Improving Depression Level Estimation by Concurrently Learning Emotion Intensity

Syed Arbaaz Qureshi and Sriparna Saha, Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
Gaël Dias, Department of Computer Science, University of Caen Normandie, Caen, France
Mohammed Hasanuzzaman, Department of Computer Science, Cork Institute of Technology, Cork, Ireland.

Abstract—Depression is considered a serious medical condition and a large number of people around the world are suffering from it. Within this context, a lot of studies have been proposed to estimate the degree of depression based on different features and modalities, specific to depression. Supported by medical studies that show how depression is a disorder of impaired emotion regulation, we propose a different approach, which relies on the rationale that the estimation of depression level can benefit from the concurrent learning of emotion intensity. To test this hypothesis, we design different attention-based multi-task architectures that concurrently regress/classify both depression level and emotion intensity using text data. Experiments based on two benchmark datasets, namely, the Distress Analysis Interview Corpus - a Wizard of Oz (DAIC-WOZ), and the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) show that substantial performance improvements can be achieved when compared to emotion-unaware single-task and multi-task approaches.

1 Introduction

Depression is a common mental disorder that causes people to experience depressed mood, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, low energy, and poor concentration [1]. It is the predominant mental health problem worldwide, followed by anxiety, schizophrenia and bipolar disorder [2]. In 2013, depression was the second leading cause of years lived with a disability worldwide, and in 26 countries, depression was the primary driver of disability [2]. More than 300 million people are now living with depression, an increase of more than 18% between 2005 and 2015.[1]

Depression lasts between 4 and 8 months on average and can actually change one's ability to think, impair attention and memory, as well as debilitate information processing and decision-making skills. It can also lower one's cognitive flexibility and executive functioning. As a consequence, in extreme cases, depression may be characterized by thoughts of death and suicide. Approximately 800,000 people suffering from depression die due to suicide yearly and the annual number of death cases due to depression is on the rise.[2]

There are many possible causes of depression, including faulty mood regulation (for example, inability to deal with failure and rejection), genetic vulnerability, stressful life events (for example, divorce, death of a family member, childhood trauma), and medical problems. It is believed that several of these forces interact to bring on depression [3].

A depression diagnosis is often difficult to make because clinical depression can manifest in many different ways. Observable or behavioral symptoms of clinical depression also may sometimes be minimal despite a person experiencing profound inner turmoil. Diagnosis of depression has traditionally been made based on clinical criteria, including patient current symptoms and history. This process is widely used but relies on subjective interpretation. To standardize both the data obtained and data interpretation, various interview-based instruments and non-interview methods exist for screening and testing for depression in various clinical settings [4]. In particular, interview-based screening tools include 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).[3]

The Patient Health Questionnaire (PHQ) [5] has been established as a valid diagnostic and severity measure for depressive disorders [6]. In particular, PHQ-8 contains eight questions, whose answers range from 0 (not at all) to 3 (nearly every day), to provide an overall mark between 0 and 24, that estimates the level of depression. Different versions of the PHQ exist, such as PHQ-8, PHQ-9 and PHQ-15, containing 8, 9 and 15 questions respectively. The PHQ-9 is the most widely used questionnaire [6], but researchers generally use PHQ-8, which consists of all the PHQ-9 questions except for the last one (a question on suicidal thoughts). The absence of the ninth question has little effect on scoring between the PHQ-8 and PHQ-9. Studies found that scores between the two tests are highly correlated [7].

However, filling these forms is a tedious task that can be perceived as insuperable by many patients, thus leading to a

. Corresponding Author: Gaël Dias (Email: gael.dias@unicaen.fr).

1. A statistic reported by the World Health Organization, available at https://bit.ly/2rsqQoP.

2. A study by Hannah Ritchie and Max Roser in 2018, available at https://bit.ly/2mnyVZ6.

3. Recommendation of the French Haute Autorité de la Santé For more information, go to https://bit.ly/2EaOs92.

great deal of medically unfollowed patients. Moreover, due to the increasing number of patients suffering from mental health diseases, the average time for a medical consultation has drastically decreased over the last decade, leading to both patients' and therapists' frustration and limiting the number of interview-based screening acts.

Effective treatments for depression are available, however, only fewer than half of those affected in the world undergo with such treatments. In some countries, this number can go down to less than 10%. Possible reasons for this may be lack of resources, lack of trained health-care providers, social stigma associated with mental disorders and also an inaccurate assessment. Simultaneously, people who are depressed may not be correctly diagnosed, and others who do not have the disorder are too often misdiagnosed and prescribed antidepressants. The above facts prove that there is a steadily increasing global burden of depression and mental illness. Thus development of more advanced, personalized and automatic technologies for the detection and estimation of depression is highly essential.

In order to help therapists in their diagnosis, a great deal of studies have been proposed for the automatic estimation of depression level based on different features over various modalities, such as text, vision and acoustics [8], [9], [10]. All these methodologies focus on the improvement of single-task learning models, trying to increase the performance by better characterizing depression itself. However, some studies in mental health have shown that depression is a disorder of impaired emotion regulation [11], [12]. In particular, patients with major depression are often unable to control their emotional responses to negative situations, and overuse emotional expressions of sadness, disgust or fear. As a consequence, we hypothesize that the estimation of depression level can benefit from the concurrent learning of emotion intensity, which can be evaluated on a [0,3] scale for the six emotions of Ekman [13] - happiness, sadness, anger, fear, disgust and surprise. So, we propose to use the text data provided in the interviews of the different datasets (depression and emotion) to concurrently estimate depression level and emotion intensity, expecting that both tasks have common backgrounds and can boost performance over single-task processing. For illustration, we show below sentences that are indicative of depression.

Interviewer: *"How easy was it for you to get used to living in Los Angeles?"*
Participant: *"It was not easy for me. It took about three years."*
Interviewer: *"Can you tell me about that?"*
Participant: *"Umm... just the move. I moved away from my family so I was uncomfortable. I didn't know anyone here and even though I did make friends I just felt out of place."*

To test our hypothesis, we particularly explore three different multi-task architectures that concurrently regress/classify both depression level and emotion intensity using textual modality exclusively. Thus, (1) the fully-shared, (2) the shared-private and (3) the adversarial shared-private models are designed, following the ideas of [14]. However, we include an attention layer in the last two models, to let the network decide by itself the weights of the private and shared representations in the decision process. We extend the multi-task architectures to three tasks, which include depression level regression, depression level classification[4] and emotion intensity regression, thus extending the ideas of [15], who have shown that depression level classification and depression level regression can be complementary in the decision process.

An exhaustive series of experiments using these models are carried out using two benchmark datasets: the Distress Analysis Interview Corpus - a Wizard of Oz (DAIC-WOZ) [16], and the Carnegie Mellon University - Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [17]. Although both datasets contain multimodal (text, vision, acoustics) information, we exclusively focus on the text modality, as [9] showed that lexical models perform reasonably well to monitor depression level. Overall results for both depression level classification and regression show that notable performance gains can be obtained by emotion-aware models, when compared to emotion-unaware single-task (ST) and multi-task (MT) baseline approaches.

With such studies, we expect that in a near future systems can be built that automatically detect depression, thus playing a great role in supporting the therapist's diagnosis. Such applications may also help in early detection of clinical depression by suggesting the sufferer to consult a psychiatrist. We anticipate that our work may reduce cases of late treatment for clinical depression.

2 Related Works

Due to the impulse for the development of automatic technologies that can aid the detection of mental health disorders, a great deal of research studies in Computer Science have been emerging over the past few years [18]. Within this particular context, the automatic detection of depression level has received major focus.

Initial initiatives targeted the understanding of relevant non-verbal descriptors that could be used in machine learning frameworks such as gaze, smile, self-touches and heart-rate descriptors [19]. Other non-verbal descriptors include acoustics. Within this context, [8] focused their research on finding how common paralinguistic speech characteristics are affected by depression, namely prosodic, source, formant and spectral features. With respect to verbal descriptors, [20] hypothesized that researchers should look beyond the acoustic properties of speech and build features that capture syntactic structures and semantic contents. Following these ideas, [9] showed that classification performance suggests that lexical models are reasonably robust to play an important role in the diagnosis and monitoring of depression. But, the analysis also suggests that users may be able to fool algorithms by avoiding direct discussion of depression. Some other interesting work directions using text features include the study of social media [21], eventually using specific corpora tuned for such tasks [22].

