Zhang, Yazhou; Li, Xiang; Rong, Lu; Tiwari, Prayag

Multi-Task Learning for Jointly Detecting Depression and Emotion

Published in:
Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021

DOI:
10.1109/BIBM52615.2021.9669546

Published: 01/01/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Zhang, Y., Li, X., Rong, L., & Tiwari, P. (2021). Multi-Task Learning for Jointly Detecting Depression and Emotion. In Y. Huang, L. Kurgan, F. Luo, X. T. Hu, Y. Chen, E. Dougherty, A. Kloczkowski, & Y. Li (Eds.), *Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021* (pp. 3142-3149). IEEE. https://doi.org/10.1109/BIBM52615.2021.9669546

Multi-Task Learning for Jointly Detecting Depression and Emotion

1st Yazhou Zhang*
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
yzzhang@zzuli.edu.cn

2nd Xiang Li*
Shandong Computer Science Center (National Supercomputing Center in Jinan)
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
xiangli@sdas.org

3rd Lu Rong†
Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou, China
lurong2013@outlook.com

4th Prayag Tiwari
Department of Computer Science
Aalto University
Helsinki, Finland
prayag.tiwari@aalto.fi

***Abstract*—Depression is a typical mood disease that makes people a persistent feeling of sadness and loss of interest and pleasure. Emotion thus comes into sight and is tightly entangled with depression in that one helps the understanding of the other. Depression and emotion detection has been a new research task. The central challenges in this task are multi-modal interaction and multi-task correlation. The existing approaches treat them as two separate tasks, and fail to model the relationships between them. In this paper, we propose an attentive multi-modal multi-task learning framework, called AMM, to generically address such issues. The core modules are two attention mechanisms, $viz.$ inter-modal (I_e) and inter-task (I_t) attentions. The main motivation of I_e attention is to learn multi-modal fused representation. In contrast, I_t attention is proposed to learn the relationship between depression detection and emotion recognition. Extensive experiments are conducted on two large scale datasets, i.e., DAIC and multi-modal Getty Image depression (MGID). The results show the effectiveness of the proposed AMM framework, and also shows that AMM obtains better performance for the main task, i.e., depression detection with the help of the secondary emotion recognition task.**

***Index Terms*—multi-modal depression detection, emotion recognition, multi-task learning, deep learning, artificial intelligence**

I. Introduction

Depression, also known depressive disorder, is a common but serious medical illness characterized by a persistent feeling of sadness and retardation of thinking. It often causes severe symptoms that affect human's cognitive function, emotion and physical function [1]. The rapid development of modern society has accelerated the pace of life, leading to the fact that depression has become a global medical problem [2]. Since the clinical treatment of depression faces with many challenges, e.g., drug reactions, individual difference, etc., how to accurately detect depression becomes the research focus in medical and artificial intelligence community.

Early depression detection approaches mainly focused on the use of textual information for discovering depression, while other modal information had been neglected. Indeed, the clinical treatment of depression has produced a massive volume of multi-modal patient records, e.g., natural language, facial expressions, gestures, etc., which could provide vivid and accurate descriptions of depression symptoms. Hence, multi-modal depression detection, as an interesting but challenging problem, has gained popularity in the recent literature [3]–[5].

In this paper, we extend this problem into the multi-task scenario and take a further step towards exploring the use of emotional knowledge to improve depression detection. We argue that depression is tightly coupled with emotion in that one helps the understanding of the other. When expressing their symptoms, patients often use some emotional words for emphasizing their feelings. Here is an example of emotion coming into sight, a medical recored "me on a daily basis knowing how depressed and anxious I am". Hence, jointly detecting depression and emotion would bring benefits to each other where a new research topic, multi-modal depression and emotion joint detection, is brought forth.

Multi-modal depression and emotion detection aims to recognize the depressive and affective states of a human using medical knowledge, deep learning and natural language processing (NLP) methods. Different from the traditional single task or text based approaches, there are two intractable challenges in multi-task multi-modal depression detection, i.e., multi-modal fusion and multi-task correlation. Most of the recent works [6]–[8] have paid full attention to the former, while how to capture the correlations across multiple tasks, e.g., whether depression detection reaps the greater benefit from emotion recognition, has been neglected. For illustration, Fig. 1 provides an example to show the presence of both challenges.

To generically address the above two issues, we propose an attentive multi-modal multi-task learning framework, termed

*Yazhou Zhang and Xiang Li contribute equally and share the co-first authorship.

†Corresponding author: Lu Rong.

