely-used open datasets, which contains clinical interviews designed to simulate standard guidelines for identifying people with major depression, post-traumatic stress disorder (PTSD), etc. The samples in DAIC-WOZ are annotated based on a self-report (i.e., PHQ-8), and the dataset consists of verbal (text) and non-verbal (acoustic and visual) features. Another well-known open dataset, Depression Severity Interviews Database (Dibeklioğlu, Hammal, and Cohn 2017) (i.e., Pittsburgh dataset), also contains clinical interviews, however, the samples are annotated based on the clinical assessment. Audio-Visual Depressive Language corpus (AViD-Corpus), which was used in AVEC 2013 (Valstar et al. 2013) and AVEC 2014 (Valstar et al. 2014), includes videos of participants who are singing, reading a speech, etc. The samples in AViD-Corpus are also labeled based on the self-report (i.e., BDI-II). These datasets have played important roles in understanding and analyzing patterns of depression. However, since most of them are collected and developed in a laboratory setting, they may not capture the usual behavior of depressed individuals (Huang et al. 2020). On the other hand, our proposed dataset, D-Vlog, includes the non-verbal behavior of depressed individuals captured not in a laboratory but in daily lives. Table 1 compares the proposed dataset, D-Vlog, with the prior datasets developed in a laboratory setting.

Table 1: Comparison between D-Vlog dataset and publically accessible depression datasets in a laboratory setting.

Dataset	Modality	# Subjects	# Samples
DAIC-WOZ	A+V+T	189	189
Pittsburgh	A+V	49	130
AViD-Corpus	A+V	292	340
D-Vlog (Ours)	A+V	816	961

Detecting Depression Using Social Media Data

Unlike data collected in a laboratory setting (e.g., DAIC-WOZ), data collected from social media can be useful in capturing and observing usual patterns of depressed individuals revealed in daily lives (Huang et al. 2020). Therefore, scholars have developed depression detection models based on social media data (e.g., texts, images, and tags) (Wang et al. 2013; Gui et al. 2019; Tadesse et al. 2019). For example, Wang et al. (2013) proposed a depression detection model that uses tags and texts in Sina Weibo, a popular micro-blog in China. They extracted content, interaction, and behavior features based on the psychological theory regarding depression to train the model. Gui et al. (2019) applied the cooperative multi-agent reinforcement learning methods to identify depression of the Twitter user based on the text and image features. Orabi et al. (2018) proposed a method for word-embedding optimization and applied it to detect depressed individuals in Twitter. While these studies have focused on detecting depression using social media data, little attention has been paid to model and analyze how depression can be detected using video data on social media, which can capture usual behavior in daily lives (Correia et al. 2021).

Multimodal Fusion

The multimodal fusion aims at integrating information from multiple modalities to predict the target output (Baltrušaitis, Ahuja, and Morency 2018). Prior work in multimodal fusion mostly relied on the model-agnostic approach such as early fusion, late fusion, and hybrid fusion (D'mello and Kory 2015). The scholars created multimodal representations by simple concatenating, adding, or multiplying feature/decision level vectors to perform prediction tasks (Kim, Lee, and Provost 2013; Liu, Zheng, and Lu 2016; Gunes and Piccardi 2007; Liu et al. 2018). With the advancements in deep learning techniques, neural networks can be used for multimodal fusion (Wöllmer et al. 2010; Nicolaou, Gunes, and Pantic 2011; Wöllmer et al. 2013). These work employed deep neural networks (e.g., LSTMs) to characterize the target by generating joint representations from multiple modalities. To effectively utilize and fuse multimodal data, we adopt

[2]https://sites.google.com/view/jeewoo-yoon/dataset

a model-based approach with a well-known deep learning architecture, Transformer (Vaswani et al. 2017), and use a cross-attention mechanism to generate multimodal representation. To the best of our knowledge, this is the first attempt to use the cross-attention fusion method for detecting depression.

Depression Vlog (D-Vlog) Dataset

In this section, we introduce the proposed depression vlog (D-Vlog) dataset. In particular, we describe (i) how to collect the vlogs, (ii) how to annotate the collected vlogs, (iii) statistics of the labeled dataset, and (iv) how to extract acoustic and visual features.