More recently, new solutions proposed to combine verbal and non-verbal descriptors (or modalities) within a single learning model [10], [23]. Although the idea is seducing as it can be viewed as a way of avoiding fooling behaviors, the first results were mitigated [24]. But, recent studies [25]

4. For that purpose, discretization follows medical scales.

evidence successful results. It is also interesting to notice that non-Deep Learning approaches have been proposed but with less successful results [26]. This may suggest that Deep Learning techniques are able to capture high-level features and long-term dependencies at levels not seen before.

All previous related works focus on finding better descriptions of depression characteristics based on verbal and/or non-verbal indicators. In this paper, we aim to investigate the effect of simultaneously learning related tasks such as depression level and emotion intensity estimation. As stated in the study of [27], simultaneous learning of every task combination is not beneficial, but, tasks having cognitive similarities often get benefited from concurrent learning. In recent years, multi-task learning frameworks have become powerful in solving different NLP tasks [27], [28]. The possible reasons for this success (i.e. learning the decision boundaries of related tasks) are : (1) knowledge transfer across tasks in the form of generating more robust representations and (2) the use of more training data. In [29], it has also been discussed that multi-task learning can act as a regularization process which avoids overfitting by maintaining competitive performance across different tasks. In particular, a multi-task framework has recently been proposed by [15] who explore in concurrently learning depression level classification and regression. Inspired by the success of such models, we propose to compare three multi-task learning models (fully-shared, shared-private and adversarial shared-private) that combine three concurrent tasks: depression level classification, depression level regression and emotion intensity regression. Indeed, as depression can be viewed as the impaired regulation of emotion intensity, it is likely that better models can be built based on the concurrent learning of depression level and emotion intensity estimations.

3 METHODOLOGY

In order to estimate the level of depression and the intensity of emotions concurrently, we propose three different multi-task architectures that take as input the transcript files from the DAIC-WOZ [16] and the CMU-MOSEI datasets [17]. These datasets are described in detail in section 4. In the following subsections, we describe the tasks to be handled, the preprocessing steps and the multi-task architectures.

3.1 Learning Tasks

In this section, we define the three tasks used in our experiments: depression level regression, depression level classification and emotion intensity regression.

Depression Level Regression (DLR). Given the interview transcript associated with a patient, we predict its PHQ-8 score. This can be modeled as a simple regression task, where a score in the range of [0-24] must be predicted.

Depression Level Classification (DLC). In this task, we discretize the PHQ-8 score, which ranges from 0 to 24, into five classes of equal length: [0-4], [5-9], [10-14], [15-19], and [20-24].[5] We now treat this problem as multi-class classification, where a class is predicted given the interview transcripts. Note that this task is highly correlated to DLR.

Emotion Intensity Regression (EIR). In the CMU-MOSEI dataset, the emotion intensity is labeled at sentence-level, in contrast to transcript-level for depression estimation. Each sentence has a 6-D vector label, that contains scores in the range [0-3], for the six Ekman's emotions.[6] Note that in this dataset, many of the transcripts do not have labels for all its sentences. For such transcripts, we append a 0 on top of the labels of all the labeled sentences (representing that some emotion is exhibited in the utterance), and we manually label all the unlabeled sentences with [1,0,0,0,0,0,0], where 1 denotes that no emotion is exhibited in the utterance. As such, each sentence is labeled by a 7-D vector. If there are T sentences in a transcript, its label matrix is of $7\times T$ dimension. So, given the monologue transcript, we must predict this $7\times T$ matrix.

3.2 Sentence Preprocessing and Encoding

The initial step of our methodology aims to preprocess and encode each sentence of the respective transcripts.

Sentence Preprocessing. In DAIC-WOZ, many participants speak colloquially. So, we formalize all utterances by replacing contractions with corresponding full words. The sentences may also contain filler words such as "umm" or "hmm". We let them remain unchanged, as they may be important features to estimate depression. No preprocessing was required for the sentences in CMU-MOSEI, as they are already clear and formal.

Sentence Encoding Network. Inspired by the success of the universal sentence encoder [30] in finding semantic similarity between two sentences, we use its transformer variant to encode the sentences of the transcripts. It encodes a sentence using the encoding sub-graph of the transformer architecture. This sub-graph uses attention to compute context-aware representations of words in a sentence that take into account both the ordering and the identity of all the other words. The context-aware word representations are converted to a fixed-length sentence encoding vector by computing the element-wise sum of the representations at each word position. The encoder takes as input a lower-cased Penn Treebank tokenized string and outputs a 512 dimensional vector as the sentence embedding. As there are different numbers of sentences in different transcripts, we left-pad all the transcripts with 512-D zero-vectors, to a common length.

3.3 Learning Architectures

We describe three different multi-task models with respect to three tasks DLR, DLC and EIR. The three multi-task models are the Fully-Shared (FS MT.), the Shared-Private (SP MT.), and the Adversarial Shared-Private (ASP MT.). Each of the multi-task architectures has been designed for a combination of DLR, DLC and EIR, which are DLR-DLC, DLC-EIR, DLR-EIR and DLR-DLC-EIR. Note that the

5. More details about this process are given in section 4.

6. Further details of the dataset are given in section 4.

inputs of each model are the encoded sentences. We also implement a series of single-task models (ST.) for each of DLR, DLC and EIR.

Single-Task (ST.). The single-task model consists of a one-to-one Long Short Term Memory (LSTM) network [31], that encodes the transcript. The LSTM unit was chosen as the recurrent unit because it is very efficient in modeling long dependencies in time series data. In particular, LSTM networks are a special kind of recurrent neural network capable of learning long-term dependencies. As stated by [32], LSTM networks are the state-of-the-art structures for NLP tasks, as they have the ability to retain data through many time steps, a feature which no other deep neural networks have.

The output from the LSTM network (which may be the individual outputs of all sentences[7] or the sum of the outputs from all the LSTM units[8]) is fed to a set of fully connected and dropout layers. The output representation is then passed on to one or more (depending upon the task) linear regression units, in case the task is regression, or to a softmax classifier, in case the task is classification.

Fully-Shared Multi-Task (FS MT.). The fully-shared multi-task model consists of one LSTM network, that acts as the shared space for all the tasks. The outputs (or their summation, for DLC and DLR) from this LSTM network are fed to a task-specific network of fully connected and dropout layers. The representation obtained from this network is passed on to the output layer, which is a single linear regression unit for DLR, a 5-class softmax layer for DLC, or a layer of 7 regression units, in case the task is EIR.

This architecture forces the LSTM network to learn both the shared and task-specific features, as shown on the right side of Figure 1. Indeed, it does not have any facility to separate both shared and private spaces. The main drawback of this architecture is that it is bound to fail for increasingly less-correlated pairs of tasks, as the LSTM network is likely to fail to capture the task-specific features of all the tasks, if they are not enough correlated. However, in case tasks are heavily correlated, this network is expected to perform well.

Shared-Private Multi-Task (SP MT.). The shared-private multi-task model consists of three LSTM networks - two task-specific and one shared. All of the networks have the same number of units. In particular, the input of a task is fed to the task-specific and the shared LSTM network. The outputs from the task-specific and the shared LSTM layers are fused using an attention fusion mechanism [33], to obtain a fusion vector. The attention fusion network is explained further in this section. This fusion vector is then fed to a network of fully-connected and dropout layers, whose output is fed to the task-specific output layer.

This architecture, presented in Figure 2, improves over the fully-shared multi-task architecture by providing an infrastructure that has separate spaces for task-specific and shared features. But this too may have drawbacks. The shared feature space could contain some unnecessary task-specific features, while some shared features could also be mixed with the private space, thus suffering from feature redundancy as shown on the right side of Figure 2.

Adversarial Shared-Private Multi-Task (ASP MT.). Inspired by the results obtained by [14], [28], we design a similar architecture with two modifications. The adversarial shared-private multi-task architecture consists of three LSTM networks, that is, two task-specific and one shared, all of which have the same number of units. The input of a task is fed to the task-specific and the shared LSTM networks. The outputs from the task-specific and the shared LSTM layers are then fused using the attention fusion mechanism, oppositely to [14], [28], who use concatenation.

The output from the shared LSTM layer is also fed to a network $\mathrm{N_D}$ of fully-connected dropout and softmax layers. This network outputs the task label (for example, if there are two tasks T_1 and T_1, the task label for T_1 is [1, 0], the task label for T_2 is [0, 1]). The shared LSTM layers and $\mathrm{N_D}$ act as an adversarial network, the shared LSTM layer acting as the generator and $\mathrm{N_D}$ acting as the discriminator.

