A Deep Multi-task Contextual Attention Framework for Multi-modal Affect Analysis

MD SHAD AKHTAR, Indraprastha Institute of Information Technology - Delhi
DUSHYANT SINGH CHAUHAN and ASIF EKBAL, Indian Institute of Technology Patna

Multi-modal affect analysis (e.g., sentiment and emotion analysis) is an interdisciplinary study and has been an emerging and prominent field in Natural Language Processing and Computer Vision. The effective fusion of multiple modalities (e.g., *text*, *acoustic,* or *visual frames*) is a non-trivial task, as these modalities, often, carry distinct and diverse information, and do not contribute equally. The issue further escalates when these data contain noise. In this article, we study the concept of multi-task learning for multi-modal affect analysis and explore a contextual inter-modal attention framework that aims to leverage the association among the neighboring utterances and their multi-modal information. In general, sentiments and emotions have inter-dependence on each other (e.g., *anger* → *negative* or *happy* → *positive*). In our current work, we exploit the relatedness among the participating tasks in the multi-task framework. We define three different multi-task setups, each having two tasks, i.e., sentiment & emotion classification, sentiment classification & sentiment intensity prediction, and emotion classificati on & emotion intensity prediction. Our evaluation of the proposed system on the CMU-Multi-modal Opinion Sentiment and Emotion Intensity benchmark dataset suggests that, in comparison with the single-task learning framework, our multi-task framework yields better performance for the inter-related participating tasks. Further, comparative studies show that our proposed approach attains state-of-the-art performance for most of the cases.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**;

Additional Key Words and Phrases: Multi-task learning, multi-modal analysis, sentiment analysis, sentiment intensity prediction, emotion analysis, emotion intensity prediction, inter-modal attention

ACM Reference format:
Md Shad Akhtar, Dushyant Singh Chauhan, and Asif Ekbal. 2020. A Deep Multi-task Contextual Attention Framework for Multi-modal Affect Analysis. *ACM Trans. Knowl. Discov. Data* 14, 3, Article 32 (May 2020), 27 pages.
https://doi.org/10.1145/3380744

1 INTRODUCTION

In the last decade, the tremendous growth of social media platforms has overwhelmed the Internet with a variety of distinct and diverse information such as *videos*, *images*, *audios*, and *text*. Sharing a

The work was carried out while he was at IIT Patna.

Authors' addresses: Md S. Akhtar, Department of Computer Science & Engineering, Indraprastha Institute of Information Technology, Delhi (IIIT-Delhi), New Delhi, India, 110020; email: shad.akhtar@iiitd.ac.in; D. S. Chauhan and A. Ekbal, Department of Computer Science & Engineering, Indian Institute of Technology Patna (IIT Patna), Patna, India, 801106; emails: {1821CS17, asif}@iitp.ac.in.

1556-4681/2020/05-ART32 $15.00
https://doi.org/10.1145/3380744

Fig. 1. Both person are uttering "*I'm ok.*" in two separate videos with opposite emotions. The facial expression of the left person (lady) implies the "*happy*" emotion, while the facial expression of the right person (gentleman) conveys his *sadness* (or possibly *anger* as well). For the first case, the textual representation might be adequate for predicting the emotion as "*happy*"; however, for the second case, it is clearly not sufficient for the prediction of *sadness* as emotion. Multi-modal analysis aims to leverage the multiple sources for the desirable predictions.

video on social media platforms is more convenient for users than expressing their feelings through text. They freely discuss current affairs, raise their concerns, or express their opinions in an open and convenient environment facilitated by these social media platforms. On the other hand, various organizations utilize these user inputs as feedback to refine their product or services. The amount of information generated daily through different social media platforms is quite colossal, and therefore researchers are quite interested in mining this vast information.

This knowledge extracted by tracking attitudes and feelings on the web, can be used as a source of honest user opinion for services, products, brands, and people. It may also be used for understanding the conversations and identifying relevant content. Organizations want to process this information as useful feedback to improve their products and services. Hence, multi-modal analysis (e.g., emotion recognition [33], sentiment analysis [41], and questioning-answering [46]) has been an emerging field of study nowadays.

A video, in general, has all the ingredients of multi-modal information such as *visual frames*, *acoustics*, and *transcripts*. The prime challenge in the multi-modal analysis is to fuse these distinct information to have the correct predictions. However, the fusion of multi-modal information is not always effective, as different sources often bring their characteristics, and some of them may contain noise as well. For example, there might be some disturbances or noise present in a video due to which acoustic features like tone, intensity, energy, and pitch, cannot contribute equally in the actual prediction.

Further, some representations can be ambiguous, as well. For example, the textual representation "*I'm ok.*" cannot reveal the true emotion for a sad or angry person. On the other hand, the same representation can belong to a "*happy*" person as well. Therefore, in such cases, it is often desirable to have other sources of representations to resolve the ambiguity. For the above case, visual expressions (or acoustic features like tone and pitch) can assist the model in the disambiguation of the prediction. An example scenario is depicted in Figure 1.

The multi-task learning (MTL) framework aims to achieve generalization for the participating tasks. It exploits the inter-relatedness among the participating tasks to improve the individual performance through a shared representation. Overall, the MTL framework has three basic advantages over the single-task learning (STL) framework. These are, (a) to achieve generalization over the participating tasks; (b) to improve the performance of individual tasks by leveraging the richness of other participating tasks; and (c) to the reduces complexity of the overall system (i.e., multiple problems/tasks are solved in an MTL system simultaneously).

Therefore, motivated by the advantages of MTL, in our current work, we propose an MTL framework that involves the various tasks of sentiment and emotion analysis, i.e., sentiment classifica-

tion, sentiment intensity prediction, emotion classification, and emotion intensity prediction. We define the following three sets of tasks: i.e., sentiment & emotion classification (S_C, E_C), sentiment classification & sentiment intensity prediction (S_C, S_I), and emotion classification & emotion intensity prediction (E_C, E_I), for our MTL framework. For any particular set of tasks (e.g., sentiment and emotion classification), we aim to exploit the inter-relatedness among them to extract the predictions for all the tasks (e.g., sentiment prediction as *negative* or *positive* and emotion prediction as *happy*, *surprise*, *sad*, *fear*, *disgust*, or *anger*). For instance, the information about *anger* and *sad* emotions can assist in the prediction of *negative* sentiment and vice-versa. Similarly, "*happy*" emotion should assist in the prediction of "*positive*" sentiment and so on.

In our proposed framework, we utilize the contextual multi-modal information for learning the expressed sentiments and/or emotions for a sequence of utterances in a video. Since the diversity among different modalities across contextual utterances plays a very crucial role in multi-modal analysis, hence one key challenge in our contextual multi-modality framework is the effective fusion of the input modalities considering the neighboring utterances for the prediction. Therefore, we also introduce an attention mechanism to calculate attention scores for both contextual utterances and inter-modalities. Our proposed approach applies attention over both these sources of information simultaneously (i.e., contextual utterance and inter-modal information), and aims to reveal the most contributing features for the classification. We hypothesize that applying attention to contributing neighboring utterances and/or multi-modal representations may assist the network to learn in a better way. Let us assume that an utterance U_x having T_x (i.e., Textual features), V_x (i.e., Visual features), and A_x (i.e., Acoustic features). Now let contextual utterance of U_x be U_y having Textual, Acoustic, and Visual features as T_y, A_y, and V_y, respectively. To produce better and richer multi-utterance contextual-attention representation of input utterances, our model computes attention *(i.e., relatedness)* among different modalities (like T_x and T_y, T_x and V_y, and T_x and A_y) of target and contextual utterances. We evaluate our proposed approach on the CMU-Multi-modal Opinion Sentiment and Emotion Intensity (MOSEI) dataset [56], and observe that the proposed method outperforms various existing state-of-the-art models for both sentiment and emotion analysis.