Data Collection

We aim to build a dataset containing the balanced numbers of depression and non-depression vlogs thereby helping a model to learn unique depression-related features that are distinct from the non-depression features. To this end, we first collected YouTube videos, which have been posted between 1st January 2020 and 31st January 2021 (13 months), by using the following search keywords.

- **Depression Vlog:** 'depression daily vlog', 'depression journey', 'depression vlog', 'depression episode vlog', 'depression video diary', 'my depression diary', and 'my depression story'
- **Non-depression Vlog:** 'daily vlog', 'grwm (get ready with me) vlog', 'haul vlog', 'how to vlog', 'day of vlog ', 'talking vlog', and etc.

Using the two types of keyword sets, we could download 4,000 videos (i.e., 2,000 videos for each type) which are randomly selected from the search results. Note that we use YouTube Data API[3] and youtube-dl[4] for querying and downloading vlog videos from YouTube, respectively, and no common video is found in the search results from the depression and non-depression vlog keyword sets.

Data Annotation

To label the collected videos into depression and non-depression vlogs, we recruited four college students. We educate the annotators with the annotation criteria along with sample videos to ensure a similar level of annotation quality across all vlogs. We assign 1,000 videos (i.e., 500 vlogs per label) to each annotator and ask for two tasks as follows. First, annotators identify whether the given video has a 'vlog' format (Correia, Raj, and Trancoso 2018) where one person speaks directly to the camera. That is, we remove videos not in 'vlog' format (e.g., a group of people, no face on videos) so that the acoustic and visual features can be extracted from the person in videos. Next, the annotators carefully watch videos with automatically generated transcripts to identify whether speakers in the videos have depression or not. Note that we only consider videos in English since the annotators have to understand the context of the words. More specifically, the annotators label a given video as a depression vlog if the speaker shares his/her current state of depression symptoms at the moment of recording (e.g., "so many times I just want to kill myself ...", "I finally got switched on to my new medication ...", "I had suicidal thoughts ...", etc.). We do not consider speakers who (i) just share today's bad feeling without any depression-suspected phrases and (ii) suffered from depression in the past but already have overcome it, e.g., "how I overcame severe depression ...", "let me share my past experience about ...", based on the content of the vlogs.

[3]https://developers.google.com/youtube/v3

[4]https://github.com/ytdl-org/youtube-dl

Data Statistics

We present descriptive statistics of our final annotated dataset in Table 2. The dataset consists of 555 depression and 406 non-depression vlogs with an average length of 640.12 and 536.38 seconds, respectively. We find that there are twice more females than males in the depression vlogs, which is aligned with the prior work that showed major depression is more prevalent in women than men (Albert 2015). We also observe the same tendency in non-depression vlogs. This may be because our YouTube search queries include 'grwm vlog', 'haul vlog', etc., which are likely to be uploaded by female vloggers.

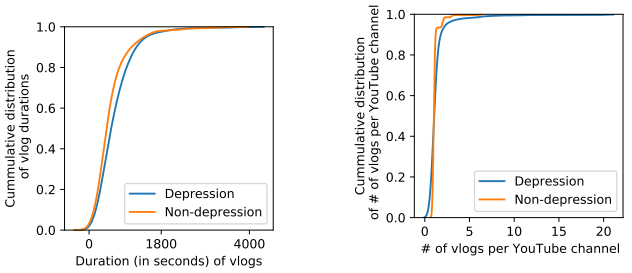
Table 2: Descriptive statistics of the D-Vlog dataset.

	Gender	# Samples	Avg. Duration
Depression	Male	182	583.74s
	Female	373	667.63s
Non-depression	Male	140	438.77s
	Female	266	587.76s

We next show the distributions of vlog duration and number of vlogs per YouTube channel in Figure 1. As shown in Figure 1a, the duration of both depression and non-depression vlogs show a heavy tail distribution. About 98.0% of both depression and non-depression vlogs have less than 30 minutes (1,800 seconds). Similarly, when we look at Figure 1b, both the distributions of the number of vlogs for depression and non-depression channels also show heavy tail distributions. Over 90.0% of YouTube channels have only one vlog. This implies that our dataset has variety in vloggers, which minimizes the chance of overfitting by identifying vloggers.

