DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

Multi-Task Representation Learning for Multimodal Estimation of Depression Level

Syed Arbaaz Qureshi
Sriparna Saha
Indian Institute of Technology Patna, India
Mohammed Hasanuzzaman
Cork Institute of Technology, Ireland
Gaël Dias
University of Caen Normandy, France
Editor:
Erik Cambria
Nanyang Technological University

Depression is a serious medical condition that is suffered by a large number of people around the world. It significantly affects the way one feels, causing a persistent lowering of mood. In this paper, we propose a novel multi-task learning attention-based deep neural network model, which facilitates the fusion of various modalities. In particular, we use this network to both regress and classify the level of depression. Acoustic, textual and visual modalities have been used to train our proposed network. Various experiments have been carried out on the benchmark dataset, namely, Distress Analysis Interview Corpus — a Wizard of Oz. From the results, we empirically justify that a) multi-task learning networks co-trained over regression and classification have better performance compared to single-task networks, and b) the fusion of all the modalities helps in giving the most accurate estimation of depression with respect to regression. Our proposed approach outperforms the state of the art by 4.93% on root mean squared error and 1.50% on mean absolute error for regression, while we settle new baseline values for depression classification, namely 66.66% accuracy and 0.53 F-score.

Depression is a serious medical illness that negatively affects how one feels. It is characterized by persistent sadness, loss of interest and an inability to carry out activities that one normally enjoys. It is the leading cause of ill health and disability worldwide. More than 300 million people are now living with depression, an increase of more than 18% between 2005 and 2015.[1]

The causes of depression are not completely known and they may not be attributed to a single source. Major depressive disorder is likely to be due to complex combinations of factors like genetics, psychology, and social surroundings of the sufferer. People who have experienced life events like divorce or death of a family member or friend, who have personality issues such as the inability to deal with failure and rejection, people with previous records of major depression, and with childhood trauma are at a higher risk of depression[1].

Depression detection is a challenging problem as many of its symptoms are covert. Since depressed people socialize less, its detection becomes difficult. Today, for the correct diagnosis of depression, patients are evaluated on standard questionnaires. Different tools for screening depression have been proposed in the literature. Some of them are the Personal Health Questionnaire Depression Scale (PHQ), the Hamilton Depression Rating Scale (HDRS), the Beck Depression Inventory (BDI), the Center for Epidemiologic Studies Depression Scale (CES-D), the Hospital Anxiety and Depression Scale (HADS), and the Montgomery and Asberg Depression Rating Scale (MADRS).[2] In particular, the eight-item PHQ-8[2] is established as a valid diagnostic and severity measure for depressive disorders in many clinical studies[3].

The steadily increasing global burden of depression and mental illness acts as an impetus for the development of more advanced, personalized and automatic technologies that aid in its detection. Affective computing is one field of research, which focuses on gathering data from faces, voices and body languages to measure human sentiment[4,5] and emotion. An important business goal of affective computing is to build human-computer interfaces that can detect and appropriately respond to an end user's state of mind. As a consequence, techniques from affective computing have been applied for the automatic detection of depression[6].

In this paper, we introduce a multi-task neural network for modality encoding and an attention-based neural network for the fusion of all the acoustic, textual and visual modalities. In particular, we encode six modalities (two acoustic, one textual and three visual). The tasks of the modality encoders are depression level regression and depression level classification. The acoustic, textual and visual modality embeddings are then fed to an attention fusion network to obtain fused vectors. These fused vectors are in turn passed to a deep regression network to predict the severity of depression based on a PHQ-8 scale. From our experiments, we show that:

- Multi-task representation learning networks for depression level regression and classification evidence better performances compared to single-task representation networks,
- The fusion of all modalities (acoustic, textual, visual) helps in better estimation of depression level, to the exception of classification,
- Our approach outperforms the previous state of the art by 4.93% on root mean squared error and 1.50% on mean absolute error for the regression of depression level, and
- The verbal input plays a predominant role in the estimation process, confirming therapists' experience.