1.1 Problem Definition

It is true that sentiments [29] and emotions [18] are closely related to each other. Most of the emotional states have a clear distinction of being a positive or negative situation. Emotional states, e.g., "*happy*" and "*surprise*[1]" indicate positive scenarios, while the emotions "*sad*," "*disgust*," "*fear*," and "*anger*" suggest negative situations. Further, in many cases, it is desirable to know the degree of sentiment (or emotion) along with the sentiment (or emotion) class as well. The degree of sentiment (or emotion) corresponds to the intensity level of the expressed sentiment (or emotion). A user might express the same sentiment (e.g., "*positive*") in two different scenarios with different intensities. For instance, one can be extreme (e.g., "*hurray... we won the world cup!*"), while the other can be mild (e.g., "*its a beautiful day today!*"). The sentiment in both the cases is "*positive*"; however, they are distant apart on the intensity spectrum and may not need similar kind of response, action, or attention. Hence, knowledge of the sentiment class along with the degree are important pieces of information and often desirable in several real-world applications, e.g., e-commerce and politics.

Motivated by the association of sentiment & emotion classification, classification & intensity prediction, and the effectiveness of the MTL paradigm, we propose a multi-task framework that learns the association among the participating tasks for enhanced performance. In particular, we

[1]In some cases, surprise can also belong to a negative situation.

define three multi-task setups considering the sentiment classification, sentiment intensity prediction, emotion classification, and emotion intensity prediction tasks as follows:

—Sentiment Classification and Emotion Classification (S_C, E_C);
—Sentiment Classification and Sentiment Intensity Prediction (S_C, S_I); and
—Emotion Classification and Emotion Intensity Prediction (E_C, E_I).

The first setup aims to exploit the inter-relatedness between the classification tasks on two dimensions of the affective analysis, while the other two setups aim to leverage the relationship among the classification and its degree on a single affect dimension (i.e., either sentiment or emotion). We furnish the details of these three setups in Section 4.3.

Our current work is an extension of one of our previous efforts on MTL for multi-modal sentiment and emotion analysis [1]. However, the main differences between our current work and [1] are as follows: (a) In [1], we solve the problem of sentiment and emotion classification in a multi-task framework, wherein our current work, we also include two other sets of tasks for MTL, i.e., sentiment classification & sentiment intensity prediction and emotion classification & emotion intensity prediction; (b) unlike [1], our current work also studies the relationship between the classification and regression problems in the MTL framework for two affect dimensions, i.e., sentiment and emotion analysis; (c) we provide a comprehensive qualitative analysis of the STL vs. MTL framework for all four tasks; and (d) in addition to the sentiment and emotion classification, we provide state-of-the-art results for the sentiment intensity prediction and emotion intensity prediction tasks and also report the improved emotion classification performance in E_C, E_I multi-task setup.

We highlight our contributions as follows: **(a)** *we leverage the inter-dependence of two related tasks in improving each other's performance using an effective multi-modal framework*; **(b)** *we propose three multi-task setups for four sentiment and emotion analysis tasks*; **(c)** *we explore the relatedness among two classes of tasks (i.e., classification-classification and classification-regression) in the multi-task framework*; **(d)** *we develop a context-based inter-modal attention module that effectively attends to the contextual utterances and the available modality inputs as per their importance to the current utterance.* For example, let a video $V = u_1, u_2, \ldots, u_10$ has 10 utterances and we want to predict the sentiment/emotion of an utterance u_i, $i \in \{1 \ldots 10\}$. Further, assume that (which is given in our case) for each utterance we have three inputs, i.e., $< u_i^T >, < u_i^A >, < u_i^V >$, corresponding to the available modalities (i.e., *text*, *acoustic* and *visual*), and to classify an utterance u_1, textual features of u_2 & u_4, acoustic features of u_1 and visual features of u_6, u_3 & u_7 are of higher importance others. Our contextual inter-modal (CIM) attention framework can attend to such diverse and distinct information; and **(e)** *we achieve state-of-the-art performance for all the four tasks.*

The organization of the article is as follows. In Section 2, we present a review of the existing literature concerning multi-modal and multi-task sentiment/emotion analysis. Various details of the proposed methodology are described in Section 3. Experimental results and detailed analysis are furnished in Section 4. Finally, in Section 5, we discuss our conclusions and potential future directions.

2 RELATED WORK

A study on multi-modal analysis suggests that it is an extended area as compared to text-based analysis. In this section, we present a brief survey on the research for multi-modal sentiment and emotion analysis.

2.1 Multi-modal Sentiment Analysis

In recent years, multi-modal sentiment analysis has become an emerging area of study and gained a lot of attention worldwide [21, 26, 34, 56, 57]. A comprehensive review of the recent multi-modal

sentiment analysis works has been studied in [32]. A cross-modality consistent regression (CCR) model is proposed in [53], which utilized both the state-of-the-art visual and textual sentiment analysis techniques. The authors fine-tuned a convolutional neural network (CNN) for extracting the visual features and then trained a distributed paragraph vector model to learn the textual features. On top of these textual and visual features, the authors learned a multi-modal regression model for the final sentiment classification. Zadeh et al. [58] proposed a multi-modal dictionary-based framework to learn the visual and acoustic features when expressing sentiment. The authors also introduced a multi-modal dataset, i.e., the CMU Multi-modal Opinion Sentiment Intensity (CMU-MOSI) dataset, the first of its kind to enable the studies of multi-modal sentiment intensity analysis, and calculated the regressor's performance based on mean absolute error. In another work, Wang et al. [49] proposed a select-additive learning approach that improves the generalizability of trained neural networks for the multi-modal sentiment analysis. An application of Tensor Fusion Network (TFN) is introduced in [54], for the effective fusion of input modalities. The multi-modality TFN framework aims at learning both the intra- and inter-dynamics of the participating modalities (i.e., *text*, *acoustic*, and *visual*). They extracted the textual features by GloVe [31] embeddings, visual features using a three-dimensional CNN [24], and acoustic features by CovaRep [14]. On the other hand, Chen et al. [11] proposed an attention-based Gated Multi-modal Embedding (GME) framework for the word-level fusion of multi-modality inputs.