Non-verbal Feature Extraction

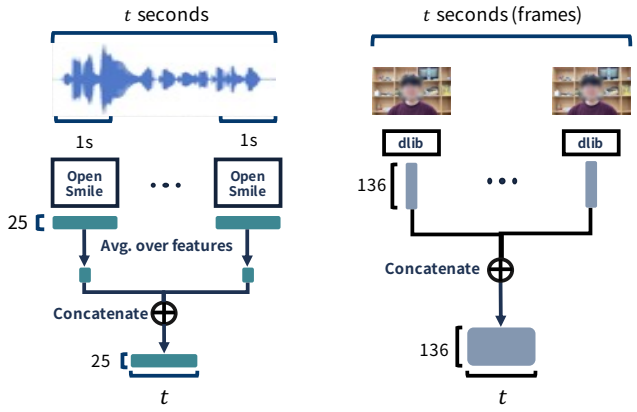
Acoustic Features: To extract audio features from a given vlog, we employ OpenSmile (Eyben, Wöllmer, and Schuller 2010), an open-source toolkit for audio processing, with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al. 2015). We first segment the audio sequence of each vlog into seconds (i.e., one segment = one second). From each audio segment, we extract 25 low-level acoustic descriptors (LLDs), which include loudness, MFCCs (Mel-frequency cepstral coefficients), spectral flux, etc. Note that we set the frame size and frame step as 0.06s



(a) Vlog duration

(b) Number of vlogs in a YouTube channel

Figure 1: Distributions of vlog duration and number of vlogs per YouTube channels in D-Vlog.



(a) Acoustic features

(b) Visual features

Figure 2: Feature extraction process of (a) acoustic and (b) visual features. We extract 25 acoustic features and 136 visual features for each second.

and 0.01s, respectively. We then average values of all 25 features in each segment and concatenate t segments to obtain audio feature vectors as illustrated in Figure 2a.

Visual Features: For visual feature extraction, we use dlib (King 2009), a well-known open-source software for computer vision tasks such as face recognition and verification. We extract 68 facial landmarks (i.e., x and y coordinates) for each frame (1 FPS) in the vlog, and obtain 136-dimensional vectors by concatenating each x and y coordinates. If the face is not detected in the frame, we replace it with zero vectors. The visual feature extraction process is illustrated in Figure 2b.

For both acoustic and visual features, we truncate timesteps or pad them with zeros to make all samples have the same length t. To de-identify each speaker in our dataset, we only provide extracted features instead of audio and video files. Hence, individuals cannot be identified with these features. Note that we followed all the anonymization processes guided by the Institutional Review Board (IRB)[5].

[5]SKKU IRB No. 2021-09-004

Multimodal Depression Detection

In this section, we first introduce the problem statement that describes the objectives of our proposed model, ***Depression Detector***. We then briefly present the overall architecture of the proposed model to detect depression in the given vlog. We finally describe details of the following components in the proposed model: (i) Unimodal Transformer Encoder, (ii) Multimodal Transformer Encoder, and (iii) Depression Detection Layer.

Problem Statement

The goal of the proposed model is to detect depression by learning non-verbal representations including acoustic and visual features of a speaker in a video. That is, our proposed model is defined as a binary classification problem that classifies a given vlog into depression or non-depression. Suppose we have a set of vlogs $P = \{p_n\}_{n=1}^{|P|}$ and each vlog can be represented as $p_n = (X_a^n \in \mathbb{R}^{t \times d_a}, X_v^n \in \mathbb{R}^{t \times d_v})$ where X_a^n, X_v^n, d_a, d_v, and t represent the acoustic features, the visual features, the dimension of acoustic features, the dimension of visual features, and the length of sequences, respectively. Note that both acoustic and visual sequences have the same length after alignment. Given a set of vlogs P, we classify the vlogs into corresponding labels (i.e., Depression or Non-depression) by learning latent features of speakers with symptoms of depression.

Overall Architecture

Figure 3 illustrates the overall architecture of the ***Depression Detector***. To leverage multimodal inputs of video, our model employs two encoders including the unimodal Transformer encoder and the multimodal Transformer encoder. More specifically, two unimodal Transformer encoders each take acoustic and visual feature vectors as inputs to generate unimodal representations. Next, the multimodal Transformer encoder colligates the acoustic and visual unimodal representations to make a final representation of the given vlog. By learning the multimodal representation, the depression detection layer finally predicts depression labels of the vlog.