[1] A statistic reported by the World Health Organization available at https://bit.ly/2rsqQoP.
[2] Recommendation of the French Haute Autorité de la Santé available at https://bit.ly/2EaOs92.

LITERATURE SURVEY

Over the last few years, a great deal of research studies in computer science have been proposed to deal with mental health disorders [7]. Within this context, the automatic detection of depression has received major focus. Some initial initiatives have targeted the understanding of relevant descriptors that could be used in machine learning frameworks. Scherer et al. [8] investigated the capabilities of automatic non verbal behavior descriptors to identify indicators of psychological disorders such as depression. In particular, they propose four descriptors that can be automatically estimated: downward angling of the head, eye gaze, duration and intensity of smiles, and self-touches. Cummins et al. [9] focused on how common para-linguistic speech characteristics (prosodic features, source features, formant features, spectral features) are affected by depression and suicidality. Morales et al. [10] argued that researchers should look beyond the acoustic properties of speech by building features that capture syntactic structure and semantic content. Within this context, Wolohan et al. [11] showed that overall classification performance suggests that lexical models are reasonably robust and well suited for a role in a diagnostic or the monitoring capacity of depression. Some other interesting work directions using text features include the study of social media [12], document modeling for personality detection from text [13], eventually using specific corpora tuned for such tasks [14].

Another promising research trend aims at leveraging all modalities into one learning model and is commonly called multimodal deep learning [15]. Multimodal deep learning approaches have been used in domains such as sentiment analysis [16,17] and depression estimation [18]. Within the context of depression estimation, a great deal of successful research studies have been proposed. He et al. [19] evaluated feature fusion and model fusion strategies via local linear regression to improve accuracy in the BDI score using visual and acoustic cues. Dibeklioglu et al. [20] compared facial movement dynamics, head movement dynamics, and vocal prosody individually and in combination, and showed that multimodal measures afford most powerful detection. More recently, Morales et al. [6,18] proposed an extensive study of fusion techniques (early, late and hybrid) for depression detection combining audio, visual and textual (especially syntactic) features, through Support Vector Machine. In particular, they showed that the syntax-informed fusion approach is able to leverage syntactic information to target more informative aspects of the speech signal, but the overall results tend to suggest that there is no statistical evidence of this finding.

METHODOLOGY

Our proposed architecture (**MT. CombAtt**) consists of three main components: a) multi-task learning modality encoders, which take unimodal features as input, and output modality embeddings, where tasks are regression and classification, b) an attention-based fusion network that fuses the individual modalities, and c) a deep neural network that outputs the estimated PHQ-8 score or classifies patients into medically-motivated classes, conditioned on the output of the attention fusion network.

Multi-Task Learning Modality Encoders

We use six different modalities for the estimation of depression. All these modalities are encoded using a multi-task learning network, where the two tasks are Depression Level Regression (DLR) and Depression Level Classification (DLC). For DLC, we discretize the PHQ-8 score following medical scales (i.e., none/minimal - [0-4] PHQ-8 score, mild - [5-9] PHQ-8 score, moderate - [10-14] PHQ-8 score, moderately severe - [15-19] PHQ-8 score, severe - [20-24] PHQ-8 score), while for regression, PHQ-8 scores are directly predicted. In particular, we trained two standard multi-task learning architectures, which are shown in Figure 1: a) the fully-shared multi-task

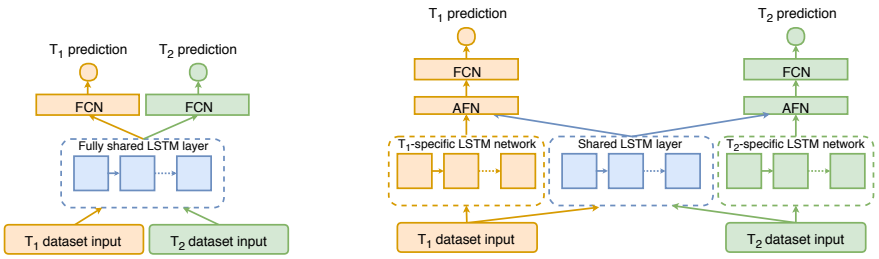

Figure 1. Fully-Shared (left) and Shared-Private (right) Multi-task Architectures.

architecture, which has a single recurrent layer (or a single stack of recurrent layers), that acts as the shared layer for both tasks, b) the shared-private architecture, which has three recurrent layers, two task-specific (one for DLR and one for DLC) and one shared.