Most of these prior research primarily focused on extracting features directly from each modality (separately), and then fuse these features for the final classification. Thus, these often ignore the deep semantic correlations between the modalities while building the models. A deep semantic network, MultiSentiNet, was proposed in [52] to model the semantics and correlation between text and images. The authors extracted the textual and visual features from a long short-time memory (LSTM) and VGGNet [45] model, respectively. They, at first, used a fusion layer for aggregating the extracted features to obtain a final multi-modal representation, and then a softmax classifier is employed at the top for sentiment classification. Further, previous works did not account for the contextual information while leaning the sentiments of utterances in a video. Poria et al. [34] proposed an LSTM based framework for the sentiment classification to capture the contextual dependencies among the utterances. In another work, a multi-kernel learning-based method is proposed to combine the three modality inputs (i.e., textual, acoustic, and visual) in [37]. A multi-attention block (MAB) framework is introduced for sentiment classification to capture the inter-modality information across modalities in [57]. Blanchard et al. [6] proposed a multi-modal fusion model that exclusively uses high-level visual and acoustic features for the sentiment classification. Sheikh et al. [44] proposed a deep canonical correlation analyzer that focuses on improving the representations of modalities (text and acoustic) for sentiment analysis. An application of the TFN [54] is found in [41], where the authors employ the TFN network to fuse the multiple modalities at each time-step in a RNN.

Recently, a quantum-inspired bi-modal (*text & image*) based sentiment analysis framework was introduced in [59] to fill the semantic gap and model the correlations between the two modalities via a density matrix. Deng et al. [15] developed a multi-modal neural architecture for emotion behavior analysis with respect to the valence and arousal on a continuous scale. Their system first integrates visual information over the time using an LSTM network and then combines it with utterance level acoustic and text cues. Cummins et al. [13] introduced a different Bag-of-Words (BoWs) paradigms to aid sentiment detection. The authors extracted the textual features by Google2SRT2[2] toolkit, the visual features by OpenFace toolkit [4], and the acoustic features by

[2]http://google2srt.sourceforge.net/en/.

DeepSpectrum [2]. Subsequently, the authors quantified the extracted features into a BoW representation using the openXBOW toolkit [43].

In another work, Cambria et al. [8] explored the speaker-dependent and speaker-independent dimensions of the multimodal sentiment analysis. They studied the generalizability of the model through a context-based framework. Chauhan et al. [10] exploit the interaction between a pair of modalities through an application of the Inter-modal Interaction Module (IIM) that closely follows the concepts of an auto-encoder. The interaction among the modalities was utilized in a RNNs based attention framework for the multi-modal sentiment and emotion analysis.

2.2 Multi-modal Emotion Recognition

Poria et al. [36] proposed a multi-kernel fusion technique for emotion prediction, where the authors employed a deep CNN for extracting the textual features and fused it with other modalities (*visual & acoustic*). An attention-based CNN for multi-modal emotion recognition has been introduced in [25]. The authors employed separate CNN's to extract the features from speech spectrograms and embedded word sequences and applied an attention mechanism to learn the multi-modal representation between speech and textual modalities. In another work, a convolutional deep belief network (CDBN) framework is proposed for multi-modal emotion recognition in [40]. They showed that CDBN learns to extract the salient multi-modal (acoustic and visual) features for emotion classification in an unsupervised manner. Another CNN-based multi-modal emotion recognition system was proposed in [48]. The author utilized a CNN for extracting the textual features, while a deep residual network of 50 layers was employed for the visual modality.

A feature level fusion based self-attention mechanism is proposed for multi-modal emotion detection in [21]. Zadeh et al. [55] proposed a memory fusion network (MFN) that explicitly accounts for both view-specific interactions and cross-view interactions. Authors employed LSTM for the view-specific interaction, whereas for the cross-view interaction, an attention mechanism has been proposed. Recently, a dynamic fusion graph (DFG) for the fusion of tri-modal inputs has been proposed in [56]. The authors extended the MFN [55] by incorporating the DFG (called as Graph-MFN) for the fusion. They also introduced a multi-modal sentiment and emotion recognition dataset (i.e., CMU-MOSEI), consisting of more than 22K utterances. In another work, Williams et al. [50] proposed an input-level fusion technique followed by a deep neural network layer (i.e., CNN, LSTM & GRU) to combine three modalities for emotion intensity prediction.

A recursive multi-attention with a shared external memory-based approach is proposed for emotion recognition in [5]. Patwardhan et al. [30] proposed a support vector machine (SVM)-based feature level fusion mechanism for mixed emotion detection. They employed openEar toolkit and FaceAPI for the feature extraction of multi-modal audio and visual continuous data, respectively. In another work, Fu et al. [19] proposed an auto-encoder and canonical correlation analysis-based approach for emotion recognition. At first, the authors employed a modality-wise sparse auto-encoder to extract the intermediate representation. Subsequently, a shared feature representation is formed based on the correlation coefficients. The shared multi-modal feature representation was then utilized for emotion recognition.

Rahdari et al. [38] proposed a multimodal emotion recognition system using facial landmarks. They employed two modalities, i.e., *affective speech* and *facial expression*, for the recognition of emotions. For affective speech, the common low-level descriptors and spectral audio features were extracted, whereas for the visual feature extraction, they exploited the displacement of specific landmarks across consecutive frames of an utterance. In [28], the authors addressed the issues related to the utterance with one or more missing modalities, i.e., their framework handled an arbitrary number of modalities during emotion recognition. They exploited the presence of three modalities (i.e., *text*, *image*, and *hashtags*) in their network, and the absence of any modality did

not have any consequences on the network architecture. Huang et al. [23] proposed a speech emotion recognition system considering the verbal and nonverbal speech sounds. At first, the authors developed an SVM-based verbal/nonverbal sound detector, and subsequently, for each verbal/nonverbal segment, the emotion and sound features were extracted. Finally, a sequence of the feature vector for the entire dialog was fed to an LSTM based architecture, which predicted the emotional sequence of the dialog.

2.3 Multi-task Learning for Sentiment and Emotion Analysis

To the best of our knowledge, the MTL for multi-modal sentiment and emotion analysis has not been studied except for our previous attempt [1]. However, there are a few studies at the intersection of the MTL and multi-modal analysis for different domains such as autonomous driving [12], semantic goal navigation, and embodied question answering [9].

In general, the MTL for sentiment and emotion analysis seems an exciting field of study [3, 3, 16, 42, 51]. A multi-task framework has been introduced for emotion recognition in two-dimensional continuous space in [51]. A recurrent neural network (RNN)-based MTL framework for fine-grained sentiment classification was proposed in [3]. The authors considered the 3-way and 5-way classification as the two tasks in the MTL framework. In another work, Deriu et al. [16] proposed a deep CNN-based MTL framework for sentiment classification of Italian Twitter messages. Their MTL framework learns three tasks related to sentiment analysis, i.e., subjectivity prediction, polarity prediction, and irony detection.

Our proposed approach differs from the various existing systems, primarily, on the basis of the contextual inter-modal attention mechanism. A few existing works [6, 36, 54, 58] do not account for the contextual information at all, while some [34] considered the context of an utterance as a straight-forward sequence, i.e., no attention mechanism was employed. In another work, Zadeh et al. [57] computed the attention weights over the multiple modalities; however, they ignored the contextual information. In contrast, in our proposed work, we compute attention weights over the contextual utterances across all the available multi-modal inputs. Thus, it ensures to reveal the contributing features across *multiple modalities* and *contextual utterances* for sentiment and emotion analysis simultaneously. Moreover, our proposed work handles two related problems, i.e., sentiment and emotion analysis, simultaneously in a multi-modal scenario—one of the very first attempts to the best of our knowledge.