Unimodal Transformer Encoder

The proposed model utilizes the unimodal Transformer encoder to generate representations of each input modality. Therefore, our proposed model is not limited to acoustic and visual features, but any other modalities can be simply added by employing the unimodal encoder. To model sequential input data, we first downsample feature vectors, process local relationships by applying 1-dimensional convolutional layers, and use a positional encoding layer. We then utilize the original Transformer encoder (Vaswani et al. 2017), which consists of self-attention and feed-forward layers. Since the self-attention layer guides the unimodal Transformer encoder to focus on significant cues within each modality, unimodal representations can benefit the model to capture useful knowledge for depression detection. The unimodal Transformer encoder generates unimodal representations $U_a^i \in \mathbb{R}^{t/4 \times d_u}$ and $U_v^i \in \mathbb{R}^{t/4 \times d_u}$ where d_u denotes

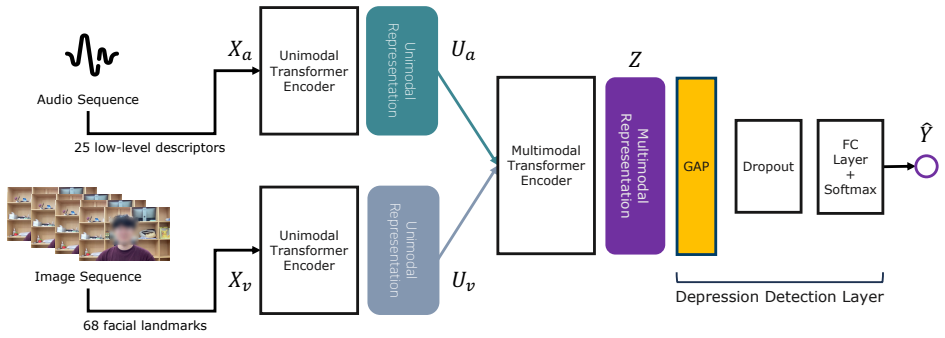


Figure 3: An illustration of the proposed model, *Depression Detector*.

the dimension of the unimodal representation. Equation 1 defines the attention module, where Q, K, V, and d_k denote query, key, value, and dimension of query/key, respectively. In self-attention, the same input feature is used as query, key, and value.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Multimodal Transformer Encoder

We next employ the multimodal Transformer encoder to learn important relationships across modalities. More specifically, we propose to use a cross-attention module, inspired by the prior work (Hasan et al. 2021), to incorporate visual and acoustic representations thereby understanding latent information between the modalities. That is, the multimodal Transformer encoder takes unimodal (i.e., acoustic and visual) representations U_a and U_v as inputs, and generates multimodal representations $Z \in \mathbb{R}^{t/4 \times 2d_u}$.

As shown in Figure 4, the encoder first takes acoustic and visual representations and passes them to the cross-attention module, where the source (query) and target (key/value) vectors are different. That is, one cross-attention (A $\rightarrow$ V) layer uses acoustic representation as its query and visual representation as its key/value, whereas another cross-attention layer (V $\rightarrow$ A) uses the opposite way. After residual connection and layer normalization, we obtain two cross representations $\tilde{U}_a$ and $\tilde{U}_v$. In this way, the cross-attention layer can characterize and emphasize the relationships across modalities.

To fuse the cross-modal information, we concatenate two different representations as follows:

$$\tilde{U}_{a,v} = \tilde{U}_a \oplus \tilde{U}_v. \quad (2)$$

Finally, the multimodal representation Z can be derived as:

$$Z = Transformer(\tilde{U}_{a,v}), \quad (3)$$

where $Transformer(\cdot)$ is a Transformer encoder that consists of self-attention and feed forward layers.

Depression Detection Layer

We finally add the depression detection layer to detect depression based on the multimodal representation Z. The ultimate logits of labels for depression detection can be inferred

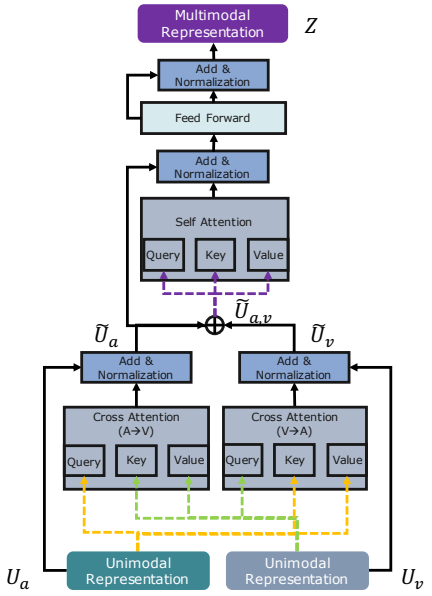


Figure 4: An illustration of the multimodal Transformer encoder.