Each of the modality encoders is trained as a separate network through backpropagation. They are jointly trained to regress and classify depression level, and the outputs of the respective recurrent layers act as the continuous representation (i.e. embedding) of each modality. To test the advantages of the multi-task learning networks over the single-task networks, we trained two sets of single-task networks, one for DLC and the other one for DLR. The results given in section 4 show that there is a clear improvement (in Root Mean Squared Error - RMSE and Mean Average Error - MAE for DLR, and in Accuracy and F-score for DLC) of the multi-task encoders over the corresponding single-task encoders.

To encode the time series data of each modality, we use long short-term memory (LSTM) networks, with forget gates [21] in the recurrent layer, because of their robustness in capturing long sequences. For the single-task and the fully-shared multi-task networks, the output from the LSTM layer acts as the encoding vector for a given modality. For the shared-private network, a concatenation of the outputs from the shared-LSTM and the DLR (or DLC) task-specific LSTM acts as the embedding for the input modality. Note that for each modality, we select the encoder that gives the lowest RMSE (or F-score for DLC) as its modality encoder. These modality embeddings are then fed to the attention fusion network.

Attention Fusion Network

The attention fusion network uses the attention mechanism to automatically weight the modality embeddings. By using an attention mechanism, the network "attends" to the most relevant part of the input to generate the output. Networks with an attention mechanism usually perform better than their counterpart without attention. As not all modalities are equally relevant for the estimation of depression level, this motivates the introduction of an attention fusion network, as an extension of the work of Poria et al. [22].

The six input modality embeddings are of different lengths. So, we pass all of them through a fully-connected layer to get a common length representation. These vectors are then stacked vertically, and passed through a fully-connected network. The output layer of this network is a 6-D softmax unit (i.e. one dimension for each modality). The elements in the 6-D output vector, which we call α-values, represent the importance of the corresponding modality embeddings. We then multiply the modality embeddings with their α-values, and take the sum of the resultant six vectors. This sum is the fusion **F** of six modality embeddings.

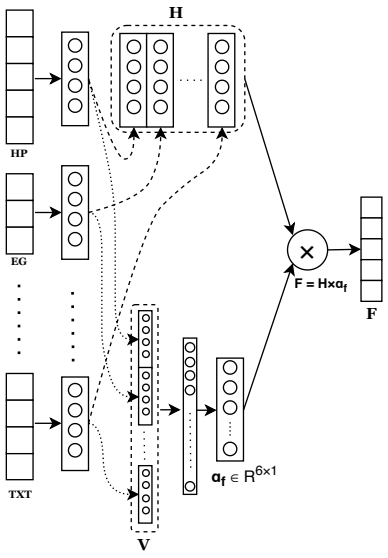

Figure 2. Attention Fusion Network.

PHQ-8 Score Estimation Network

The PHQ-8 score estimation network is a deep network conditioned on the fusion vector **F**, the output of the attention fusion network. In particular, **F** is fed to a few dense and dropout layers. The resultant vector is finally fed to a) a linear regression unit, which outputs the PHQ-8 score for regression or b) a linear classification unit, which outputs the PHQ-8 class for classification.

DAIC-WOZ DEPRESSION DATASET

The DAIC-WOZ depression dataset[3] is part of a larger corpus, the Distress Analysis Interview Corpus[23], that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and non-verbal indicators of mental illness. The data collected include audio and video recordings, and extensive questionnaire responses from the interviews conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. The data has been transcribed and annotated for a variety of verbal and non-verbal features.