In [46], a related attention mechanism is studied for visual question-answering. In contrast to our contextual inter-modal attention mechanism, the authors apply attention to the spatial domain over different positions of the image. In one of our earlier works, we have proposed an inter-modal attention framework for the multi-modal sentiment analysis [20]. However, the key differences are, (a) the system [20] addressed only the sentiment classification, while our current work addresses four different tasks, i.e., sentiment classification, emotion classification, sentiment intensity prediction, and emotion intensity prediction; and (b) since only one task (i.e., sentiment classification) was addressed in [20], the underlying framework was an STL framework. In comparison, we solve four related problems in three MTL setups.

3 MULTI-TASK LEARNING FOR MULTI-MODAL ANALYSIS

In our proposed framework, we aim to leverage the multi-modal and contextual information in a MTL framework for the prediction of multiple tasks simultaneously. As stated earlier, a video consists of a sequence of utterances, and their semantics often have inter-dependencies on each other. For each utterance in a video, we aim to predict either the sentiment and emotion, sentiment and intensity, or emotion and intensity together. We utilize all three available modalities, i.e., text, acoustic, and visual, for learning the model. Initially, extracted features from each modality are fed

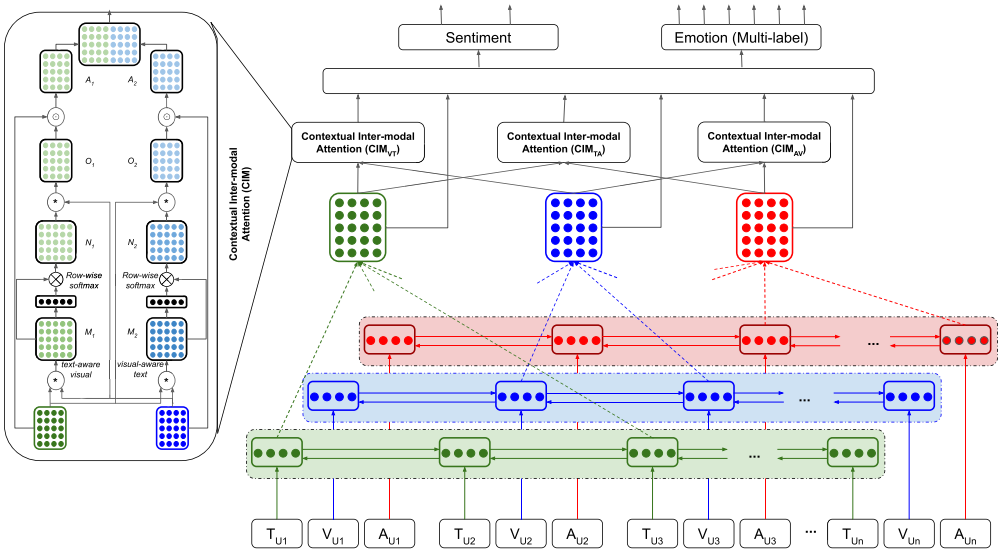


Fig. 2. Overall architecture of the proposed framework. CIM attention computation between *visual* and *text* modality.

to three parallel bidirectional gated recurrent units (Bi-GRU) layers (i.e., one Bi-GRU per modality input). Each Bi-GRU layer captures the modality-wise sequential pattern for all the utterances in the sequence. Once the contextual information is captured, we compute the joint-association among the utterances and the participating modalities, through our contextual inter-modal (CIM) attention mechanism. The intuition behind the CIM mechanism to leverage the contributing features in the contextual vicinity of an utterance. Also, it enables the model to filter the redundant or noisy features to participate in the final prediction. Our CIM attention module operates on a pair of input modalities, i.e., *text-acoustic*, *visual-acoustic*, and *text-visual*. Residual skip-connection [22] is an efficient strategy to regulate the flow of the gradient down to the lower layers. It offers a bypass connection to the lower layers by skipping the complex modules within the network. Therefore, motivated by the residual skip connection [22], the outputs of pair-wise attentions along with the representations of individual modalities are concatenated. Finally, the concatenated representation is shared across the two branches of our proposed network—corresponding to two tasks, i.e., one for each task in the multi-task framework. The shared representation will receive gradients of error from both the branches and accordingly adjust the weights of the models. Thus, the shared representations will not be biased to any particular task, and it will assist the model in achieving generalization. The empirical evidence supports our hypothesis (c.f. Table 3). A high-level architecture is depicted in Figure 2.

3.1 Contextual Inter-modal Attention Framework

Our contextual inter-modal attention framework works on a pair of modalities. At first, we capture the cross-modality information by computing a pair of matching matrices $M_1, M_2 \in \mathbb{R}^{u \times u}$, where *"u"* is the number of utterances in the video. Further, to capture the contextual dependencies, we compute the probability distribution scores ($N_1, N_2 \in \mathbb{R}^{u \times u}$) over each utterance of cross-modality matrices M_1, M_2 using a softmax function. This essentially computes the attention weights for the contextual utterances. Subsequently, we apply soft attention over the contextual inter-modal matrices to compute the modality-wise attentive representations ($O_1 \& O_2$). Finally, a multiplicative

gating mechanism [17] ($A_1\&A_2$) is introduced to attend the important components of multiple modalities and utterances. The concatenated attention matrix of $A_1\&A_2$ then acts as the output of our CIM attention framework. The entire process is repeated for each pair-wise modalities, i.e., *text-visual*, *acoustic-visual*, and *text-acoustic*. We illustrate and summarize the proposed methodology in Algorithm 1.

ALGORITHM 1: Multi-task Multi-modal Emotion and Sentiment (MMES)

procedure MMES(t, v, a)
  $d \leftarrow 100$ ▷ GRU dimension
  $T \leftarrow biGRU_T(t, d)$
  $V \leftarrow biGRU_V(v, d)$
  $A \leftarrow biGRU_A(a, d)$
  $Atn_{TV} \leftarrow CIM\text{-}Attention(T, V)$
  $Atn_{AV} \leftarrow CIM\text{-}Attention(A, V)$
  $Atn_{TA} \leftarrow CIM\text{-}Attention(T, A)$
  $Shared_{Rep} \leftarrow [Atn_{TV}, Atn_{AV}, Atn_{TA}, T, V, A]$ ▷ Concatenation and residual connections
  $polarity \leftarrow Sentiment(Shared_{Rep})$
  $emotion \leftarrow Emotion(Shared_{Rep})$
  return $[polarity, emotion]$

procedure CIM-Attention(X, Y)
  /*Inter-modality information*/
  $M_1 \leftarrow X.Y^T$
  $M_2 \leftarrow Y.X^T$
  /*Contextual Inter-modal attention*/
  for $i, j \in 1, \ldots, u$ **do** ▷ $u = \#utterances$
    $N_1(i,j) \leftarrow \frac{e^{M_1(i,j)}}{\sum_{k=1}^{u} e^{M_1(i,k)}}$
    $N_2(i,j) \leftarrow \frac{e^{M_2(i,j)}}{\sum_{k=1}^{u} e^{M_2(i,k)}}$
  $O_1 \leftarrow N_1.Y$
  $O_2 \leftarrow N_2.X$
  /*Multiplicative gating*/
  $A_1 \leftarrow O_1 \odot X$ ▷ $\odot$: Element-wise multiplication
  $A_2 \leftarrow O_2 \odot Y$
  return $[A_1, A_2]$

4 DATASETS, EXPERIMENTS, AND ANALYSIS

In this section, we describe the datasets used for our experiments and report the results along with necessary analysis.