as:

$$\hat{Y} = Softmax(\mathcal{F}(Dropout(GAP(Z)))), \quad (4)$$

where $GAP(\cdot)$, $Dropout(\cdot)$, $\mathcal{F}(\cdot)$, and $Softmax(\cdot)$ denote global average pooling, dropout, fully connected layer, and softmax activation function, respectively.

For depression detection, we define the task as a binary classification problem and utilize the cross entropy (Goodfellow, Bengio, and Courville 2016) as the loss function:

$$\text{loss} = \sum_{c \in [0,1]} P(c \mid \mathbf{y}) \log P(c \mid \hat{\mathbf{y}}), \quad (5)$$

where 0 and 1 are the Non-depression and the Depression classes, respectively; $P(c \mid \mathbf{y})$ is the ground truth label distribution; $P(c \mid \hat{\mathbf{y}})$ is the estimated probability for the label c by the logits $\hat{\mathbf{y}}$ and a softmax function.

Experiments

Experimental Settings

We use Tensorflow (Abadi et al. 2016) to implement our proposed model. As shown in Table 3, we split the dataset into the train, validation, and test sets with a 7:1:2 ratio. Note that these folds share no YouTube channel with each other. For training the model, Adam optimizer (Kingma and Ba 2014) is used, and we set batch size, epochs, learning rate, and sequence length (t) to 32, 50, 0.0002, and 596 (i.e., mean duration of the vlogs), respectively. All weights are randomly initialized in both proposed and baseline models.

Table 3: Number of samples in train, validation, and test folds of D-Vlog dataset.

Gender	Train	Val	Test
Male	216	40	66
Female	431	62	146

Baseline Models

To conduct extensive experiments for performance comparisons, we consider two categories of baseline methods: (i) Fusion and (ii) Model.

Fusion Baselines: Our proposed model suggests to apply the cross attention mechanism to effectively fuse multiple modalities. To evaluate the proposed fusion method, we compare with three commonly used multimodal fusion operations: (i) ***Concat*** (Kim, Lee, and Provost 2013; Liu, Zheng, and Lu 2016; Bargal et al. 2016), (ii) ***Add*** (Gunes and Piccardi 2007; Castellano, Kessous, and Caridakis 2008), and (iii) ***Multiply*** (Liu et al. 2018; Mittal et al. 2020) in fusing audio and visual representations. More precisely, we first simply replace the multimodal Transformer encoder in our proposed model with the above fusion operations. We then add the same depression detection layer in the ***Depression Detector***.

Model Baselines: To evaluate the overall performance of the proposed model, we compare with the following six methods: (i) Logistic Regression (***LR***), (ii) Support Vector Machines (***SVM***), (iii) Random Forest (***RF***), (iv) K-Nearest Neighbors based Fusion (***KNN-Fusion***) (Pampouchidou et al. 2017), (v) Bi-directional LSTM (***BLSTM***), and (vi) Tensor Fusion Network (***TFN***) (Zadeh et al. 2017). For ***LR***, ***SVM***, and ***RF***, we aggregate all features by flattening and concatenating vectors. ***KNN-Fusion*** is a machine-learning based fusion method for depression detection proposed in prior work (Pampouchidou et al. 2017). The model employs the decision-level fusion method, where classification results of each audio and visual classifier are combined through OR operands. For ***BLSTM***, we concatenate outputs of two different (i.e., acoustic and visual) bidirectional LSTM encoders, and add a fully connected layer with softmax activation function to detect depression. Note that recent studies on depression detection (Yin et al. 2019; Ray et al. 2019) have mostly employed the ***BLSTM*** in their models. ***TFN*** creates a multi-dimensional tensor to capture the unimodal, bimodal, and even trimodal interactions at once. We implemented ***TFN*** based on the open-source code available at: https://github.com/A2Zadeh/TensorFusionNetwork.