The dataset contains 189 sessions of interviews. We discarded a few interviews, as some of them were incomplete and others had interruptions. Each interview is recognized by a unique ID assigned to it. Each interview session contains a raw audio file of the interview session, files containing the coordinates of 68 facial landmarks of the participant, HoG (Histogram of oriented Gradients) features of the face, head pose, eye gaze features of the participant (recorded over the entire duration of interview using a framework named OpenFace[24]), a file containing the continuous facial action units of the participant's face extracted using the facial action coding software CERT[25], the COVAREP and formant feature files of the participant's voice extracted using a framework named COVAREP[26], and a transcript file of the interview. All the features, leaving the transcript file, are time series data.

[3]http://dcapswoz.ict.usc.edu/.

The training, development and test split files are provided with the dataset. The training and development split files comprise of interview IDs, PHQ-8 binary labels, PHQ-8 scores, participant's gender, and single responses to every question of the PHQ-8 questionnaire. The test split file comprises of interview IDs and participant's gender. We use only the training and the development split, as the labels are not provided for the test split.

EXPERIMENTAL RESULTS

We used RMSE and MAE to evaluate the regression performance (DLR), and Accuracy and F-score to evaluate the classification performance (DLC). Table 1 illustrates a) the advantages of multi-task representation learning on DLC and DLR over the single-task encodings, and b) the estimation gains by the fusion of the six modalities using the attention fusion network.

Multi-task learning on DLR and DLC shows steady improvements in terms of performance metrics compared to the corresponding single-task networks, implying that better encodings can be obtained from coupling both classification and regression into a joint model. We observe this behaviour for all the modality encoders, where the fully-shared or the shared-private architectures systematically outperform the single-task model. In particular, this motivates us to test this approach on different datasets in order to verify generalization issues.

In order to test multimodal issues, we designed four attention-based networks: two based on single-task representation learning (**ST. DLR. CombAtt** and **ST. DLC. CombAtt**) and two others based on multi-task representation learning (**MT. DLR. CombAtt** and **MT. DLC. CombAtt**). In particular, the models were fed with the best corresponding embedding for each of the modality (based on RMSE for DLR and F-score for DLC). So, for instance, **MT. DLC. CombAtt** received the following encodings as input: **FS. MT. DLC. HP** (Head Pose), **FS. MT. DLC. EG** (Eye Gaze), **SP. MT. DLC. AU** (Action Unit), **FS. MT. DLC. COV** (COVAREP), **FS. MT. DLC. FMT** (Formant) and **FS. MT. DLC. TXT** (Text). On the one hand, regression results show that **MT. DLR. CombAtt** evidences an improvement of 4.06% in RMSE and 4.91% in MAE over the single-task encoding **ST. DLR. CombAtt**. On the other hand, classification results show that **MT. DLC. CombAtt** presents an increase of 3.03% in Accuracy and an improvement of 4.25% in F-score over the single-task-based representation **ST. DLC. CombAtt**. Note however, that while the combination of all modalities improves results for regression, best results for classification are obtained on text features only. Overall, results evidence that the verbal input plays a predominant role in the estimation process, confirming therapists' experience.

Finally, we compare our best architecture for regression (**MT. DLR. CombAtt**) with three state-of-the-art (SOTA) approaches: **VFSC$_{\text{sem}}$** [27], **AW$_{\text{bhv}}$** [28], and **MMD** [29]. Note that we are the first to include depression classification as an extra task. The SOTA methodologies are briefly explained for the sake of comparison. In **VFSC$_{\text{sem}}$**, the authors derive biomarkers from visual, acoustic and textual modalities. They define semantic context indicators, which use the transcripts to infer a subject's status with respect to four conceptual classes. The semantic context feature is the sum of points accrued from all four indicators. This approach is the state of the art on the official development split of DAIC-WOZ. **AW$_{\text{bhv}}$** is the winning approach in AVEC 2017 [30] depression sub-challenge. The authors use feature extraction methods on acoustic and text features, and recurrent neural networks on visual features. It is the current state of the art on the test split of DAIC-WOZ. The authors developed four models, but we compare our approach with the best of these four models (**AW$_{\text{bhv}}$**). In **MMD**, the authors propose a multimodal fusion framework composed of deep convolutional neural network (DCNN) and deep neural network (DNN) models. The framework considers acoustic, textual and visual streams of data. For each modality, handcrafted feature descriptors are fed to a DCNN that learns high-level global features with compact dynamic information. Then, the learned features are fed to a DNN to predict the PHQ-8 scores.