4.1 Datasets

We perform all experiments on the CMU-MOSEI dataset [56]. It consists of a total of 3,229 videos from 1,000 speakers, which results in approximately 23,000 utterances. The training, validation, and test splits are 16,216, 1,835, and 4,625 utterances, respectively.

There are six emotion values associated with each utterance, which represent the degree of emotion for *surprise*, *happy*, *sad*, *fear*, *disgust*, and *anger*. Each utterance can have zero (no emotion),

Table 1. Dataset Statistics for CMU-MOSEI

	Speakers	Videos	Utterance	Sentiment		Emotion					
				Positive	Negative	Anger	Disgust	Fear	Happy	Sad	Surprise
Train		2,250	16,216	11,499	4,717	3,506	2,946	1,306	8,673	4,233	1,631
Dev	1,000	300	1,835	1,333	502	334	280	163	978	511	194
Test		679	4,625	3,281	1,344	1,063	802	381	2,484	1,112	437

(a) Each utterance contains multi-modal information.

	Emotion(s)						
	No	One	Two	Three	Four	Five	Six
# Utterances	3,372	11,050	5,526	2,084	553	84	8

(b) Statistics of multi-label emotions.

one (single emotion), or more than one emotion values (multi-label). For the experiment, we take 7-classes (6 *emotions* + 1 *no emotion*), and we try to optimize the binary-cross entropy loss for each of the classes. While in the case of sentiment prediction, the classes are disjoint, i.e., $value < 0$ represents the negative (Neg) sentiment, and $value \geq 0$ represents the positive (Pos) sentiment. Detailed statistics of the CMU-MOSEI dataset are shown in Table 1. The dataset is available to download through the CMU Multi-modal Data SDK[3].

4.2 Feature Extraction

The CMU-MOSEI dataset contains both raw data as well as pre-computed feature vectors for every word in an utterance. For the experiments, we make use of the pre-computed feature vectors. The feature vectors correspond to the *GloVe* [31] embeddings, *CovaRep* [14] representation, and *Facets*[4] representation for the *textual*, *acoustic*, and *visual* features, respectively. To obtain feature representation for an utterance, we compute the average over all words in the utterance. The resulting feature vectors have dimensions of 300, 74, and 35 for the *text*, *acoustic*, and *visual*, respectively.

4.3 Multi-task Frameworks

As mentioned in Section 1, the MTL paradigm aims to leverage the inter-dependence among the participating tasks. In this work, we employ an MTL framework for three different problems (sentiment, emotion, and intensity) involving four tasks, i.e., sentiment classification, sentiment intensity prediction, emotion classification, and emotion intensity prediction. Though the basic architecture for all three sets of multi-tasks is the same, the loss function and activations at the output layer differ according to the underlying tasks. Below, we define our three multi-task setups.

(1) **Sentiment Classification and Emotion Classification (S_C, E_C):** In this multi-task setup, we perform two classification tasks, i.e., sentiment classification and emotion classification, together. The shared representation $Shared_{Rep}$ is fed to both sentiment and emotion branches for the respective predictions. We use the *softmax* layer for the sentiment prediction, whereas a *sigmoid* layer with seven neurons (corresponding to 6 emotions + 1 no emotion) is employed to predict the mixed emotions. For emotion prediction, we define a *threshold* and take all the emotion classes whose respective values are above the *threshold*. We optimize *categorical cross-entropy* and *binary cross-entropy* losses for the sentiment and emotion classification, respectively.

[3] https://github.com/A2Zadeh/CMU-MultimodalDataSDK.
[4] https://pair-code.github.io/facets/.

(2) **Sentiment Classification and Sentiment Intensity Prediction (S_C, S_I):** Knowledge of sentiment class along with the degree of sentiment are important piece of information in many cases. In this framework, we learn the sentiment classification and its intensity in a multi-task framework. Similar to the earlier case, we utilize *softmax* and *categorical cross-entropy* loss for the sentiment prediction. For sentiment intensity prediction, we employ *tanh* activation function with *binary cross-entropy* for predicting the intensity in the range [−1, +1] (we normalize the sentiment intensity range [−3, +3] to [−1, +1] for the experiments).

(3) **Emotion Classification and Emotion Intensity Prediction (E_C, E_I):** Similar to the multi-task setup for sentiment classification and sentiment intensity prediction, in this multi-task setup, we perform emotion classification and emotion intensity prediction together. For emotion intensity prediction, we normalize the intensity score in the range [0, 1] and use the *sigmoid* activation function with *binary cross-entropy* loss for the prediction.

4.4 Experiments

We use the Python-based Keras[5] library with TensorFlow[6] as the backend for the implementation of our models. For evaluation, we compute *F1-score* and *accuracy* values for sentiment classification and *F1-score* and *weighted-accuracy* (*W-Acc*) [47] for emotion classification. For emotion classification, we choose weighted accuracy as an evaluation metric due to unbalanced samples across various emotions, and it is also in line with the other existing works [56]. The formulation of *W-Acc* is as follows:

$$W\text{-}Acc = \frac{TP \times N/P + TN}{2N}$$

where TP and TN refer to the *true positive* and *true negative* predictions, whereas P and N signify the total number of positive (i.e., *true positive* + *false negative*) and negative (i.e., *false positive* + *true negative*) samples in the dataset. To measure the performance of intensity prediction tasks, i.e., sentiment intensity prediction and emotion intensity prediction, we compute the mean-squared-error (MSE), mean-absolute-error (MAE), Pearson correlation score (PEAR), and Cosine similarity (COS) as the evaluation metrics. While higher values of Pearson score and Cosine similarity are the indicators of better performance, lower values of MSE and MAE correspond to better performance.

We use Bi-directional GRUs having 300 neurons, each followed by a fully-connected layer consisting of 100 neurons. Utilizing the fully-connected layer, we project the input features of all the three modalities to the same dimension. We set $dropout = 0.3$ as a measure of regularization for the experiments. In addition, we use $dropout = 0.3$ for the recurrent layers. We employ *rectified linear unit (ReLU)* as the activation function in the intermediate layers. For training the network we set the batch size = 32 and use *Adam* optimizer with *cross-entropy* as the loss functions. We run the experiments for 50 epochs while saving the best epoch seen so far. We run each experiment 5 times and report the average of the accuracies in the article. We pad all the utterances up to the maximum length, i.e., 98. A summary of the hyper-parameters used in the experiments are listed in Table 2. Code of the article is available at [https://bit.ly/3azSPs1].

Since our proposed approach requires at least two modalities to compute the contextual inter-modal attention, we experiment with bi-modal and tri-modal input combinations for the proposed approach, i.e., taking *text-visual*, *text-acoustic*, *acoustic-visual,* and *text-visual-acoustic* at a time. As our proposed framework does not account for uni-modal inputs, for completeness, we also experiment with a variant of the proposed approach where we apply self-attention on the utterances

[5] https://keras.io.
[6] https://www.tensorflow.org/.

Table 2. Model Configurations

Parameters	Values
Bi-GRU	2×200 *neurons*, *dropout = 0.3*
Dense layer	100 *neurons*, *dropout = 0.3*
Activation	*ReLU*
Output	*Softmax* (Sent) & *Sigmoid* (Emo)
Optimizer	*Adam* (*lr = 0.001*)
Loss	*Binary cross-entropy*
Threshold	0.4 (F1) & 0.2 (W-Acc) for multi-label
Batch	*16*
Epochs	*50*

of input modality. Similar to the bi-modal and tri-modal scenarios, the uni-modal input is, at first, passed through a Bi-GRU layer, which learns the contextual information from the sequence of utterances in a video. Subsequently, we compute the self-attention on the hidden representation (i.e., $u \times d$), and forward it to the upper layers for prediction.