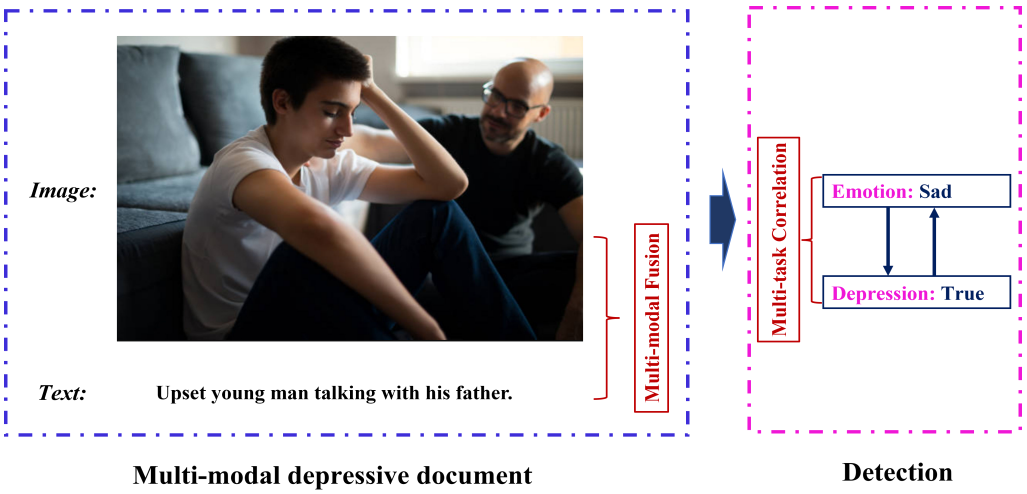

Fig. 1: An example of a multi-modal depressive document.

AMM. The core modules are two attention mechanisms, $viz.$ inter-modal (I_e) and inter-task (I_t) attentions. The main motivation of I_e attention is to learn multi-modal fused representation. The multi-modal representation of the target document is obtained and fed into two LSTM sub-networks to produce its depressive and emotional results. In contrast, I_t attention is proposed to learn the mutual influence between depression detection and emotion recognition. Extensive experiments are conducted on two benchmarking datasets, i.e., DAIC and multi-modal Getty Image depression (MGID) dataset, in comparison with a wide range of strong baselines, including support vector machine (SVM), multi-modal deep convolutional neural network (CNN), bidirectional GRU, hierarchical attention networks (HAN), multi-head attention-based bidirectional long-short memory (MHA-BiLSTM) network, bidirectional encoder representations from transformers (BERT), and two state-of-the-art multi-modal approaches. The results show the effectiveness of the proposed AMM framework. The evaluation also shows that AMM obtains better performance for the main task, i.e., depression detection with the help of the secondary emotion recognition task.

The major innovations of the work presented in this paper can be summarized as follows.

- A attention based multi-task multi-modal model is presented.
- The inter-dependence among related tasks is explicitly modeled.
- We verify the effectiveness of our model by applying it to the task of multi-modal depression and emotion detection.

II. Related Work

We review related work in multi-modal depression detection and multi-modal emotion recognition in this section.

A. Multi-Modal Depression Detection

Generally, multi-modal depression detection targets to identify the depressive attitude of a human expressed via multi-modal language that may involve visual, audio and textual information. As an early example, Senoussaoui [9] presented many different multi-modal depression level predictors using a model fusion approach. Then, Alghowinem et al. [10] developed a machine learning and feature fusion based multi-modal depression detection approach, which showed a remarkable improvement compared to unimodal approach. However, the multi-modal correlation had been ignored. To improve the detection performance, many researchers had proposed a wide and diverse variety of multi-modal fusion approaches. For instance, Ye et al. [4] proposed a deep learning based multi-modal depression detection approach to handle multi-modal feature fusion. Zheng and Yan [3] designed a multi-modal graph attention model embedded with medical knowledge to address the multi-modal data sufficiency problem. The proposed approach significantly improved the classification and prediction performance. In [11], a topic modeling based multi-modal feature vector building approach was proposed to capture effective temporal information in clinical interviews. Zhang et al. [12] presented a decision fusion method that was used to combine a BERT based textual depression detection sub-model and a LSTM based visual depression sub-model.

The above-mentioned works focused on multi-modal feature extraction and multi-modal feature fusion. Few approaches had also attempted to use sentiment or emotion information to enhance the classification performance. In [13], [14] and [15], the sentimental or emotional knowledge were treated as a supplementary information to the multi-modal representation. However, they still treated sentiment analysis and depression detection as two tasks.

B. Multi-Modal Emotion Recognition

Emotion could be seen as fine-grained sentiment, involving complex psychological states such as fear, anger, happiness, etc. Multi-modal emotion recognition aims to identify the emotional polarity expressed in multi-modal documents. In earlier times, Chuang and Wu [16] constructed a multi-modal

emotion recognition framework based on speech signals and textual content. Datcu and Rothkrantz [17] fused early acoustic features with facial expressions for emotion recognition. Zhang et al. [18], [19] proposed a quantum-inspired multimodal sentiment analysis model. Recently, CNN, RNN and their multifarious variants were commonly used to extract visual and sequential features and built multi-modal emotion recognition framework [20], [21].