Table 1. Overall results. ST: Single-task, MT: Multi-task, FS: Fully-Shared, SP: Shared-Private, DLC: Depression Level Classification, DLR: Depression Level Regression, HP: Head Pose, EG: Eye Gaze, AU: Action Units, COV: COVAREP, FMT: Formant, TXT: Text.

	Architectures	RMSE	MAE	Acc (%)	F-score
Unimodal	**ST. DLR. HP**	6.89	5.67	-	-
	ST. DLC. HP	-	-	54.54	0.41
	FS. MT. HP	6.75	**5.48**	**60.60**	**0.43**
	SP. MT. HP	**6.65**	5.53	54.54	0.42
	ST. DLR. EG	6.67	4.72	-	-
	ST. DLC. EG	-	-	54.54	0.37
	FS. MT. EG	**6.50**	**4.60**	**57.57**	**0.41**
	SP. MT. EG	6.59	5.16	57.57	0.39
	ST. DLR. AU	6.49	5.55	-	-
	ST. DLC. AU	-	-	54.54	0.42
	FS. MT. AU	**6.28**	**5.03**	54.54	0.44
	SP. MT. AU	6.46	5.42	**57.57**	**0.45**
	ST. DLR. COV	6.64	5.72	-	-
	ST. DLC. COV	-	-	51.51	0.36
	FS. MT. COV	**6.55**	**5.67**	**54.54**	**0.40**
	SP. MT. COV	6.59	5.71	54.54	0.37
	ST. DLR. FMT	6.91	5.89	-	-
	ST. DLC. FMT	-	-	51.51	0.34
	FS. MT. FMT	6.72	**5.77**	**54.54**	**0.36**
	SP. MT. FMT	**6.69**	5.79	51.51	0.34
	ST. DLR. TXT	4.90	3.99	-	-
	ST. DLC. TXT	-	-	60.60	0.45
	FS. MT. TXT	4.96	3.90	**66.66**	**0.53**
	SP. MT. TXT	**4.70**	**3.81**	60.61	0.42
Multimodal	**ST. DLR. CombAtt**	4.42	3.46	-	-
	MT. DLR. CombAtt	**4.24**	**3.29**	-	-
	ST. DLC. CombAtt	-	-	57.57	0.46
	MT. DLC. CombAtt	-	-	**60.61**	**0.48**
SOTA	**VFSC$_{sem}$**	4.46	3.34	-	-
	AW$_{bhv}$	5.54	4.73	-	-
	MMD	4.65	3.98	-	-

For multimodal fusion, the estimated PHQ-8 scores from the three modalities are integrated in another DNN to obtain the final PHQ-8 score. In our experiment, comparative results illustrate that although **VFSC$_{sem}$** is the best performing SOTA model, our **MT. DLR. CombAtt** methodology evidences increased results of 4.93% on RMSE and 1.50% on MAE. Note that on conducting statistical significance tests, we observe that our results are statistically significant.

CONCLUSION

In this paper, we showed that multi-task representation learning, where tasks are regression and classification performs better than single-task representation learning for the estimation of depression level. Moreover, this situation stands both for uni-modal and multimodal inputs. In particular, this motivates us to further explore the multi-task learning approach with regression and classification tasks on other datasets to verify whether generalization issues can be found. But mostly, we evidenced that our models, either multimodal for regression and text-based for classification, present the new state of the art results over the DAIC-WOZ dataset.