In a single-task framework, we build separate systems for each task at hand, i.e., sentiment classification, sentiment intensity prediction, emotion classification, and emotion intensity prediction, whereas in multi-task framework a joint-model is learned for the multiple tasks. As mentioned in Section 4.3, we experiment with all three set of tasks for our MTL, i.e., sentiment & emotion classification (S_C, E_C), sentiment classification & intensity prediction (S_C, S_I) and emotion classification & intensity prediction (E_C, E_I). We report the experimental results of our STL and MTL frameworks in Table 3.

To enable convenient performance analysis of the STL and MTL frameworks for a specific task, we prefer to report the results in a task-specific order against the framework-specific order. The first row of Tables 3(a)–(d) reports the obtained results for the four problems (i.e., S_C, E_C, S_I, and E_I) in a STL framework. Whereas, the other rows in the tables (i.e., except the first row) report results for one of the multi-task frameworks (c.f. Section 4.3). For instance, the second and third rows of the Table 3(a) report the sentiment classification (S_C) results obtained in the (S_C, E_C) and (S_C, S_I) MTL frameworks, respectively. The results of other tasks in the (S_C, E_C) and (S_C, S_I) MTL frameworks (i.e., E_C and S_I) are reported in their respective tables (i.e., second rows of Table 3(b) and 3(c) for E_C and S_I, respectively). Furthermore, each of these rows reports results for different evaluation metrics and seven combinations of the available input modalities (i.e., *text* (T), *acoustics* (A), and *visual* (V)).

In sentiment classification (S_C) task, our single-task framework reports an F1-score of 77.67% and accuracy value of 79.8% for the tri-modal inputs, as depicted in Table 3(a). Our proposed MTL framework obtains an improved performance for both the setups that include sentiment classification as one of the tasks, i.e., (S_C, E_C) and (S_C, S_I) MTL framework. The (S_C, E_C) framework yields an F1-score and accuracy value of 78.8% and 80.5% with an improvement of 1.2% and 0.7% over the STL framework, respectively. We also observe that the multi-task framework reports better score for the uni-modal and bi-modal input combinations along with the tri-modal input combination. Similarly to the (S_C, E_C), we obtain the performance improvement for all the input combinations in the (S_C, S_I) MTL framework as well.

In Table 3(b), we report the performance of our proposed single-task and multi-task approaches for emotion classification. To obtain multi-labels for emotion classification, we set the *threshold* as 0.4 & 0.2 for F1-score and weighted accuracy, respectively. We obtain 77.71% F1-score, and 60.88%

Table 3. STL and MTL Frameworks for the Proposed Approach

Tasks		F1-score							Accuracy						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	S_C	75.1	67.9	66.3	77.0	76.5	69.6	77.6	78.2	74.8	75.8	79.4	79.7	76.6	79.8
MTL	S_C, E_C	77.5	72.1	69.1	78.7	78.6	75.8	78.8	79.7	75.7	76.5	80.4	80.2	77.4	80.5
	S_C, S_I	77.6	72.5	69.3	78.6	78.6	75.9	78.9	79.9	76.5	76.4	80.0	80.0	77.9	80.1

(a) Sentiment Classification (S_C)

Tasks		F1-score							Weighted-Accuracy						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	E_C	75.9	72.3	73.6	77.5	76.8	76.0	77.7	58.0	56.7	53.7	60.1	59.6	58.0	60.8
MTL	S_C, E_C	76.9	74.6	75.4	78.5	77.6	77.0	78.6	60.2	56.2	57.5	62.5	60.5	59.3	62.8
	E_C, E_I	78.0	74.8	77.1	79.1	78.2	78.1	79.2	61.5	58.3	56.7	63.0	61.6	60.3	63.3

(b) Emotion Classification (E_C)

Tasks		Pearson Correlation							Cosine similarity						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	S_I	0.544	0.416	0.349	0.557	0.549	0.436	0.559	0.553	0.424	0.364	0.567	0.561	0.451	0.568
MTL	S_C, S_I	0.554	0.421	0.365	0.566	0.558	0.456	0.568	0.560	0.431	0.382	0.574	0.566	0.469	0.576

Tasks		MSE							MAE						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	S_I	0.871	1.045	1.088	0.854	0.860	0.999	0.848	0.702	0.793	0.799	0.699	0.700	0.768	0.699
MTL	S_C, S_I	0.863	1.032	1.072	0.841	0.853	0.978	0.838	0.701	0.791	0.792	0.698	0.698	0.762	0.697

(c) Sentiment Intensity Prediction (S_I)

Tasks		Pearson Correlation							Cosine similarity						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	E_I	0.249	0.165	0.134	0.260	0.219	0.228	0.298	0.436	0.334	0.281	0.437	0.422	0.405	0.464
MTL	E_C, E_I	0.259	0.180	0.154	0.269	0.277	0.235	0.323	0.439	0.367	0.301	0.448	0.462	0.419	0.478

Tasks		MSE							MAE						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
STL	E_I	0.136	0.128	0.126	0.108	0.117	0.112	0.104	0.163	0.167	0.164	0.159	0.165	0.161	0.162
MTL	E_C, E_I	0.119	0.120	0.118	0.098	0.114	0.105	0.094	0.175	0.170	0.171	0.166	0.173	0.162	0.169

(d) Emotion Intensity Prediction (E_I)

weighted accuracy for emotion classification in the single-task framework. In comparison, similar to the sentiment classification, when the emotion classification task is learned and evaluated in the multi-task frameworks along with the sentiment classification (i.e., S_C, E_C) and the emotion intensity prediction (i.e., E_C, E_I), we observe a performance increase in both the F1-score and weighted-accuracy of the emotion classification task. We obtain an improvement of 1% in F1-score and 2% in weighted-accuracy for the (S_C, E_C) multi-task framework, whereas for the (E_C, E_I) framework, we observe improvements of 1.5% and 2.5% in F1-score and weighted-accuracy, respectively.

For sentiment and emotion intensity prediction tasks, we report Pearson correlation score, cosine similarity, mean-squared-error and mean-absolute-error as the performance metrics in Table 3(c) and 3(d). Similar to the sentiment classification (c.f. Table 3(a)) and emotion classification (c.f. Table 3(b)) tasks, we observe the performance improvement for sentiment intensity and emotion intensity tasks in multi-task learning framework as well. Both the Pearson score (0.568, 0.323) and cosine similarity (0.576, 0.478) for the MTL frameworks (i.e., (S_C, S_I), (E_C, E_I)) are better than their respective single-task frameworks. Similarly, MSE and MAE for the sentiment intensity prediction are lower in comparison to the single-task framework. For emotion intensity prediction, our proposed MTL framework (E_C, E_I) yields lower MSE than the single-task framework; however, the system reports lower MAE for STL.