Finally, a L_{diff} loss function acts as an orthogonality constraint between private and shared layers and differs from the one used in [14], [28]. It is defined in Equation 1, where $\|.\|_1$ is the L_1 norm, H and S are two matrices, whose rows are each unit output of the task-specific LSTM network and the shared LSTM network, respectively, and m and n are the first and second dimensions of $H^\top S$ respectively. This definition of L_{diff} was empirically settled after testing other definitions. The architecture is shown in Figure 3.

$$L_{\text{diff}} = \frac{\|H^\top S\|_1}{m \times n}. \tag{1}$$

This architecture ensures that the task-specific and shared spaces are as separate as possible, as shown on the right side of Figure 3. The introduction of the adversarial network (shared LSTM - $\mathrm{N_D}$ pair) removes the possibility of task-specific features creeping into the shared-LSTM space. The orthogonality constraint ensures that the task-specific and shared spaces are as orthogonal as possible, which means the task-specific LSTM space should not contain any of the shared features as its space should be orthogonal to the shared LSTM space. Note that when the tasks are highly correlated, this architecture tends to perform poorly, as it would be very tough for the shared LSTM (generator) to create such a representation that can fool $\mathrm{N_D}$ (discriminator).

Attention Fusion Network (AFN.). In attention fusion, we first concatenate the outputs from the task-specific and the shared layers, pass them to a network of fully-connected and dropout layers, the output of which is passed to a softmax layer. This softmax layer outputs two values: α_{task} and α_{shared}, which weight the task-specific LSTM network and the shared LSTM network, respectively, in calculating the final output. So, α_{task} is multiplied with the output of the task-specific LSTM network, α_{shared} is multiplied with the output of the shared-LSTM network, and the corresponding products are summed. This summation represents the fusion vector. The attention fusion network is shown in 4.

7. This is the case for EIR as labels are given at sentence level.
8. This is the case for DLR and DLC as labels are given at text level.

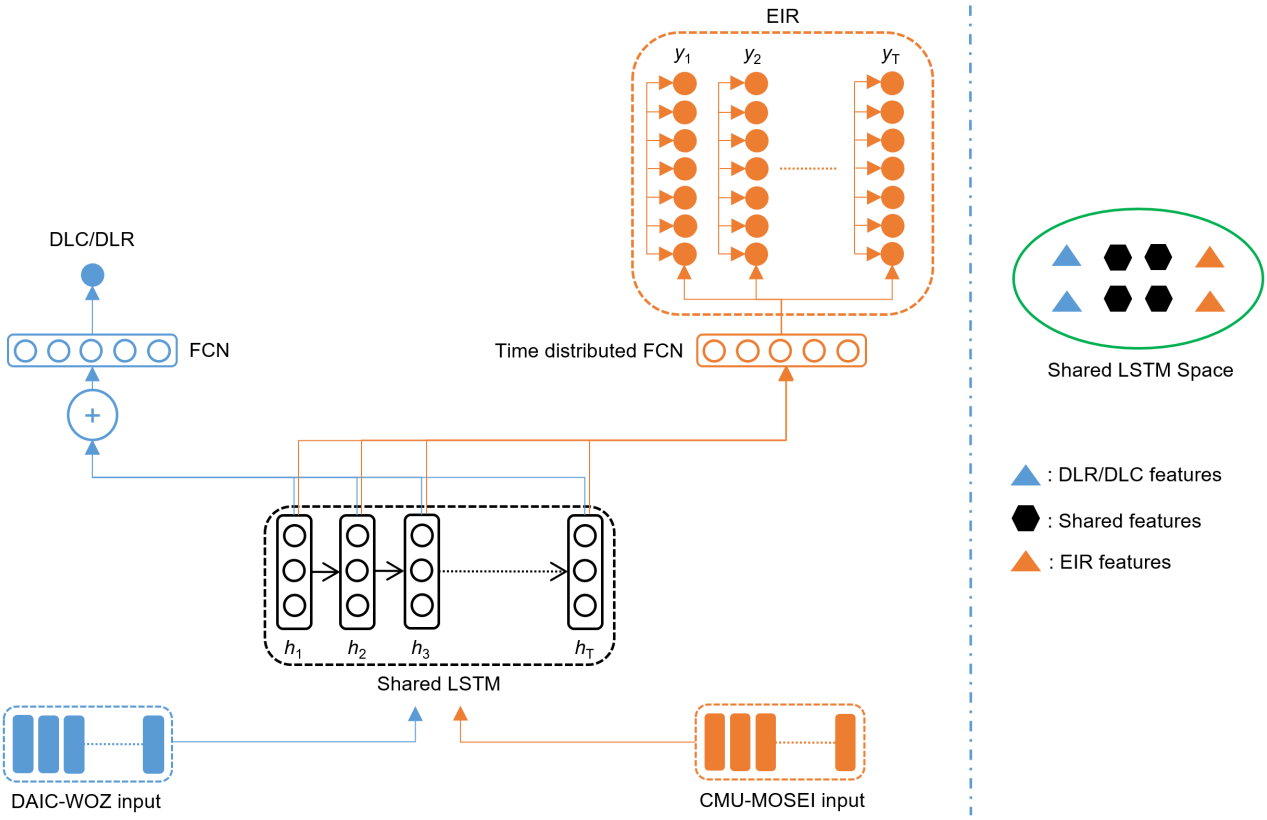

Fig. 1. Fully-Shared Multi-task model (FS MT.). FCN stands for Fully Connected Network, EIR for Emotion Intensity Regression and DLC (resp. DLR) for Depression Level Classification (resp. Depression Level Regression).

We particularly included the attention mechanism to better understand the behavior of each of the task-specific and shared features in the decision process. For estimating depression, if task-specific embeddings are more important than the shared embeddings, then α_{task} would have a value greater than 0.5, and α_{shared} would be less than 0.5. This would allow the network to learn the importance of the shared and task-specific embeddings by itself, in order to estimate depression/emotion levels. Moreover, networks with an attention mechanism usually perform better than their counterpart without attention [34].

Three-task Architectures. The definition of multi-task architectures that contain more than two tasks (here DLC+DLR+EIR) may not be straightforward in all cases. In the case of the fully-shared model, the definition is simple. Each task is solved using the single shared LSTM layer. With respect to the shared-private and the adversarial shared-private models, different strategies are possible. In our case, we take advantage of previous findings, namely that highly related tasks should perform better when fully-shared architectures are used. So, as DLC and DLR are highly correlated, we choose to combine them using a fully-shared architecture and combine the pair of tasks with EIR using the other two possible architectures (SP MT. and ASP MT.). The architecture for ASP MT. on three tasks is shown in Figure 5, and the SP MT. architecture can easily be inferred from the same illustration, by removing the discriminator and the orthogonality constraints.

4 Datasets and Learning Setups

In this section, we present two benchmark datasets, namely DAIC-WOZ for depression estimation and CMU-MOSEI for emotion intensity detection, as well as we define the learning setups of our different architectures.

4.1 DAIC-WOZ Dataset

The DAIC-WOZ depression dataset[9], that is used in the current study, is a subset of the DAIC corpus [16] containing clinical interviews of situations of psychological distress, which was generated by scientists from University of Southern California. These interviews were taken by a computer agent controlled by a human (wizard-of-oz virtual interviewer) who interacted with common people asking about their mental states and identified different verbal and non-verbal indicators for the same. The audio and video recordings and extensive questionnaire responses from the interviews are a part of the dataset. The data is annotated with a variety of verbal and non-verbal features.

189 sessions of dialogues are in the dataset, out of which, 45 are affiliated with the official test split, whose labels are not given. Out of the remaining 144, 6 of them were rejected as they had partial recording and interruptions, prompting to a final number of 138 samples. The accompanying features are (1) a raw audio document of the dialogue session combined with its transcript, (2) files gathering coordinates of 68 facial indicators, the histogram of oriented gradients

9. http://dcapswoz.ict.usc.edu/.