Emotion recognition in conversation (ERC) has become an active research topic. Majumder [22] described a DialogueRNN model that kept track of the individual party states throughout the conversation and used this information for ERC. Poria et al. [23] created the first multimodal conversational dataset, namely, the multimodal emotionlines dataset (MELD), to facilitate the development of conversational sentiment analysis. Zhang and Li [24] designed a quantum-inspired interactive network model for textual conversational sentiment analysis and showed its effectiveness. However, they did not take the interactions among different modalities into consideration. Liu et al. [25] also chose the multi-task learning to jointly modeling the relations between sarcasm and sentiment. They have achieved the state-of-the-art performance. Xing et al [26] proposed an adapted dynamic memory network where self and inter-speaker influences were modelled individually.

In general, remarkable progress has been made in the current state-of-the-art. Different from them, we tackle all these two problems in a multi-modal depressive scenario with a multi-task learning framework.

III. Methodology

In this section, we detail the proposed AMM framework, which aims to capture both multi-modal fusion and multi-task correlation in an unified paradigm.

A. Task Description

Task Description. Suppose the dataset contains N multi-modal depressive samples, the k^{th} sample X_k could be represented as $\{X_k = (M_k, Y_k)\}$, where M_k, Y_k represent the target multi-modal utterance and the label respectively, and $k \in [1, 2, ..., N]$. In this work, we only consider the textual and vision modalities, e.g., $M_k = (M_k^t, M_k^v)$. But we argue that the proposed framework could be extended into the triple-modal task. Given a multi-modal utterance M_k, how to jointly detect the depression and emotion polarities, i.e., $Y_k = \left(Y_k^{dep}, Y_k^{emo}\right)$. We formulate the problem as follows:

$$\zeta = \prod_k p\left(Y_k | M_k, \Theta\right) \tag{1}$$

where Θ represents the parameter set. In the following sections, we will present each component of our model. The architecture of the AMM model is shown in Fig. 2.

B. Attention based Textual and Visual Encoder

1) Textual Encoder: For text, we assume that there are n words in the k^{th} target document, i.e., $M_k^t = \{tw_1, tw_2, ..., tw_n\}$. Each word $tw \in \mathcal{R}^{d_t}$ is initialized with pre-trained BERT embeddings [27]. We thus feed them into a bidirectional Gated Recurrent Unit (BiGRU) to learn the contextual relationship between the words and the hidden states $H = [h_1^t, h_2^t, ..., h_n^t]$. To measure the contribution of the words, we use the attention mechanism and produce a weighted feature representation h^t, which can be formulated as:

$$\begin{aligned} D &= tanh\left(W_d H + b_d\right) \\ \alpha &= softmax\left(w^T D\right) \\ h^t &= \alpha H \end{aligned} \tag{2}$$

2) Visual Encoder: For video, suppose that there are n clips in the k^{th} target video, i.e., $M_k^v = \{tv_1, tv_2, ..., tv_n\}$. Each input video clip is scaled to 480×360 and its feature vectors are extracted by the pre-trained EfficientNet network [28]. We use average pooling to get a 768 dimensional feature representation. We thus feed each video clip into the GRU unit to produce its hidden state $H = [h_1^v, h_2^v, ..., h_n^v]$. Based on Eq. 2, the attention mechanism is also adopted to get a weighted visual representation h^v. If the target visual document is a static image, we choose to divide the whole image into n visual zones from top to bottom.

C. External Knowledge Augmentation

The participant gender information is the necessary and latent knowledge for document representation, which enables the model to understand the influence of gender difference. Because the research implies that female is easier to be disturbed by depression against male [29]. Inspired by this, we represent the gender information using the pre-trained BERT embedding h^{gen} and thus merge it with the textual hidden representation h^t, to elaborate refined textual representation, i.e., $h^{tg} = h^t \oplus h^{gen}$.

Similarly, since depression patients tend more to deliver darker images, while non-depressive users more likely publish colorful images, we regard the color distribution as another supplement knowledge to improve the visual representation. We extract the HSV space based color histogram of each visual document, denoted as h^{col}, and concatenate it with the original visual representation h^v to obtain the final visual vector, i.e., $h^{vc} = h^v \oplus h^{col}$.

The final textual and visual vectors h^{tg}, h^{vc} are fed into next layer to perform multi-modal fusion.

D. Multi-Head Inter-Modal Attention Fusion

We have obtained the textual and visual representations, i.e., h^{tg}, h^{vc}. Then, an inter-modal multi-head self attention based multi-modal fusion layer is designed to obtain the multi-modal representation $M_k^{(m)}$.