Further, for most of the above scenarios, the performance of a tri-modal input combination is better than the performance of bi-modal inputs, which are, in turn, better than the performance of uni-modal inputs. Thus, it signifies that the effective combination of multi-modal inputs, obtained by the contextual inter-modal attention mechanism, has a significant effect on the performance of the underlying system. Also, these results suggest that the MTL framework successfully leverages the inter-dependence of multiple tasks in improving the overall performance. In Figure 3, we present a graphical comparison among the STL and MTL frameworks for different input combinations.

To verify the efficacy of our proposed model, we also experiment with another multi-modal dataset, i.e., CMU-MOSI [58]. The CMU-MOSI dataset has approximately 2,199 utterances, and each utterance has an associated sentiment score. We follow our (S_C, S_I) MTL framework (c.f. Section 4.3) for the prediction of sentiment class and intensity values in the MOSI dataset. In Table 4, we depict the evaluation results for tri-modal input. The STL framework for the sentiment classification task obtains an F1-score of 76.67% and an accuracy value of 76.74%. In comparison, the proposed model yields an improved F1-score (78.24%) and accuracy value (78.42%) for the sentiment classification task in the MTL framework. Similarly, we observe better performance when the sentiment intensity prediction task is learned alongside sentiment classification in the MTL framework. The performance of the MTL framework in MOSI dataset (as well as in MOSEI dataset) further supports our claim that the inter-relatedness among the participating tasks indeed assist each other for performance improvement.

4.5 Qualitative Analysis of the Single-task and Multi-task Learning Frameworks

As discussed in Section 1, MTL framework aims to leverage the inter-relatedness of multiple tasks for individual performance improvement. Through experiments, we establish the efficacy of MTL framework for the four tasks (c.f. Table 3 and Figure 3). We further analyze the outputs of our proposed STL and MTL frameworks from the qualitative perspective. In Table 5, we present few example scenarios, where the tasks in our proposed MTL framework exploit the information of other tasks for the correct classification (or better intensity score prediction). In comparison, the single-task framework finds it non-trivial for correct prediction of the same instances. For the four tasks at hand, we compare the predictions of all three multi-task setups (c.f. Section 4.3) with the single-task framework. For the first example (i.e., u_1) in Table 5, the actual sentiment is *negative* with the intensity −2.67, whereas the emotions are *anger* and *disgust* with intensities 0.66 and 1.0, respectively. In the STL framework, the sentiment classification model (S_C) misclassifies the sentiment as *positive*, while the sentiment intensity prediction obtains the score of −0.11 (an error of 2.56 points). In comparison, the sentiment classification and intensity prediction (S_C, S_I) MTL framework correctly predicts the sentiment polarity to *positive*, and also improves the intensity prediction by 1.25 points at −1.35. Thus, we can argue that the MTL framework, indeed, learns better than the STL framework, as these two tasks assist each other for the performance improvement in the MTL setup.

Fig. 3. Comparative study of STL and MTL frameworks. Accuracy, Weighted-Accuracy, F1-score, Pearson, Cosine: Higher the better; MSE, MAE: Lower the better.

Similarly, the single task emotion classification model obtains one correct (i.e., disgust) and two incorrect predictions (i.e., *happy* and *sad*). Furthermore, it also fails to identify the *anger* emotion. Comparatively, both the MTL setups involving emotion classification, i.e., (E_C, E_I) and (S_C, E_C), yield improved performance (i.e., $precision = 0.66$ & $recall = 1.0$ for MTL (E_C, E_I) and

Table 4. MOSI: STL and MTL Frameworks for the Proposed Approach on the Tri-modal Inputs (T+A+V)

Tasks	Sentiment Class		Sentiment Intensity			
	F1-score	Acc	MAE	MSE	Pearson	Cosine
STL	76.67	76.74	0.874	1.236	0.716	0.727
MTL	78.24	78.42	0.867	1.218	0.719	0.731

$precision = 0.5$ & $recall = 1.0$ for MTL (S_C, E_C)) than the single-task emotion classification system ($precision = 0.33$ & $recall = 0.5$). Similarly, for the emotion intensity prediction, the MTL setup (E_C, E_I) obtains better score for the two actual emotion classes *anger* and *disgust*. Further, we observe that the two tasks in MTL (E_C, E_I) are in synchronization most of the time, i.e., for all the predicted emotion classes (e.g., *anger*, *disgust* and *happy* for u_1), the system yields better intensity scores for all the predicted classes against the non-predicted classes (e.g., *fear*, *sad* and *surprise*). It can also be observed in other utterances, where the MTL (E_C, E_I) setup clearly differentiates between the predicted and non-predicted emotion classes in intensity prediction as well (e.g., in utterance u_2, it yields intensity score of 0.93 for the predicted *happy* emotion and ∼0 intensity scores for the other emotion classes). These observations suggest that the correct emotion classification has a direct effect on the improved intensity prediction.

For the sentiment and emotion classification (S_C, E_C) MTL setup, the knowledge of sentiment helps in identifying the correct emotion label in MTL framework. For example, in utterance u_3, the presence of *negative* sentiment drives the system to ignore the *happy* emotion for the final prediction, which was predicted by both STL and emotion classification and emotion intensity prediction (E_C, E_I) MTL frameworks in the absence of information regarding the *negative* sentiment. It suggests that our MTL framework identifies the relationship between sentiment and emotion, and leverages the predicted sentiment for the correct classification of emotion. Once again, we argue that this is an example of inter-dependence between two related tasks and our MTL framework successfully exploits it for the correct prediction.

4.6 Comparative Analysis

In Table 6, we report the comparative results for all four tasks, i.e., sentiment classification (S_C), sentiment intensity prediction (S_I), emotion classification (E_C), and emotion intensity prediction (E_I). In particular, we compare our proposed single-task and multi-task approaches with the following existing systems:

A. **Sheikh et al. [44]** proposed a deep canonical correlation analyzer to focus on improving the representations of the input modalities (i.e., text and acoustic). These improved representations are fed to the softmax classifier for the sentiment classification.
B. **Blanchard et al. [6]** proposed a multi-modal fusion technique that accounts for visual and acoustics input modalities. At first, they extracted mid-level acoustic and visual features from trained bag-of-words (BoW) models and subsequently learned the confidence scores using a separate classifier (SVM) for each modality. Finally, they employed two fusion techniques, i.e., score-level fusion and output-level fusion, for the final prediction.
C. **Zadeh et al. [55]** proposed a MFN that explicitly accounts for both view-specific interactions and cross-view interactions. Authors employed LSTM for the view-specific interaction, whereas for the cross-view interaction, an attention mechanism had been proposed.
D. **Nojavanasghari et al. [27]** employed a deep feed-forward neural network (DNN) for combining the three input modalities. Initially, they computed a confidence score for each input