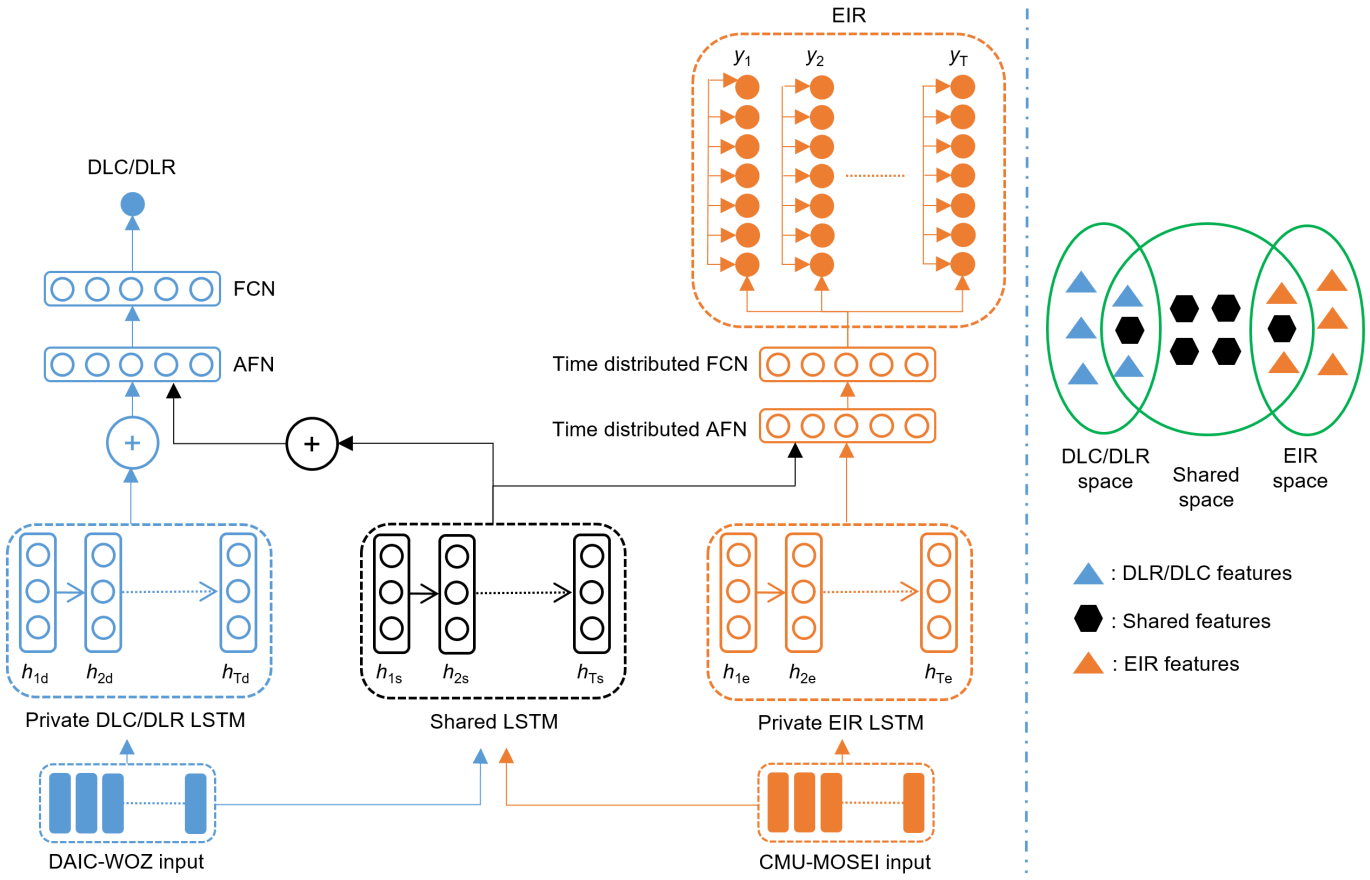

Fig. 2. Shared-Private Multi-task model (SP MT.). AFN refers to Attention Fusion Network, FCN to Fully Connected Network, EIR to Emotion Intensity Regression and DLC (resp. DLR) to Depression Level Classification (resp. Depression Level Regression).

(HoG) characteristics of the face, head pose and gaze directionality characteristics (extracted with OpenFace [35]), (3) a document containing the continuous facial activity units extracted with CERT [36], and (4) files with the COVAREP and formant voice characteristics computed with the COVAREP sotware [37].

As we are only focusing on the text modality, we only retain the transcript files that contain the sentences spoken by the virtual interviewer and the participant. The class-wise distribution of our training, development and test splits is summarized in Table 1. Note that medical studies [38] state that a PHQ-9 score in the interval [0-4] stands for None-minimal depression, in [5-9] for Mild, in [10-14] for Moderate, in [15-19] for Moderately severe, and in [20-27] for Severe depression. In the particular case of the PHQ-8 score, one question about suicidal condition is missing. As a consequence, the exact same discretization can be used, where severe depression is in the range of [20-24].

TABLE 1
Distribution of the DAIC-WOZ dataset by depression class.

Class	Train + Dev.	Test
None-minimal - [0-4] PHQ-8 score	47	16
Mild - [5-9] PHQ-8 score	28	5
Moderate - [10-14] PHQ-8 score	19	5
Moderately severe - [15-20] PHQ-8 score	7	6
Severe - [20-24] PHQ-8 score	4	1

The DAIC-WOZ dataset, however, has some limitations. The number of samples in the entire dataset is small and not evenly distributed, with just one sample of the "severely depressed" category in the test set. It is clear that further efforts are needed to increase such a dataset, although this remains out of the scope of this paper. In all cases, all obtained results of our study will have to be put in perspective relatively to this small amount of learning instances.

4.2 CMU-MOSEI Dataset

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset[10] comprises 3,228 videos from 1,000 different speakers over 250 topics [17]. The videos were gathered from an online video platform, where users emit their opinions in the form of monologues. Each video contains a unique person, who discusses in front of the camera about a given topic. Each video can be transformed into three information sources: language (spoken utterances), visual (gesture analysis), and acoustics (intonations and prosody). During data acquisition, videos were analyzed by automatic face detection to verify whether a unique speaker is present. Moreover, only the videos where the speaker's attention is exclusively towards the camera were kept. The number of videos acquired from each channel was restricted to 10 to avoid bias and all videos must have correct transcriptions provided by the speaker. The

10. https://github.com/A2Zadeh/CMU-MultimodalSDK.