REFERENCES

1. A. T. Beck and B. A. Alford, *Depression: Causes and treatment*, University of Pennsylvania Press, 2009.
2. K. Kroenke et al., "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, 2008, pp. 163–73.
3. K. Kroenke, "Enhancing the clinical utility of depression screening," *Canadian Medical Association Journal*, vol. 184, no. 3, 2012, pp. 281–282.
4. E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, 2016, pp. 102–107.
5. E. Cambria et al., "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, 2017, pp. 74–80.
6. M. Morales, S. Scherer, and R. Levitan, "A linguistically-informed fusion approach for multimodal depression detection," *CLPsych: From Keyboard to Clinic*, 2018, pp. 13–24.
7. G. Andersson and N. Titov, "Advantages and limitations of Internet-based interventions for common mental disorders," *World Psychiatry*, vol. 13, 2014, pp. 4–11.
8. S. Scherer et al., "Automatic behavior descriptors for psychological disorder analysis," *IEEE FG*, 2013, pp. 1–8.
9. N. Cummins et al., "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, 2015, pp. 10–49.
10. M. R. Morales and R. Levitan, "Speech vs. text: a comparative analysis of features for depression detection systems," *IEEE SLT*, 2016, pp. 136–143.
11. J. Wolohan et al., "Detecting linguistic traces of depression in topic-restricted text: attending to self-stigmatized depression with NLP," *LCCM*, 2018, pp. 11–21.
12. M. De Choudhury et al., "Predicting depression via social media," *ICWSM*, 2013, pp. 128–137.
13. N. Majumder et al., "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, 2017, pp. 74–79.
14. D. E. Losada and F. Crestani, "A test collection for research on depression and language use," *CLEF*, 2016, pp. 28–39.
15. J. Ngiam et al., "Multimodal deep learning," *ICML*, 2011, pp. 689–696.
16. S. Poria et al., "Multimodal sentiment analysis: addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, 2018, pp. 17–25.
17. A. Zadeh et al., "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, 2016, pp. 82–88.
18. M. R. Morales, *Multimodal Depression Detection: An Investigation of Features and Fusion Techniques for Automated Systems*, Ph.D. thesis, City University of New York, 2018.
19. L. He, D. Jiang, and H. Sahli, "Multimodal depression recognition with dynamic visual and audio cues," *ACII*, 2015, pp. 260–266.
20. H. Dibeklioğlu et al., "Multimodal detection of depression in clinical interviews," *ICMI*, 2015, pp. 307–310.
21. F. A. Gers, J. Schmidhuber, and F. A. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, 2000, pp. 2451–2471.
22. S. Poria et al., "Multi-level multiple attentions for contextual multimodal sentiment analysis," *ICDM*, 2017, pp. 1033–1038.
23. J. Gratch et al., "The Distress analysis interview corpus of human and computer interviews," *LREC*, 2014, pp. 3123–3128.
24. T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," *WACV*, 2016, pp. 1–10.
25. G. Littlewort et al., "The computer expression recognition toolbox (CERT)," *IEEE FG*, 2011, pp. 298–305.

26. G. Degottex et al., “COVAREP-A collaborative voice analysis repository for speech technologies,” *ICASSP*, 2014, pp. 960–964.
27. J. R. Williamson et al., “Detecting depression using vocal, facial and semantic communication cues,” *AVEC*, 2016, pp. 11–18.
28. E. Stepanov et al., “Depression Severity Estimation from Multiple Modalities,” *HEALTHCOM*, 2018, pp. 1–6.
29. L. Yang et al., “Multimodal measurement of depression using deep learning models,” *AVEC*, 2017, pp. 53–59.
30. F. Ringeval et al., “Avec 2017: real-life depression, and affect recognition workshop and challenge,” *AVEC*, 2017, pp. 3–9.

ABOUT THE AUTHORS

Syed Arbaaz Qureshi is a Graduate Student in computer science and engineering at the Indian Institute of Technology Patna, India, and researcher at the University of Caen Normandie, France. Contact him at arbaaz.qureshi@unicaen.fr.

Sriparna Saha is an Associate Professor in computer science and engineering at the Indian Institute of Technology Patna, India. Contact her at sriparna@iitp.ac.in.

Mohammed Hasanuzzaman is an Assistant Professor in computer science at Cork Institute of Technology, Ireland. Contact him at mohammed.hasanuzzaman@adaptcentre.ie.

Gaël Dias is a Full Professor in computer science at the University of Caen Normandy, France. Contact him at gael.dias@unicaen.fr.