Inspired by Multi-modal Transformer [30], we aim to fuse multi-modal information by learning a latent adaptation across modalities. Given textual and visual modalities t and v with their vectors h^{tg} and h^{vc}, we treat textual modality as $Query$, i.e., $Q_\mu^t = W^{Q_\mu^t} h^{tg}$, and visual modality as as $Keys$ and $Values$, i.e., $K_\mu^v = W^{K_\mu^v} h^{vc}$ and $V_\mu^v = W^{V_\mu^v} h^{vc}$, where

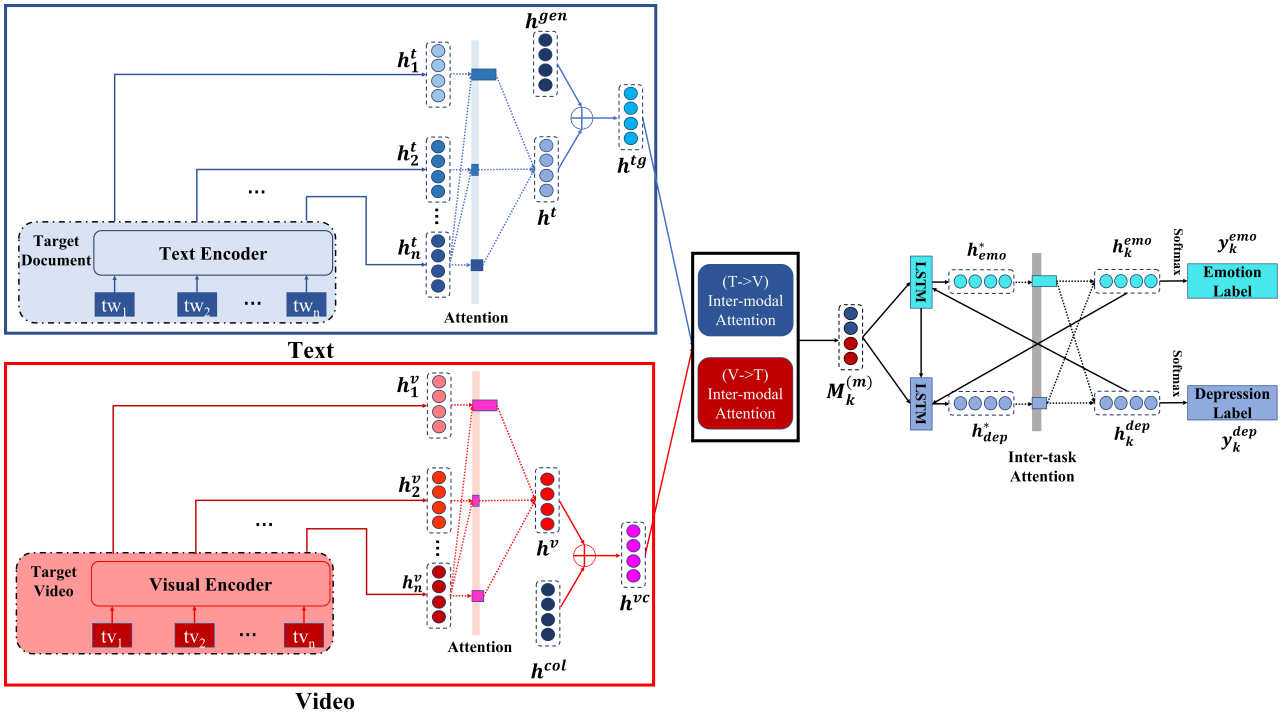

Fig. 2: The overall architecture of the AMM model.

$\mu \in [1, 2, ..., H]$, H is the number of heads, $W_\mu^{Q^t}$, $W_\mu^{K^v}$ and $W_\mu^{V^v}$ are weights. The mapping from t to v is defined as:

$$M_\mu^{t \to v} = softmax\left(\frac{Q_\mu^t K_\mu^v}{\sqrt{d_k}}\right) V_\mu^v \tag{3}$$

Correspondingly, the mapping from v to t is defined as:

$$M_\mu^{v \to t} = softmax\left(\frac{Q_\mu^v K_\mu^t}{\sqrt{d_k}}\right) V_\mu^t \tag{4}$$

Eq. 3 and Eq. 4 will yield H output values respectively. We clarify that the reason to use multi-head attention is that different tasks may focus on different words/clips. Then, these values are concatenated as: $M^{t \to v} = [M_1^{t \to v}, M_2^{t \to v}, ..., M_H^{t \to v}]$ and $M^{v \to t} = [M_1^{v \to t}, M_2^{v \to t}, ..., M_H^{v \to t}]$. In this work, H is set to eight.

Now, we merge them together to obtain the bidirectional multi-modal representation, as shown in Eq. 5

$$M_k^{(m)} = \left[M^{t \to v}; M^{v \to t}\right] \tag{5}$$

E. Inter-Task Attention Mechanism

We design two attention based LSTM sub-networks to capture the correlation between depression and emotion, and predict both depression and emotion labels. Since our focus is depression detection, we regard it as the main task, while emotion recognition is seen as the secondary task, e.g., $(emo \to dep)$. This action ensures the proposed model leverage knowledge from other tasks.