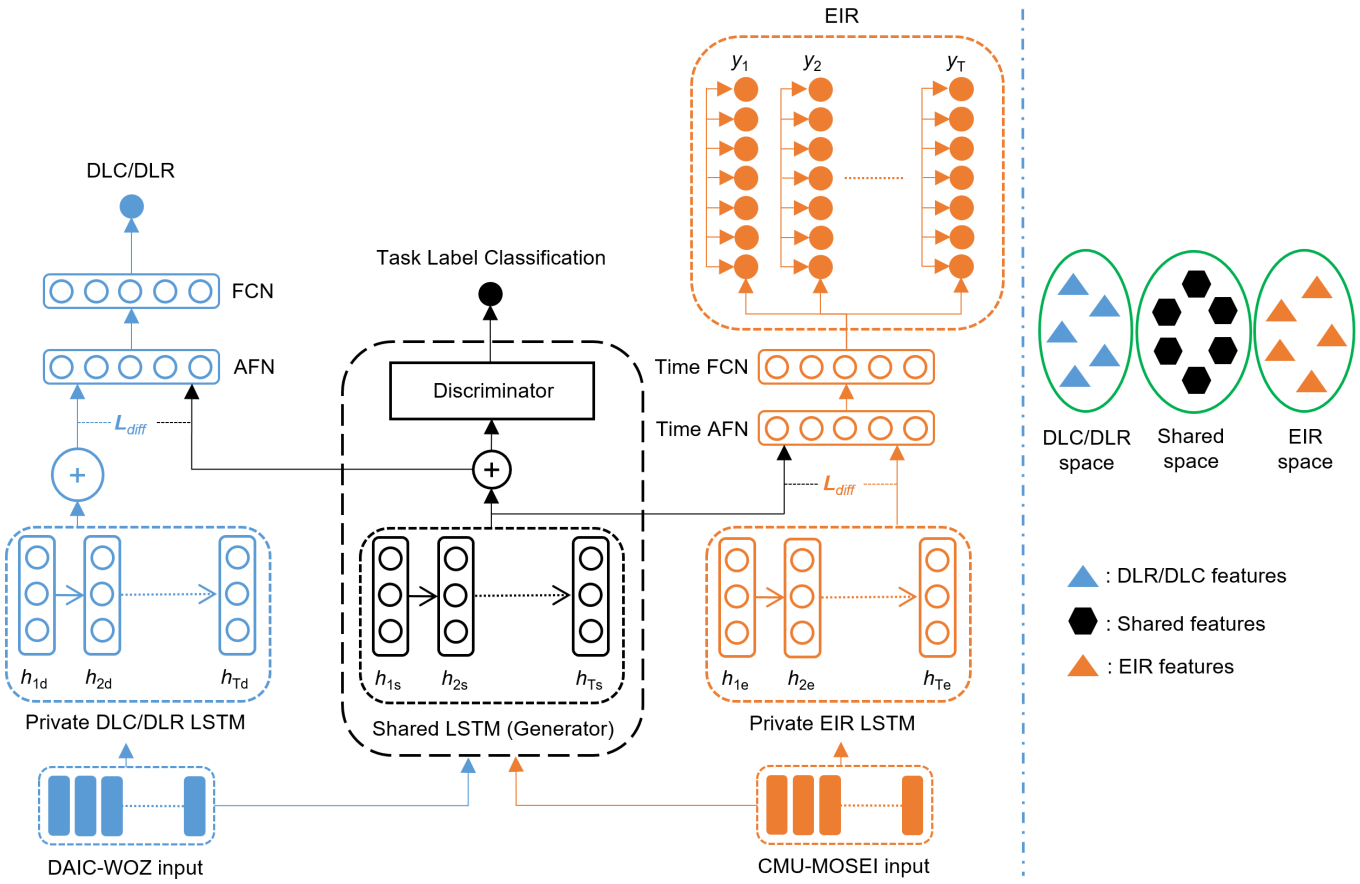

Fig. 3. Adversarial Shared-Private Multi-task model (ASP MT.). AFN refers to Attention Fusion Network, FCN to Fully Connected Network, EIR to Emotion Intensity Regression and DLC (resp. DLR) to Depression Level Classification (resp. Depression Level Regression).

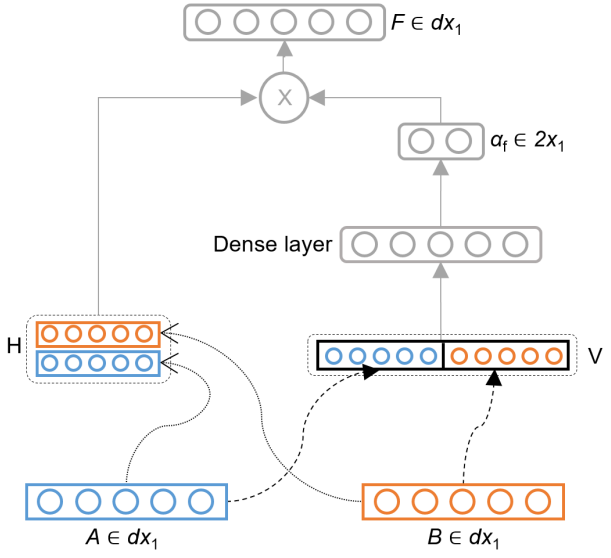

Fig. 4. Attention Fusion Network (AFN).

quality inspection has been made by 14 expert judges, and 3,228 videos were selected from the 5,000 initially gathered.

The 3,228 videos were then segmented into 23,453 annotated pieces, where each segment contains a manual transcription aligned with audio to phoneme level. The annotation of CMU-MOSEI closely follows the annotation rules of the CMU-MOSI [39] dataset. In particular, sentences were annotated for Ekman's six emotions, that is happiness, sadness, anger, fear, disgust and surprise, on a [0,3] Likert scale for the presence of emotion. As such, 0 stands for no evidence of x, 1 for weakly x, 2 for x, and 3 for highly x. With respect to sentiment evaluation, a [-3,3] Likert scale was used such that: -3 is highly negative, -2 is negative, -1 is weakly negative, 0 is neutral, 1 is weakly positive, 2 is positive, and 3 is highly positive. Note that in this paper, we do not use the annotation for sentiment evidence. As stated in [17], the annotation was carried out by 3 crowd-sourced judges from Amazon Mechanical Turk platform, where judges were provided with a 5 minutes training video on how to use the annotation system in order to avoid extreme annotation, and all judges were master workers with an approval rate higher than 98%.

Note that as in CMU-MOSEI each of the 3,228 video transcripts contains an average of 7.3 utterances, and in DAIC-WOZ, the 138 interview transcripts contain an average of 90 utterances, we randomly selected 517 transcripts from CMU-MOSEI to reduce imbalance between datasets.

4.3 Learning Setups

With respect to multi-task learning, the task-specific LSTM layers are trained alternatively using the entire training split. As an example, consider the training of the shared-private multi-task network for depression level regression and emotion intensity regression: SP MT. DLR+EIR.

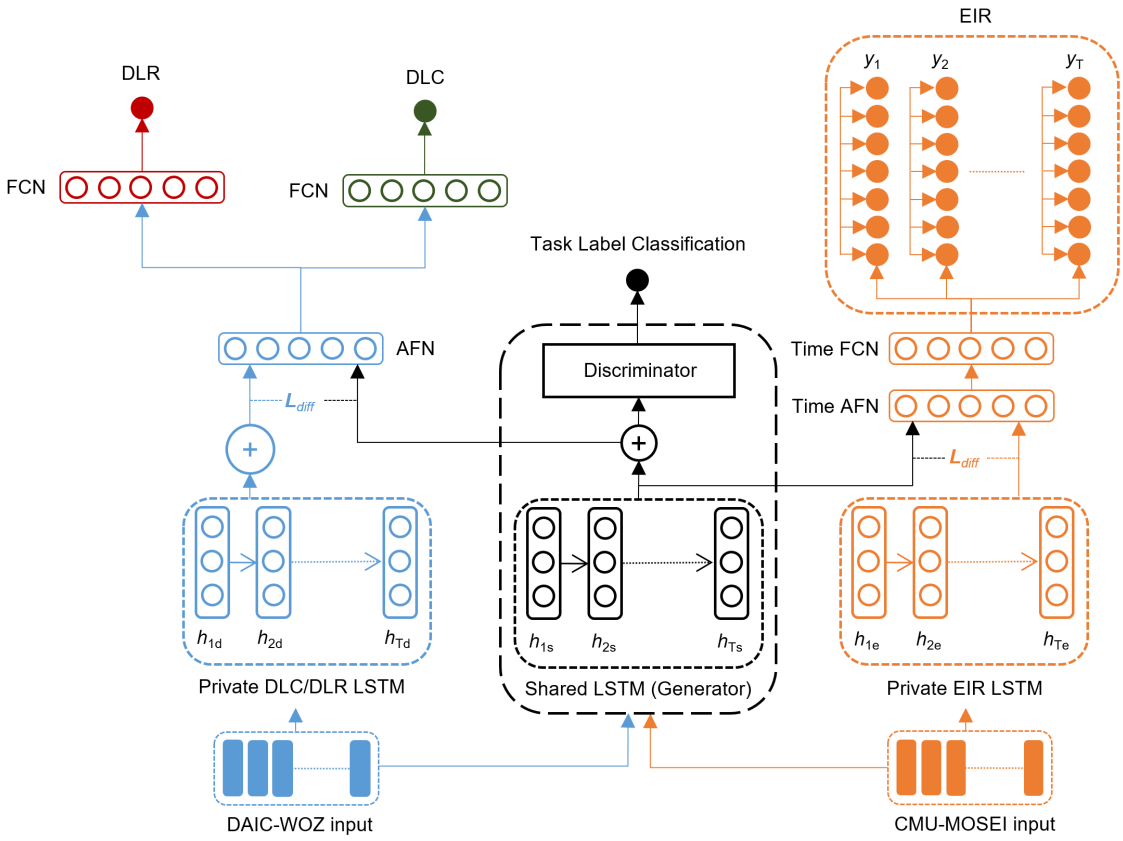

Fig. 5. Adversarial Shared-Private Multi-task model (ASP MT.) for three tasks. AFN refers to Attention Fusion Network, FCN to Fully Connected Network, EIR to Emotion Intensity Regression and DLC (resp. DLR) to Depression Level Classification (resp. Depression Level Regression).

The DLR-specific LSTM layer, the shared-LSTM layer, and the corresponding attention fusion network and fully-connected network are trained for N_{DLR} epochs without updating the weights of the EIR-specific layers. For the next N_{EIR} epochs, the EIR-specific LSTM layer, the shared-LSTM layer, and the corresponding attention fusion network and fully-connected network are trained without updating the weights of the DLR-specific layers. Here, N_{DLR} and N_{EIR} are treated as hyperparameters. We go on training the network in this alternating fashion till a maximum number of iterations N_{total} (the total number of times the shared-LSTM layer is trained) is reached. The model that shows best performance on the development split over all iterations is chosen for testing. The pseudo-code for our training procedure is shown in algorithm 1.