Experimental Results

Overall Performance: Table 4 shows the weighted average precision/recall/F1-score of the baseline models and the proposed model, respectively. As shown in Table 4, all fusion baseline models have lower performance than the proposed model. This indicates that the proposed fusion method using cross-attention can capture useful knowledge across the multiple modalities. Among the fusion baselines, ***Multiply*** model achieves higher performance at 63.09% of weighted average F1-score than other fusion baseline methods. This implies that multiplicative fusion can generate more informative multimodal representation than other fusion methods. Comparing with model baselines, we find that ***BLSTM*** and ***TFN*** have higher performance than the traditional machine-learning methods such as ***LR***, ***SVM***, and ***RF***. This is because ***BLSTM*** and ***TFN*** are designed to consider sequential information of the input features whereas the traditional machine-learning models fail to learn such information. Overall, the proposed model outperforms all the baselines by achieving 65.57% of weighted average recall and 63.50% of weighted average F1-score. This suggests that the proposed two-level Transformer encoders and the multimodal fusion method can generate representations that capture distinct characteristics of depressed speakers.

Analysis on Different Modalities: To analyze the importance of each modality (i.e., audio and visual) for detecting depression, we conduct a performance analysis of the unimodal models trained with each modality. For unimodal depression detection, we only use the unimodal Transformer encoder in Figure 3. We simply add global average pooling and fully connected layers with softmax activation function to the unimodal Transformer encoder. As shown in Table 5, the model trained with acoustic features achieves higher performances (58.55% of weighted average F1-score) than the model trained with visual features. That is, the acoustic features are more useful than the visual features in depression detection which can be linked to the results of the prior study (Pampouchidou et al. 2017). This implies that depressive speakers have distinctive features in their speech (Mundt et al. 2007). Although facial expressions are less effective to detect depression than acoustic features, we find that considering both modalities significantly improves the performance. This reveals that learning acoustic characteristics and facial expressions, as well as their relationships, is more effective than relying only on one modality.

Analysis on Different Gender: We next investigate how gender affects our depression detection. To this end, we first train and validate our model with a specific gender (i.e., male or female). Consistent with prior work (Yang et al. 2017b), the model achieves higher depression detection performance with non-verbal features in the male set than in the female set. This implies that male depressive speakers in vlogs tend to show distinct indicators in their speech and facial expressions which are different from male non-depressive speak-

Table 4: Performance comparisons between nine baseline models and the proposed model.

Model Type	Model	Precision	Recall	F1-Score
Fusion Baselines	***Concat***	62.51	63.21	61.10
	Add	59.11	60.38	58.11
	Multiply	63.48	64.15	63.09
Model Baselines	***LR***	54.86	54.72	54.78
	SVM	53.10	55.19	52.97
	RF	57.69	58.49	57.84
	KNN-Fusion	57.86	59.43	54.25
	BLSTM	60.81	61.79	59.70
	TFN	61.39	62.26	61.00
Proposed Model	***Depression Detector***	**65.40**	**65.57**	**63.50**

Table 5: Performance comparisons on unimodal depression detection models.

Modality	Precision	Recall	F1-Score
Audio	60.99	61.79	58.55
Visual	60.92	61.32	56.38
Both	65.40	65.57	63.50

ers, while it is difficult to differentiate the characteristics between female depressive and non-depressive speakers. We next evaluate our model by training with both genders and test with a specific gender. As shown in Table 6, we find that learning features from both genders increases the overall performance. This suggests that non-verbal features from different gender are useful in detecting depression.

Table 6: Gender differences in the model performance.

Train	Test	Precision	Recall	F1-Score
Male	Male	75.49	75.76	74.93
Female	Female	53.94	54.79	54.00
Both	Male	79.06	77.27	75.41
	Female	57.25	58.22	54.46

Cross-Corpus Validation with DAIC-WOZ: In this validation, we use our proposed dataset, D-Vlog, to detect depressions in DAIC-WOZ, which is a clinically labeled (e.g., PHQ-8) dataset, to analyze how the proposed D-Vlog can contribute to research on depression. We first extract the acoustic and visual features from DAIC-WOZ following the same feature extraction process used in D-Vlog for fair evaluations. We then conduct the following four experiments with the proposed model: (i) train and test the model with D-Vlog, (ii) train the model with the DAIC-WOZ (train fold), and test the model with the D-Vlog (test fold), (iii) train and test the model with DAIC-WOZ, and (iv) train the model with D-Vlog (train fold) and test with the DAIC-WOZ (test fold). Note that we randomly split DAIC-WOZ into 112, 37, and 38 samples for train, validation, and test folds, respectively.