To explicitly learn the relationship between the classes of all the tasks, we design a self-attention mechanism. This is:

$$\begin{aligned} h_{emo}^* &= LSTM_e\left(M_k^{(m)}\right) \\ h_{dep}^* &= LSTM_d\left(M_k^{(m)}\right) \\ h_k^{emo} &= Attention\left(h_{emo}^*, h_{dep}^*\right) \\ h_k^{dep} &= Attention\left(h_{dep}^*, h_{emo}^*\right) \end{aligned} \tag{6}$$

where $Attention(\alpha, \beta) = softmax\left(Q^\alpha K^\beta\right) V^\beta$.

F. Classification

The outputs, e.g., h_k^{dep}, h_k^{emo} are forwarded through the softmax functions to yield both depression and emotion labels. The cross entropy with L2 regularization is used as the loss functions for training each task.

IV. Experimental Setup

In this section, we validate the effectiveness of the AMM framework from an experimental viewpoint.

A. Experiment Settings

Datasets. Given that multi-modal multi-task depression detection is a new topic, the benchmark datasets are relatively limited. In this work, we first choose the distress analysis interview corpus (DAIC) to evaluate the proposed model. DAIC[1] [31], [32] contains 621 multi-modal clinical interviews, which are designed to support the diagnosis of psychological

[1]https://dcapswoz.ict.usc.edu/

TABLE I: Dataset statistics.

Dataset	*Task*	*Classes*	*No. of Utter.*	*RC(%)*
DAIC	Depression	Dep.	397	63.9
		Non.	224	36.1
	Emotion	Sad	372	59.9
		Non.	249	40.1
MGID	Depression	Dep.	2500	50.00
		Non.	2500	50.00
	Emotion	Neutral	1340	26.8
		Happiness	500	10.0
		Fear	500	10.0
		Surprise	500	10.0
		Depressed	1660	33.2
		Anger	500	10.0

TABLE II: Model configurations.

Hyper-parameters	**DAIC**	*MGID*
Epoch	100	
Activations	Relu	
Batch	64	
Learning rate	0.001	0.005
Task order	$(emo \rightarrow dep)$	
dropout	0.3	0.2

TABLE III: Comparison of different models.

Dataset	**Method**	**Depression Detection**		
		P	R	$\mathrm{M_i}$-F1
DAIC	SVM	51.3	49.7	50.1
	Multi-modal DCNN	55.1	56.6	56.5
	BiGRU	53.9	52.5	52.8
	HAN	58.7	59.2	59.0
	MHA-BiLSTM	57.4	58.5	58.3
	BERT	67.2	66.8	66.8
	RCNN-RoBERTa	67.3	66.9	67.1
	Multi-modal Transformer	66.5	67.5	67.5
	Text-AMM	66.7	67.2	67.2
	Image-AMM	63.6	64.0	64.0
	AMM	**69.4**	**68.6**	**68.5**
	△SOTA	(+2.9%)	(+1.4%)	(+1.4%)
MGID	SVM	59.4	59.7	59.5
	Multi-modal DCNN	61.2	61.2	61.1
	BiGRU	62.7	61.6	61.9
	HAN	65.3	65.7	65.6
	MHA-BiLSTM	66.2	66.7	66.4
	BERT	71.3	71.8	71.6
	RCNN-RoBERTa	72.4	73.0	72.8
	Multi-modal Transformer	72.5	71.7	72.2
	Text-AMM	72.1	72.3	72.1
	Image-AMM	64.5	64.7	64.6
	AMM	**74.4**	**74.0**	**74.1**
	△SOTA	(+2.6%)	(+1.3%)	(+1.5%)

distress conditions such as anxiety and depression. Each utterance is segmented at boundaries with at least 300 milliseconds of silence. The number of positive and negative samples are 397 and 224 respectively. As for emotion, we manually annotate the sad feelings of the depressive patients.

Moreover, we also create a weakly labeled multi-modal depression and emotion dataset to support our experiment. As what other researchers did in [1], [18], we set a list of keywords with strongly depression and emotions to query Getty Images, and use the labels of these words to label the retrieved multi-modal documents of the first thirty pages. Based on these documents, we construct a new multi-modal Getty Image depression and emotion (MGID) dataset containing 2500 depressive and 2500 non-depressive samples. We have made our datasets freely downloadable [2]. Table I shows a statistical summary of the datasets.

Evaluation metrics. We adopt *precision* (P), *recall* (R) and *micro-F1* ($\mathrm{M_i}$-F1) as evaluation metrics in our experiments. We also introduce a *balanced accuracy* metric for an ablation test.

Hyper-parameter Setup. The textual and visual inputs are initialized with BERT and EfficientNet. The dimensionality of the embeddings is set to 768. All weight matrices are given their initial values by sampling from a uniform distribution $U(-0.1, 0.1)$, and all biases are set to zeros. We use the Adam algorithm to train the network, and the number of epochs is set to 100. The optimal hyper-parameters are listed in Table II.