Algorithm 1 FS/SP/ASP MT. T_1+T_2 training

$n_{total} \leftarrow 1$
while $n_{total} < N_{total}$ **do**
  for $n_{T_1} \leftarrow 1$ to N_{T_1} **do**
    Update T_1-specific and Shared weights
    $n_{total} \leftarrow n_{total} + 1$
  for $n_{T_2} \leftarrow 1$ to N_{T_2} **do**
    Update T_2-specific and Shared weights
    $n_{total} \leftarrow n_{total} + 1$

The architectures have been implemented with Keras[11] and hyperparameters have been optimized through grid search. Note that all learning models are trained on the basis of stratified 5-cross validation, thus keeping the data distribution between training, development and test datasets. In particular, the best of the 5 models over the development set is applied to classify/regress the examples in the test set.

5 Results

In order to test our hypothesis, we perform a series of experiments for three different tasks: (1) Depression Level Regression (DLR), which aims to assign a value between 0 to 24 (that is, the PHQ-8 score) to a given patient interview transcript, (2) Depression Level Classification (DLC), whose objective is to identify the correct discrete class of depression level (None-minimal, Mild, Moderate, Moderately severe, Severe), and (3) Emotion Intensity Regression (EIR), which regresses a [0-3] value for each of the six Ekman emotions (happiness, sadness, anger, fear, disgust and surprise) for a given user transcript.

Five different models serve as baselines. That is, each task is first modeled as a single-task problem, and two unaware-emotion multi-task (fully-shared and shared-private) architectures are implemented that combine both DLR and DLC.[12] Three different combinations of emotion-aware multi-task frameworks are tested, for each one of the three theoretical models (fully-shared, shared-private and adversarial shared-private): (1) DLC combined with EIR, (2)

11. https://keras.io.

12. These baselines correspond to the 5 first rows of Table 2.

TABLE 2
Overall classification results including single-task (ST.) models as well as Fully-Shared Multi-Task (FS MT.), Shared-Private Multi-Task (SP MT.) and Adversarial Shared-Private Multi-Task (ASP MT.) models. Acc., Ov. and Un. metrics are given in % and respectively correspond to Accuracy, Over and Under. F1 stands for F1 score, MCC for Matthews Correlation Coefficient. RMSE refers to Root Mean Squared Error, MAE to Mean Average Error, R^2 to Coefficient of Determination, SM. to the Symmetric Mean Absolute Percentage Error and $\overline{Ov.}$ and $\overline{Un.}$ to over-evaluation and under-evaluation metrics for regression. $\overline{MSE}$ stands for average Mean Squared Error.

Models	**Evaluation Metrics**													
	DLC							DLR						EIR
	Acc.	F1	MCC	RMSE	MAE	Ov.	Un.	RMSE	MAE	R^2	SM.	$\overline{Ov.}$	$\overline{Un.}$	$\overline{MSE}$
Baselines without Emotion Intensity Regression														
ST. DLC	60.61	0.54	0.38	1.31	0.75	3.03	36.36	-	-	-	-	-	-	-
ST. DLR	-	-	-	-	-	-	-	4.90	3.99	0.46	0.97	3.21	5.18	-
ST. EIR	-	-	-	-	-	-	-	-	-	-	-	-	-	7.15
FS MT. DLC+DLR	**66.66**	**0.62**	**0.49**	1.23	0.66	3.03	30.31	4.96	3.89	0.44	0.98	2.81	5.19	-
SP MT. DLC+DLR	60.61	0.51	0.39	1.26	0.72	0.00	39.39	4.70	3.81	0.50	0.99	3.39	4.32	-
Multi-task Results with Emotion Intensity Regression														
FS MT. DLC+EIR	60.61	0.51	0.42	1.58	0.90	0.00	39.39	-	-	-		-	-	6.98
SP MT. DLC+EIR	57.57	0.50	0.35	1.27	0.76	6.07	36.36	-	-	-	-	-	-	7.05
ASP MT. DLC+EIR	60.61	0.54	0.38	1.26	0.73	9.09	30.30	-	-	-	-	-	-	7.19
FS MT. DLR+EIR	-	-	-	-	-	-	-	4.60	3.74	0.52	0.99	3.16	4.63	6.88
SP MT. DLR+EIR	-	-	-	-	-	-	-	**4.51**	3.89	**0.54**	**0.94**	3.91	3.85	**6.82**
ASP MT. DLR+EIR	-	-	-	-	-	-	-	4.72	3.96	0.50	**0.94**	3.80	4.15	7.08
FS MT. DLC+DLR+EIR	57.57	0.46	0.38	1.36	0.82	3.04	39.39	4.83	4.03	0.47	0.97	3.13	5.11	6.96
SP MT. DLC+DLR+EIR	**63.64**	0.58	**0.48**	**0.94**	**0.51**	24.24	12.12	4.56	3.79	0.53	0.97	3.20	4.59	7.02
ASP MT. DLC+DLR+EIR	60.61	**0.60**	0.42	1.14	0.64	12.12	27.27	4.61	**3.69**	0.52	0.95	2.87	4.81	7.11

DLR combined with EIR, and (3) DLC combined with both DLR and EIR.[13]

To evaluate regression/classification results, we use well-known evaluation metrics that are standard for depression level estimation [40]: (1) Accuracy, F1 score and Matthews Correlation Coefficient (MCC) for classification; (2) Root Mean Square Error (RMSE), Mean Average Error (MAE), Coefficient of Determination (R^2) and Symmetric Mean Absolute Percentage Error (SMAPE) for regression. In particular, we include two other metrics (Over and Under), that complement Accuracy and evaluate how much a learning model over-evaluates (Over) or under-evaluates (Under) the correct result. Such metrics are important to understand the behavior of learning models. But, as far as we know, they are not presented in related works. For classification, Accuracy, Over and Under sum to 100% and are defined in equations 2 and 3. For regression, $\overline{Over}$ and $\overline{Under}$ metrics quantify average continuous over-evaluation and under-evaluation and are defined in equations 4 and 5.

$$Over = \frac{\sum_{y_i<\hat{y_i}} 1}{\sum_{y_i} 1} \tag{2}$$

$$Under = \frac{\sum_{y_i>\hat{y_i}} 1}{\sum_{y_i} 1} \tag{3}$$

$$\overline{Over} = \frac{\sum_{y_i<\hat{y_i}} \hat{y_i} - y_i}{\sum_{y_i<\hat{y_i}} 1} \tag{4}$$

$$\overline{Under} = \frac{\sum_{y_i>\hat{y_i}} y_i - \hat{y_i}}{\sum_{y_i>\hat{y_i}} 1}. \tag{5}$$

Finally, for emotion intensity regression, we present a global metric $\overline{MSE}$ that averages the squared errors over the indicator of the presence of emotion, and all six emotions. It is defined in Equation 6. As our main focus is on depression, we do not compute emotion-wise metrics, and $\overline{MSE}$ acts as a global indicator.

$$\overline{MSE} = \frac{1}{|y|} \sum_{y} \sum_{i=1}^{i=7} (y_i - \hat{y}_i)^2. \tag{6}$$

Overall evaluation results are given in Table 2. Note that we provide all confusion matrices as supplementary online material[14] to show the overall sketch for DLC.

5.1 Results by Task

DLC can be seen as a coarse-grain task compared to DLR. In this paper, we study both tasks contrarily to previous related works, which only focus on the fine-grained task.