Table 7 shows the results of four experiments. We find that the model trained with D-Vlog achieves higher depression detection performance than that with DAIC-WOZ in both datasets. This suggests that the features extracted in D-Vlog are more useful than those in the DAIC-WOZ dataset. This may be because D-Vlog contains more various acoustic and facial features captured in daily lives than DAIC-WOZ which has been developed in a laboratory setting. We believe the features extracted from D-Vlog can extend the research opportunity in detecting depressions from daily lives at an early stage.

Table 7: Cross-corpus validation results between D-Vlog and DAIC-WOZ datasets. DV and DW denote D-Vlog and DAIC-WOZ, respectively.

Train	Test	Precision	Recall	F1-Score
DW	DV	60.14	60.38	60.24
DV	DV	65.40	65.57	63.50
DW	DW	62.57	52.63	55.45
DV	DW	69.45	55.26	57.73

Conclusion

In this paper, we presented the multimodal depression dataset, D-Vlog, to detect depressed individuals based on non-verbal signals. Our dataset consists of acoustic and visual features extracted from 961 vlogs (i.e., around 160 hours) with 816 distinct speakers. We also proposed the Transformer-based multimodal deep learning model that uses a cross-attention mechanism to generate multimodal representations to detect depression. Our model takes acoustic and visual features extracted from a vlog as inputs and classifies whether the given vlog is a depressed individual or not. The proposed model outperforms other baseline models. We believe that the proposed dataset and the multi-modal depression detection model are useful for screening depressed individuals in social media at an early stage so that individuals who may need appropriate treatment can receive proper and timely clinical interventions.

Acknowledgement

This research was supported by the framework of international cooperation program managed by the National Research Foundation of Korea (NRF-2020K2A9A2A11103842) and the National Research Foundation (NRF) of Korea Grant funded by the Korean Government (MSIT) (No. 2021R1A4A3022102).

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Albert, P. R. 2015. Why is depression more prevalent in women? *Journal of Psychiatry & Neuroscience: JPN*, 40(4): 219.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *2018 IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(2): 423–443.

Bargal, S. A.; Barsoum, E.; Ferrer, C. C.; and Zhang, C. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, 433–436.

Becker, J. T.; Boiler, F.; Lopez, O. L.; Saxton, J.; and McGonigle, K. L. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6): 585–594.

Castellano, G.; Kessous, L.; and Caridakis, G. 2008. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, 92–103. Springer.

Correia, J.; Raj, B.; and Trancoso, I. 2018. Querying depression vlogs. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 987–993. IEEE.

Correia, J.; Teixeira, F.; Botelho, C.; Trancoso, I.; and Raj, B. 2021. The in-the-Wild Speech Medical Corpus. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6973–6977. IEEE.

Dibeklioğlu, H.; Hammal, Z.; and Cohn, J. F. 2017. Dynamic multimodal measurement of depression severity using deep autoencoding. *2017 IEEE Journal of Biomedical and Health Informatics*, 22(2): 525–536.

D'mello, S. K.; and Kory, J. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3): 1–36.

Edition, F.; et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21.

Eyben, F.; Scherer, K. R.; Schuller, B. W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L. Y.; Epps, J.; Laukka, P.; Narayanan, S. S.; et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *2015 IEEE Transactions on Affective Computing*, 7(2): 190–202.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia (MM)*, 1459–1462.

Ghosh, S.; Ekbal, A.; and Bhattacharyya, P. 2021. A multitask framework to detect depression, sentiment and multilabel emotion from suicide notes. *Cognitive Computation*, 1–20.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Gratch, J.; Artstein, R.; Lucas, G. M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, 3123–3128.

Gui, T.; Zhu, L.; Zhang, Q.; Peng, M.; Zhou, X.; Ding, K.; and Chen, Z. 2019. Cooperative multimodal approach to depression detection in Twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 110–117.

Gunes, H.; and Piccardi, M. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4): 1334–1345.

Hasan, M. K.; Lee, S.; Rahman, W.; Zadeh, A.; Mihalcea, R.; Morency, L.-P.; and Hoque, E. 2021. Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12972–12980.