[2]This dataset can be accessed on the web page: https://github.com/yzzhang2008/MGID-Dataset

B. Comparison Models

We compare the proposed AMM model with a number of baselines. They are listed as follows.

SVM: It represents the textual document using GloVe vectors [33] and visual image using bag of visual words method [34]. Then, it concatenates them to be the multi-modal representation and feeds into a SVM classifier.

Multi-modal DCNN: It contains two deep CNNs, one is composed of two convolutional layers and a fully connected layer, while the other is composed of six convolutional layers and one fully connected layer. The former is used to extract textual features, the latter is designed for visual features extraction. It takes joint text-level and image-level representations as input, and trains a logistic regression classifier to identify the depression polarity.

BiGRU: It leverages a bidirectional GRU network for learning utterance representations of texts and images and then uses a classification layer to make prediction of depression.

HAN: It [35] uses GloVe vectors to extract low-level representations of the words, and employs a hierarchical attention-based network for depression detection.

MHA-BiLSTM: It [21] extracts the most significant features and builds a multi-head attention-based bidirectional long short term memory network to detect depressive utterances.

BERT: It represents the textual utterances using BERT vectors [27] and visual document using a deep CNN. Then, it merges them together and feeds into the softmax function for depression detection.

RCNN-RoBERTa [36]: It utilizes pre-trained RoBERTa vectors combined with a RCNN in order to capture contextual information.

Multi-modal Transformer [30]: It is a multi-modal pre-trained architecture that combines cross model attention with Transformer, which aims to address multi-modal alignment in an end-to-end manner.

C. Comparative Analysis

DAIC. The experimental results are summarized in Table III. Since DAIC is an unbalanced dataset, we will pay more attention to micro-F1 score here. We observe that SVM performs worst, because low-level feature representation is not effective to alleviate "the semantic gap". Multi-modal DCNN and BiGRU are superior than SVM, showing that learning deep mid-level features can improve classification performance. Since introducing the attention mechanism to capture the valuable features, HAN and MHA-BiLSTM perform better than the above three baselines. This proves the importance of the use of attention in depression detection. By stacking six layers of attentive modules, BERT performs very well, which overcomes HAN and MHA-BiLSTM by a large margin. The reason is that the pre-trained features provided by BERT have stronger discrimination ability. Through presenting improved modifications for training BERT models, RCNN-RoBERTa obtains slightly improvement over BERT. Multi-modal Transformer achieves the best classification results among all baselines. Compared with BERT, the micro-F1 result increase by 1.0%. One possible reason is that it builds an effective cross-modal mapping mechanism.

Text-AMM and Image-AMM perform not very well against other BERT based models, demonstrating that text or visual modalities cannot be treated independently for multi-modal depression detection. The proposed AMM model achieves the best micro-F1 of 68.5% as compared to micro-F1 of 67.5% of the state-of-the-art system (i.e., Multi-modal Transformer). This shows that the proposed AMM framework successfully leverages the advantages of inter-modal and inter-task attentions in modeling multi-modal fusion and multi-task correlation.

MGID. Further, we have evaluated the proposed AMM model on the MGID dataset, which are collected from Getty Image. We can see that the performance differences between all models are not as contrasting as they are on DAIC. We notice that SVM, Multi-modal DCNN, BiGRU, HAN and MHA-BiLSTM perform worse, which do not exceed 70% in term of micro-F1. BERT and its variants obtain superior results, which shows that pre-trained language models (PLMs) have excellent generalization abilities. AMM remarkably overcomes all baselines, and achieves the state-of-the-art performance with the micro-F1 of 74.1%. We attribute the main improvements to both PLMs and two attention mechanisms, which ensures that AMM can model inter-modality fusion and multi-task interaction, and thus refine the final features.

D. STL v/s MTL Framework

We outline the comparison results between the multi-task (MTL) and single-task (STL) learning frameworks in Table IV. Bi-modal (T+V) shows a better performance over unimodal setups.

For depression detection, MTL outperforms STL by a large margin in bi-modality instead of text and visual modalities. The reason is that depression detection involves a higher level of abstraction, where both textual and visual information play key roles. MTL learns more supplementary information when using multi-modal information. For emotion recognition, MTL achieves better performance than STL on text and bi-modal. Because textual information contributes more on understanding semantics. The proposed AMM framework could learn the inter-dependence between two related tasks and improves performance.

E. Ablation Test

We perform the ablation experiments to further analyze the effectiveness of different components of AMM: (1) *No* I_e *Attention* that replaces the cross-modal attentive fusion with multi-modal feature concatenation; (2) *No* I_t *Attention* that only handles depression detection without modeling multi-task correlation; (3) *No Attention* that removes both I_e and I_t attentions from AMM.