With respect to **DLC**, the best results in terms of Accuracy are obtained by the emotion-unaware multi-task baseline that combines both DLC and DLR, outdoing the best emotion-aware model by 3.03%. However, best results in terms of RMSE and MAE are evidenced by the emotion-aware shared-private multi-task model that concurrently learns all tasks DLC, DLR and EIR. In this case, improvements respectively reach 23.57% for RMSE and 22.7% for MAE. So, although the baseline tends to produce more accurate results, incorrect guesses largely deviate from the correct answer. Moreover, baseline decisions tend to under-evaluate the degree of depression. Indeed, for the best baseline model, 90.9% (Under=30.31%) of the incorrect guesses are under-evaluated, compared to only 9.1% (Over=3.03%), which are over-evaluated. In comparison, the emotion-aware model tends to over-evaluate depression levels in 66.6% (Over=24.24%) of the incorrect cases, and under-evaluates them in 33.3% (Under=12.12%), showing a more balanced behavior. In terms of medical decisions, this phenomenon can be an important issue, as under-evaluating the degree of depression of a given patient may

13. These models correspond to the 9 last rows of Table 2.

14. http://dias.users.greyc.fr/cm.pdf.

have worst consequences than over-evaluating it, although none of these cases should be encountered.[15]

With respect to **DLR**, the best results overall are obtained for the emotion-aware models. In this case, a minimum RMSE=4.51 is obtained by the shared-private model that combines DLR and EIR, and a minimum MAE=3.69 is achieved by the adversarial shared-private model that combines DLR, DLC and EIR. Note that the best evidenced model for DLC (that is, shared-private multi-task model combining DLC, DLR and EIR) shows very similar results with RMSE=4.56 and MAE=3.79 for DLR. As a consequence, an improvement of 4.04% in terms of RMSE and 3.14% in terms of MAE can be achieved over the best baseline, embodied by the shared-private multi-task model that combines DLC and DLR. Interestingly, the emotion-aware models tend to show that in case of over-evaluation, the exceeding values are smaller for the baselines, a situation that also occurs for under-evaluation, although values of under-evaluation are larger than figures evidenced by over-evaluation. As a consequence, there is a tendency of under-evaluation of all models, which may be a drawback in terms of medical issue as mentioned above.

With respect to **EIR**, best results are unexpectedly obtained for depression-aware models, suggesting that emotion intensity regression may also benefit from depression level regression/classification. In particular, the best improvement is evidenced by the shared-private two-task model, which learns DLR and EIR concurrently, with $\overline{MSE}$=6.82, closely followed by the fully-shared model that combines DLR and EIR with $\overline{MSE}$=6.88, evidencing the second best result. As such, an improvement of 4.6% can be obtained compared to the baseline.[16]

The first results show that emotion-aware models can improve the performance of the depression level estimation. In particular, the shared-private multi-task model combining DLC, DLR and EIR seems the more regular architecture to improve over all three tasks on average, as it is highly ranked for all tasks individually across all evaluation metrics. Nevertheless, in order to better understand these results, we propose a class-wise analysis.

5.2 Results by Class

The overall idea of the class-wise analysis is to verify whether some classes of depression are better handled by the classifiers than others. Note that as far as we know, previous related works do not incorporate such an analysis and rely exclusively on overall results, thus failing to take into account important medical issues. The overall results by class are given in Table 3. Note that we do not show all evaluation metrics as it has been evidenced in Table 2 that they are all highly correlated.

For that purpose, we present the exact same results of Table 2 down-described by the 5 classes of depression, which are, none-minimal, mild, moderate, moderately severe and severe. Although this information is interesting, it must be carefully interpreted as the number of test examples is small and not equally distributed. For example, there is only one test example for severe depression, and the DAIC-WOZ dataset contains only four such cases. Overall results are presented in Table 3.

Within this context, overall results show high inequalities between class. The **none-minimal class** seems to be well-handled with high accuracy values and respectively low RMSE and MAE on average for both DLC and DLR for all models, including baselines. Note that the best three-task multi-task model evidences the lowest RMSE and MAE values for this particular class, although it fails to correctly classify all examples. Moreover, there is a clear tendency for over-evaluation, which is understandable as many examples have a PHQ-8 score equals to 0. These observations are clearly positive indicators that strong classification/regression results can be obtained for the class with more patients involved both at training and test splits.

On the other side, the **severe class** shows worst class-wise results as none of the models is capable of correctly classifying the single example present in the test set. Moreover, almost all models fail to correctly estimate this example by a large margin: two classes difference for DLC, and large RMSE and MAE values for DLR, although less expressive values are obtained for emotion-aware architectures. Of course, these concluding remarks can not be generalized due to the lack of statistical evidence over more examples.

As for the **moderately severe class**, all models perform like-wise in terms of DLC accuracy. However, the emotion-aware models evidence lower RMSE and MAE values than baseline models, thus showing more accurate classification estimations. However, in terms of DLR, huge average under-evaluation values are shown by all models, thus showing the difficulty to handle this class in terms of regression. Note that this class is the one that evidences worst results overall in terms of RMSE and MAE for DLR over all models. In fact, some patients within this class can easily be classified, but others are rather difficult to estimate in terms of depression level, and odd low values are usually given by the learning model to these cases.

The **mild class** receives best accuracy levels for the baseline model, and the best three-task model clearly fails within this class, showing worst results overall in terms of DLC. In this case, emotion-aware models do not benefit from the introduction of the concurrent learning of emotion intensity. Moreover, almost no improvement is obtained in terms of DLR by emotion-aware models, to the exception of the shared-private multi-task models combining DLR and EIR, with minor improved results. This class is certainly the one where our initial hypothesis does not clearly stand.

Finally, the **moderate class** receives highest classification results with the three-task model by a large margin. In this case, it clearly outperforms all emotion-aware and emotion-unaware models, for accuracy, RMSE and MAE. With respect to DLR, the best performing model is still an emotion-aware model, but the two-task model. In this case, it clearly outperforms all other tested models. Note that in all cases, there is a clear tendency for under-evaluation as no estimator over-evaluates any patient's level of depression.

Although, as explained before, no strict concluding remarks can be drawn from this analysis due to the small number of test examples, this class-wise analysis should systematically be included in related works of depression

15. We will see in this section that most DLR models under-evaluate estimations.

16. Stronger analysis is out of the scope of this paper.

TABLE 3
Detailed classification/regression results by depression level class: 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). Results for the best performing architecture only are given. Acc., Ov. and Un. metrics are given in % and respectively correspond to Accuracy, Over and Under. RMSE refers to Root Mean Squared Error, MAE to Mean Average Error, and $\overline{Ov.}$ and $\overline{Un.}$ to over-evaluation and under-evaluation for regression.

Models	**Evaluation Metrics**								
	DLC					DLR			
	Acc.	RMSE	MAE	Ov.	Un.	RMSE	MAE	$\overline{Ov.}$	$\overline{Un.}$
	Best for DLC without EIR: FS MT. DLC+DLR					**Best for DLR without EIR: SP MT. DLC+DLR**			
None-minimal	100	0.00	0.00	0	-	3.97	3.22	3.51	1.14
Mild	40	1.10	0.80	20.00	40	3.80	3.11	3.82	2.05
Moderate	40	1.34	1.00	0.00	60	4.04	3.50	0.00	3.50
Moderately severe	33.33	2.27	1.83	0.00	66.67	6.78	5.75	0.47	6.81
Severe	0	2.00	2.00	-	100	6.81	6.81	0.00	6.81
	Best for DLC+EIR: ASP MT. DLC+EIR					**Best for DLR+EIR: SP MT. DLR+EIR**			
None-minimal	100	0.00	0.00	0	-	4.28	3.85	4.05	0.74
Mild	20	1.18	1.00	40	40	3.51	3.07	3.56	2.32
Moderate	20	1.61	1.40	20	60	2.94	2.60	0.00	2.60
Moderately severe	33.33	2.16	1.67	0	66.67	6.70	6.05	2.77	6.71
Severe	0	2.00	2.00	-	100	2.03	2.03	0.00	2.03
	Best for DLC+DLR+EIR: SP MT. DLC+DLR+EIR								
None-minimal	93.75	0.50	0.13	6.25	-	3.42	2.89	2.97	1.79
Mild	0	1.00	1.00	60	40	3.78	3.49	3.89	2.88
Moderate	80	0.89	0.40	0	20	3.84	3.37	0.00	3.37
Moderately severe	33.33	1.41	1.00	0	66.67	7.54	6.78	4.67	7.21
Severe	0	2.00	2.00	-	100	3.85	3.85	0.00	3.85

level estimation. Indeed, it seems that some classes are more difficult to handle than others, and also models do not perform equally over all classes, although there is a tendency, which confirms the initial hypothesis that emotion intensity estimation can be beneficial to depression level classification/regression.