Huang, Z.; Epps, J.; Joachim, D.; Stasak, B.; Williamson, J. R.; and Quatieri, T. F. 2020. Domain adaptation for enhancing Speech-based depression detection in natural environmental conditions using dilated CNNs. *Interspeech*, 4561–4565.

Kim, Y.; Lee, H.; and Provost, E. M. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP*, 3687–3691. IEEE.

King, D. E. 2009. Dlib-Ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.*, 10: 1755–1758.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *Proceedings of 3rd International Conference on Learning Representations (ICLR)*.

Lejuez, C. W.; Hopko, D. R.; and Hopko, S. D. 2001. A brief behavioral activation treatment for depression: Treatment manual. *Behavior Modification*, 25(2): 255–286.

Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*.

Liu, W.; Zheng, W.-L.; and Lu, B.-L. 2016. Emotion recognition using multimodal deep learning. In *International Conference on Neural Information Processing (ICONIP)*, 521–529. Springer.

Maddage, N. C.; Senaratne, R.; Low, L.-S. A.; Lech, M.; and Allen, N. 2009. Video-based detection of the clinical depression in adolescents. In *2009 IEEE Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3723–3726. IEEE.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1359–1367.

Mundt, J. C.; Snyder, P. J.; Cannizzaro, M. S.; Chappie, K.; and Geralts, D. S. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of neurolinguistics*, 20(1): 50–64.

Nicolaou, M. A.; Gunes, H.; and Pantic, M. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *2011 IEEE Transactions on Affective Computing*, 2(2): 92–105.

Orabi, A. H.; Buddhitha, P.; Orabi, M. H.; and Inkpen, D. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology (CLPsych: From Keyboard to Clinic*, 88–97.

Pampouchidou, A.; Simantiraki, O.; Vazakopoulou, C.-M.; Chatzaki, C.; Pediaditis, M.; Maridaki, A.; Marias, K.; Simos, P.; Yang, F.; Meriaudeau, F.; et al. 2017. Facial geometry and speech analysis for depression detection. In *2017 IEEE Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1433–1436. IEEE.

Ray, A.; Kumar, S.; Reddy, R.; Mukherjee, P.; and Garg, R. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, 81–88.

Rodrigues Makiuchi, M.; Warnita, T.; Uto, K.; and Shinoda, K. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, 55–63.

Senoussaoui, M.; Sarria-Paja, M.; Santos, J. F.; and Falk, T. H. 2014. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 57–63.

Sharifa, M.; Goecke, R.; Wagner, M.; Epps, J.; Breakspear, M.; Parker, G.; et al. 2012. From joyous to clinically depressed: Mood detection using spontaneous speech. In *Twenty-Fifth International FLAIRS Conference*.

Sobin, C.; and Sackeim, H. A. 1997. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1): 4–17.

Tadesse, M. M.; Lin, H.; Xu, B.; and Yang, L. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7: 44883–44893.

Üstün, T. B.; Ayuso-Mateos, J. L.; Chatterji, S.; Mathers, C.; and Murray, C. J. 2004. Global burden of depressive disorders in the year 2000. *The British Journal of Psychiatry*, 184(5): 386–392.

Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; and Pantic, M. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge (AVEC)*, 3–10.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and Pantic, M. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (AVEC)*, 3–10.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, X.; Zhang, C.; Ji, Y.; Sun, L.; Wu, L.; and Bao, Z. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 201–213. Springer.

Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; and Rigoll, G. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2): 153–163.

Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; and Narayanan, S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Interspeech*, 2362–2365.

Yang, L.; Jiang, D.; He, L.; Pei, E.; Oveneke, M. C.; and Sahli, H. 2016. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th international workshop on audio/visual emotion challenge (AVEC)*, 89–96.

Yang, L.; Jiang, D.; Xia, X.; Pei, E.; Oveneke, M. C.; and Sahli, H. 2017a. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 53–59.

Yang, L.; Sahli, H.; Xia, X.; Pei, E.; Oveneke, M. C.; and Jiang, D. 2017b. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 45–51.

Yin, S.; Liang, C.; Ding, H.; and Wang, S. 2019. A multimodal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, 65–71.

Yingthawornsuk, T.; Keskinpala, H. K.; France, D.; Wilkes, D. M.; Shiavi, R. G.; and Salomon, R. M. 2006. Objective estimation of suicidal risk using vocal output characteristics. In *9th International Conference on Spoken Language Processing (ICSLP)*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*.