The results in Table V show that the inter-task I_t attention contributes the most to overall performance. Because we treat depression detection as the main task, AMM will share emotional knowledge with depression detection via the connectivity from LSTM, for improving the performance. Hence, the contribution of inter-modal attention is less than inter-task attention. In addition, *No Attention* achieves the worst results, which shows that inter-modal I_e attention also plays an important role in AMM. In summary, both of them are indispensable components for AMM.

F. Misclassification Cases

We check the dataset and show a few misclassification cases (text+image), including the cases that MTL predicts correctly while STL fails, and that both setups fails to predict correctly. These cases are shown in Fig. 3.

For the main task, we notice that misclassification for STL framework happens in the situation where the literal meaning of the text differs from its visual counterpart. For example, the text "A middle aged woman sitting in the kitchen at the glass table" shows neutral attitude while its corresponding image clearly expresses depressive feeling. The proposed AMM model leverages the negative emotion to make correct judgment. However, both setups often fail when they have to distinguish depression from the dark and negative scenarios. They might require external information.

For emotion recognition, we see that both setups succeed to model multi-modal incongruity, thanks to the use of inter-model I_e attention. STL wrongly classifies the third sample, but MTL makes the right decision. Because MTL leverages the sharing knowledge from depression. Both setups fail to handle complex scenarios, e.g., the fourth multi-modal document.

TABLE IV: Comparison with single-task learning (STL) and multi-task (MTL) learning frameworks.

Task	Setups	T		V		T+V	
		M_i-F1	Acc	M_i-F1	Acc	M_i-F1	Acc
Depression	STL	71.5	71.2	65.0	65.0	72.3	72.0
	MTL	72.1	72.0	64.6	64.2	74.1	74.5
Emotion	STL	52.7	52.9	50.7	50.5	54.4	54.4
	MTL	55.5	55.4	51.6	51.7	58.1	58.1

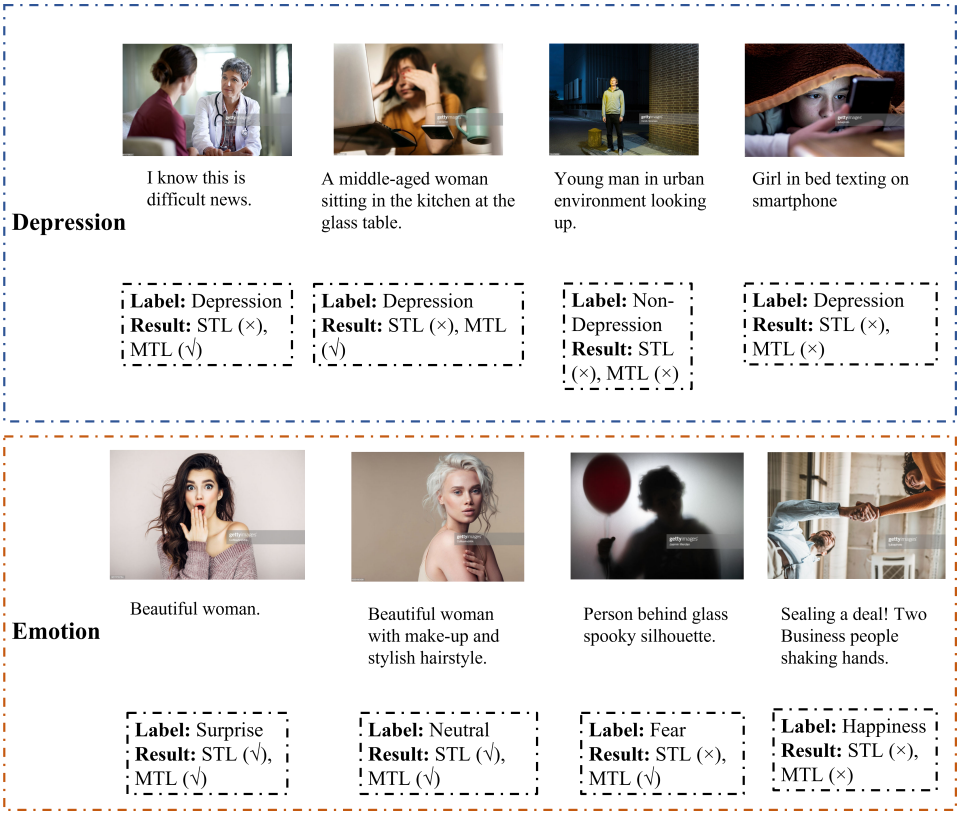

Fig. 3: Wrongly classified multi-modal samples where MTL framework performs better than the STL framework.

TABLE V: Ablation experiment results.