5.3 Results by Learning Models

Finally, we analyze the behavior of each multi-task model in terms of the fully-shared, shared-private and adversarial shared-private architectures. In particular, improved results were expected by the adversarial shared-private models following initial results reported in [14], [28]. However, this architecture never reaches the highest results, with the exception of the two-task model that includes DLC and EIR, even though it is with a tiny margin over the shared-private architecture. In fact, the adversarial shared-private framework relies on a generator, which learns a shared representation that is capable of fooling the discriminator in terms of task label. This architecture can indeed be beneficial when the concurrent tasks are closely related and in particular when they share some ambiguous features. However, this is not really the case in our experiments as the length of the transcripts is unequal for depression and emotion levels, as well as the vocabulary may not highly overlap. As a consequence, finding a shared representation that can discriminate between both tasks is not a difficult problem, and the learned representation may not be informative enough to handle the concurrent tasks individually. So, the shared-private models regularly evidence stronger results both for DLC and DLR, to the exception of the baseline model, which combines both DLC and DLR. In this case, the fully-shared model shows the best results. This can easily be understood, as (1) the same training dataset is used twice in the training step enforcing the generalization process in terms of shared representation and (2) DLC can be seen as a subtask of DLR, thus including a strong regularization process within the model. Note that this finding is at the origin of the proposed shared-private three-task architecture, that includes a fully-shared layer between DLC and DLR, and globally evidences more stable results overall.

6 Conclusion

In this paper, we tested the hypothesis that depression level classification/regression can leverage from the concurrent learning of emotion intensity. For that purpose, we implemented a series of emotion-aware and emotion-unaware multi-task architectures over combinations of three tasks: depression level classification, depression level regression and emotion intensity regression. Strong evaluation including new metrics and class-wise results shows that emotion-aware models outperform emotion-unaware baselines in a vast majority of tested situations over the standard benchmarks DAIC-WOZ and CMU-MOSEI. We anticipate that our work will help to reduce the number of cases of late treatment of depression, as one can always get an estimate of his/her PHQ-8 score, without needing to consult a psychiatrist, especially considering the stigma surrounding this illness. However, current results are not accurate enough to help the therapist in his diagnosis as model performance is still too low. This should be a great motivation for future work in depression level estimation. Such research directions include (1) the combination of text, visual and acoustic modalities following the ideas of [10], [17], [33], (2) the study of different concurrent tasks for depression estimation and (3) the creation of larger datasets to better evaluate depression models in terms of class-wise results, that may also include new biomarkers or descriptors.

ACKNOWLEDGMENTS

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

REFERENCES

[1] T. Vos, R. M. Barber, B. Bell, and A. Bertozzi-Villa, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study," *The Lancet*, vol. 386, no. 9995, pp. 743-800, August 2015.

[2] A. Ferrari, F. Charlson, R. Norman, S. Patten, G. Freedman, C. Murray, T. Vos, and H. Whiteford, "Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study," *PLOS Medicine*, vol. 10, no. 11, pp. 1-12, November 2013.

[3] A. Beck and B. Alford, *Depression: Causes and Treatment*, University of Pennsylvania Press, 2009.

[4] K. Smith, P. Renshaw, and J. Bilelloa, "The diagnosis of depression: Current and emerging methods," *Comprehensive Psychiatry*, vol. 54, no. 1, pp. 1-6, 2013.

[5] K. Kroenke, T. Strine, R. Spitzer, J. Williams, J. T. Berry, and A. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, pp. 163-73, 2008.

[6] S. El-Den, T. Chen, Y-L. Gan, E. Wong, and C. O'Reilly, "The psychometric properties of depression screening tools in primary healthcare settings: A systematic review," *Journal of Affective Disorders*, vol. 225, pp. 503-522, 2018.

[7] C. Shin, S.-H. Lee, K.-M. Han, H.-K. Yoon, and C. Han, "Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: Analysis of psychiatric outpatient Data," *Psychiatry Investigation*, vol. 16, no. 4, pp. 300-305, 2019.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10-49, 2015.

[9] J. T. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP," *Proc. of International Workshop on Language Cognition and Computational Models*, pp. 11-21, Santa Fe, August 2018.

[10] 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.

[11] J. Joormann and I. Gotlib, "Emotion regulation in depression: relation to cognitive inhibition," *Cognition and Emotion*, vol. 24, no. 2, pp. 281-298, 2010.

[12] R. Thompson, M. Boden, and I. Gotlib, "Emotional variability and clarity in depression and social anxiety," *Cognition and Emotion*, vol. 31, no. 1, pp.98-108, 2017.

[13] P. Ekman and R. Davidson, *The Nature of Emotion: Fundamental Questions.* Oxford University Press, 1994.

[14] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 1-10, Vancouver, July 30-August 4, 2017.

[15] S-A. Qureshi, S. Saha, M. Hasanuzzaman, and G. Dias, "Multitask representation learning for multimodal estimation of depression level," *IEEE Intelligent Systems*, vol. 34, no. 5, pp. 1-8, 2019.

[16] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, and S. Marsella, "The distress analysis interview corpus of human and computer interviews," *Proc. of International Conference on Language Resources and Evaluation*, pp. 3123-3128, Reykjavik, May 26-31, 2014.

[17] A. Zadeh, P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 2236-2246, Melbourne, July 15-20, 2018.

[18] N. Dewan, J. Luo, and N. Lorenzi, *Mental Health Practice in a Digital World: A Clinicians Guide*. Springer Publishing Company Incorporated, 2015.

[19] M. Chatterjee, G. Stratou, S. Scherer, and L-P. Morency, "Context-based signal descriptors of heart-rate variability for anxiety assessment," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3631-3635, Florence, May 4-9 2014.

[20] M. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," *Proc. of IEEE Spoken Language Technology Workshop*, pp. 136-143, San Diego, May 13-16, 2016.

[21] D. Hovy, M. Mitchell, and A. Benton, "Multitask learning for mental health conditions with limited social media data," *Proc. of Conference of the European Chapter of the Association for Computational Linguistics*, pp. 152-162, Valencia, April 3-7, 2017.

[22] D. Losada and F. Crestani, "A test collection for research on depression and language use," *Proc. of International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 28-39, Évora, September 5-8, 2016.

[23] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. Cohn, "Multimodal detection of depression in clinical interviews," *Proc. of ACM International Conference on Multimodal Interaction*, pp. 307-310, Seattle, November 9-13, 2015.

[24] M. Morales, S. Scherer, and R. Levitan, "A linguistically-informed fusion approach for multimodal depression detection," *Proc. of Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 13-24, New Orleans, June, 2018.

[25] S-A. Qureshi, M. Hasansuzzaman, S. Saha, and G. Dias, "The verbal and non verbal signals of depression – combining acoustics, text and visuals for estimating depression level," *arXiv:1904.07656 [cs]*, April, 2019.

[26] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," *Proc. of Annual Workshop on Audio/Visual Emotion Challenge*, pp. 61-68, Mountain View, October 23-27, 2017.

[27] J. Bingel and A. Søgaard, "Identifying beneficial task relations for multi-task learning in deep neural networks," *Proc. of Conference of the European Chapter of the Association for Computational Linguistics*, pp. 164-169, Valencia, April 3-7, 2017.

[28] S. Yadav, A. Ekbal, S. Saha, P. Bhattacharyya, and A. Sheth, "Multi-task learning framework for mining crowd intelligence towards clinical treatment," *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 271-277, New Orleans, June 1-6, 2018.

[29] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41-75, 1998.

[30] D. Cer, Y. Yang, S-Y. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant, M. Guajardo-Cespedes, S. Yuan, and C. Tar, "Universal sentence encoder," *arXiv:1803.11175 [cs]*, Mars, 2018.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term Memory". *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[32] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *arXiv:1708.02709 [cs]*, August, 2017.

[33] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," *Proc. of IEEE International Conference on Data Mining*, pp. 1033-1038, New Orleans, November 18-21, 2017.

[34] T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 1412-1421, Lisbon, September 17–21, 2015.

[35] T. Baltrušaitis, P. Robinson, and L-P. Morency, "Openface: An open source facial behavior analysis toolkit," *Proc. of IEEE Winter Conference on Applications of Computer Vision*, pp. 1-10, Lake Placid, March 7-9, 2016.

[36] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," *Proc. of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 298-305, Santa Barbara, March 21-25, 2011.

[37] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 960-964, Florence, May 4-9, 2014.

[38] K. Kroenke, R. Spitzer, and J. Williams "The PHQ-9: Validity of

a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606-613, 2001.

[39] A. Zadeh, R. Zellers, E. Pincus, and L-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv:1606.06259 [cs]*, June, 2016.

[40] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," *Proc. of Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3-9, Mountain View, October 23-27, 2017.