Dataset	Models	Metrics	
		M_i-F1	Acc
DAIC	No I_e Attention	67.1	66.9
	No I_t Attention	65.6	65.6
	No Attention	64.4	64.3
	AMM	68.5	68.5
MGID	No I_e Attention	73.4	73.3
	No I_t Attention	72.5	72.7
	No Attention	71.3	71.1
	AMM	74.1	73.8

V. Conclusions and Future Work

Multi-modal depression and emotion joint detection is an important and challenging AI task. In this paper, we propose a multi-modal deep attentive multi-task learning model, termed AMM. The main idea is to use two attention mechanisms, i.e., inter-modal (I_e) and inter-task (I_t) attentions, to address the problems of multi-modal fusion and multi-task interaction. Comprehensive experiments on two benchmark datasets, DAIC and MGID, show that the effectiveness of AMM over state-of-the-art baselines. To support the development of this research topic, we have created a weakly labeled and large scale MGID dataset. Since there are closely relationships between sentiment, emotion and depression, future works will focus on designing an unified multi-task learning model to capture the correlation among triple or more tasks.

Acknowledgment

This work is supported by National Science Foundation of China under grant No. 62006212, the fund of State Key Lab. for Novel Software Technology in Nanjing University under grant No.KFKT2021B41.

References

[1] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 110–117.

[2] X. Xu, H. Peng, M. Z. A. Bhuiyan, Z. Hao, L. Liu, L. Sun, and L. He, "Privacy-preserving federated depression detection from multi-source mobile health data," *IEEE Transactions on Industrial Informatics*, 2021.

[3] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Graph attention model embedded with multi-modal knowledge for depression detection," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[4] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu, "Multi-modal depression detection based onemotional audio and evaluation text," *Journal of Affective Disorders*, 2021.

[5] N. Abaeikoupaei and H. Al Osman, "A multi-modal stacked ensemble model for bipolar disorder classification," *IEEE Transactions on Affective Computing*, 2020.

[6] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, 2020.

[7] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 344–350.

[8] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, and B. Wang, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Information Fusion*, vol. 62, pp. 14–31, 2020.

[9] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 57–63.

[10] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2016.

[11] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 69–76.

[12] Y. Zhang, Y. Wang, X. Wang, B. Zou, and H. Xie, "Text-based decision fusion model for detecting depression," in *2020 2nd Symposium on Signal Processing Systems*, 2020, pp. 101–106.

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

[14] R. Francese and P. Attanasio, "Supporting depression screening with multimodal emotion detection," in *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 2021, pp. 1–8.

[15] Y. Li, M. Cai, S. Qin, and X. Lu, "Depressive emotion detection and behavior analysis of men who have sex with men via social media," *Frontiers in Psychiatry*, vol. 11, p. 830, 2020.

[16] Z.-J. Chuang and C.-H. Wu, "Multi-modal emotion recognition from speech and text," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 2004, pp. 45–62.

[17] D. Datcu and L. J. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," *Emotion recognition: a pattern analysis approach*, pp. 411–435, 2014.

[18] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, and B. Wang, "A quantum-inspired multimodal sentiment analysis framework," *Theoretical Computer Science*, vol. 752, pp. 21–40, 2018.

[19] Y. Zhang, D. Song, X. Li, and P. Zhang, "Unsupervised sentiment analysis of twitter posts using density matrix representation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 316–329.

[20] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription." in *Interspeech*, 2019, pp. 3302–3306.

[21] Y. Zhang, Y. Liu, Q. Li, P. Tiwari, B. Wang, Y. Li, H. M. Pandey, P. Zhang, and D. Song, "Cfn: A complex-valued fuzzy network for sarcasm detection in conversations," *IEEE Transactions on Fuzzy Systems*, 2021.

[22] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2019, pp. 527–536.

[24] Y. Zhang, Q. Li, D. Song, P. Zhang, and P. Wang, "Quantum-inspired interactive networks for conversational sentiment analysis," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 5436–5442.

[25] Y. Liu, Y. Zhang, Q. Li, B. Wang, and D. Song, "What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 871–880.

[26] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, 2020.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018, pp. 4171–4186.

[28] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[29] E. S. Williams, C. E. Manning, A. L. Eagle, A. Swift-Gallant, N. Duque-Wilckens, S. Chinnusamy, A. Moeser, C. Jordan, G. Leinninger, and A. J. Robison, "Androgen-dependent excitability of mouse ventral hippocampal afferents to nucleus accumbens underlies sex-specific susceptibility to stress," *Biological psychiatry*, vol. 87, no. 6, pp. 492–501, 2020.

[30] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[31] 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," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[32] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.

[33] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[34] Y. Zhang, D. Song, P. Zhang, X. Li, and P. Wang, "A quantum-inspired sentiment representation model for twitter sentiment analysis," *Applied Intelligence*, vol. 49, no. 8, pp. 3093–3108, 2019.

[35] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," 2019.

[36] R. A. Potamias, G. Siolas, and A.-G. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," *Neural Computing and Applications*, pp. 1–12, 2020.