Table 5. Qualitative Analysis of the STL and MTL Frameworks

	Utterances		Sentiment		Emotion						
			Class	Intensity	Class	Intensity					
						An	Dg	Fr	Ha	Sd	Sr
u_1	*richard gere and susan umm you i really didn't enjoy this movie at all it kinda boring for*	Actual	**Neg**	**−2.67**	**An**, **Dg**	**0.66**	**1.00**	0	0	0	0
		STL	*Pos*	*−0.11*	Dg, *Hp*, *Sd*	*0.26*	*0.15*	*0*	*0.39*	*0.15*	*0.03*
		MTL (E_C, E_I)	-	-	An, Dg, *Hp*	***0.39***	***0.32***	*0.03*	*0.53*	*0.26*	*0.12*
		MTL (S_C, S_I)	Neg	*−1.35*	-	-					
		MTL (S_C, E_C)	Neg	-	An, Dg, *Hp*, *Sd*	-					
u_2	*laughter and applause still there though..*	Actual	**Pos**	**0.66**	**Hp**	0	0	0	**1.33**	0	0
		STL	*Neg*	*−0.16*	Hp, *Sr*	*0.10*	*0.03*	*0.03*	*0.79*	*0.11*	*0.23*
		MTL (E_C, E_I)	-	-	Hp	*0.07*	*0.06*	*0.03*	***0.93***	*0.07*	*0.04*
		MTL (S_C, S_I)	Pos	*0.23*	-	-					
		MTL (S_C, E_C)	Pos	-	Hp	-					
u_3	*is in love with some other person so you know the story*	Actual	**Neg**	**−2.66**	**An**, **Dg**, **Sd**	**1.33**	**0.33**	0	0	**1.00**	0
		STL	*Pos*	*−0.49*	Dg, *Hp*, Sd	*0.12*	*0.10*	*0*	*0.15*	*0.10*	*0*
		MTL (E_C, E_I)	-	-	An, Dg, *Hp*, Sd	***0.61***	***0.32***	*0.04*	*0.20*	***0.30***	*0.03*
		MTL (S_C, S_I)	Neg	*−0.85*	-	-					
		MTL (S_C, E_C)	Neg	-	An, Dg, Sd	-					
u_4	*i can say unfortunately i don't think it's a serious program*	Actual	**Neg**	**−1.00**	**Dg**, **Sd**, **Sr**	0	**0.33**	0	0	**0.33**	**0.33**
		STL	*Pos*	*−0.06*	*An*, *Hp*, Sd	*0.07*	*0.01*	*0*	*0.23*	*0.09*	*0.01*
		MTL (E_C, E_I)	-	-	Dg, *Hp*, Sd	*0.13*	***0.29***	*0.05*	*0.28*	***0.31***	***0.07***
		MTL (S_C, S_I)	Neg	*−0.36*	-	-					
		MTL (S_C, E_C)	Neg	-	*An*, Dg, *Hp*, Sd	-					
u_5	*the last administration bought into just as much as this one does unfortunately*	Actual	**Neg**	**−1.33**	**An**, **Dg**, **Sd**	**0.33**	**0.33**	0	0	**1.00**	0
		STL	*Pos*	*0.34*	An, *Hp*	*0.04*	*0.01*	*0*	*0.35*	*0.01*	*0.01*
		MTL (E_C, E_I)	-	-	An, Dg, Sd	***0.18***	***0.28***	*0*	*0.09*	***0.13***	*0*
		MTL (S_C, S_I)	Neg	*−0.52*	-	-					
		MTL (S_C, E_C)	Neg	-	An, Dg, *Hp*, Sd	-					
u_6	*it's just too great of a risk and it is socially unacceptable*	Actual	**Neg**	**−0.33**	**An**, **Dg**, **Hp**	**1.00**	**0.33**	0	**0.33**	0	0
		STL	*Pos*	*0.03*	An, Hp	*0.24*	*0.05*	*0.04*	*0.27*	*0.12*	*0.03*
		MTL (E_C, E_I)	-	-	An, Dg, Hp, *Sd*	***0.32***	*0.10*	*0.07*	***0.28***	*0.12*	*0.05*
		MTL (S_C, S_I)	Neg	*−0.21*	-	-					
		MTL (S_C, E_C)	Neg	-	An, Dg, Hp	-					
u_7	*had a robot here at hopkins since the year longer than most institutions in this country and around the world*	Actual	**Pos**	**0.33**	**Hp**	0	0	0	**0.33**	0	0
		STL	Pos	*0.81*	Hp, *Sd*	*0.02*	*0.01*	*0*	*0.17*	*0.03*	*0.01*
		MTL (E_C, E_I)	-	-	Hp	*0*	*0*	*0*	***0.20***	*0.06*	*0*
		MTL (S_C, S_I)	Pos	*0.72*	-	-					
		MTL (S_C, E_C)	Pos	-	Hp	-					

Few error cases where multi-task learning framework performs better than the single-task framework. **An:** Anger, **Dg:** Disgust, **Fr:** Fear, **Hp:** Happy, **Sd:** Sad and **Sr:** Surprise. The *red colored text* shows error in classification, while the *blue colored text* reflects predicted intensity values.

Table 6. **Comparative Results:** Proposed Multi-task Framework Attains Better Performance as Compared to the State-of-the-Art (SOTA) Systems in Both the Tasks i.e., Emotion Recognition (Average) and Sentiment Analysis

			Exsisting Systems														Proposed	
			A [44]	B [6]	C [55]⋆	D [27]⋆	E [39]⋆	F [54]⋆	G [58]⋆	H [57]⋆	I [56]	J [50]	K [5]	L [34]	M [35]		STL	MTL
S_C	Pos	*F1*	-	-	-	-	-	-	-	-	-	-	-	-	-		94.2	**95.5**
		Acc	-	-	-	-	-	-	-	-	-	-	-	-	-		90.6	**91.6**
	Neg	*F1*	-	-	-	-	-	-	-	-	-	-	-	-	-		61.0	**62.1**
		Acc	-	-	-	-	-	-	-	-	-	-	-	-	-		69.0	**69.4**
	Avg	*F1*	*63.3*	*63.2*	76.0	-	76.4	-	-	-	77.0	65.6	-	76.1	75.8		77.6	**78.8**
		Acc	*63.4*	*60.0*	76.0	-	76.4	-	-	-	76.9	74.0	-	77.6	76.1		79.8	**80.5**
S_I	Pos	PEAR	-	-	-	-	-	-	-	-	-	-	-	-	-		**0.30**	0.28
		MAE	-	-	-	-	-	-	-	-	-	-	-	-	-		**0.59**	0.62
	Neg	PEAR	-	-	-	-	-	-	-	-	-	-	-	-	-		0.80	**0.84**
		MAE	-	-	-	-	-	-	-	-	-	-	-	-	-		0.79	**0.76**
	Avg	PEAR	-	0.30	-	-	-	-	-	-	0.54	-	-	-	21.8		0.55	**0.56**
		MAE	-	0.91	-	-	-	-	-	-	0.71	-	-	-	1.46		**0.69**	**0.69**
E_C	An	*F1*	-	-	-	71.4	-	-	-	-	72.8	-	-	-	72.1		75.6	**75.9**
		W-Acc	-	-	-	-	56.0	60.5	-	-	62.6	-	-	-	49.8		64.5	**66.8**
	Dg	*F1*	-	-	71.4	-	-	-	-	-	76.6	-	-	-	73.2		81.0	**81.9**
		W-Acc	-	-	65.2	67.0	-	-	-	-	69.1	-	-	-	49.9		72.2	**72.7**
	Fr	*F1*	-	-	**89.9**	-	-	-	-	-	**89.9**	-	-	-	94.2		87.7	87.9
		W-Acc	-	-	-	-	-	-	60.0	-	62.0	-	-	-	49.9		51.5	**62.2**
	Hp	*F1*	-	-	-	-	-	66.6	-	**71.0**	66.3	-	-	-	23.7		59.3	67.0
		W-Acc	-	-	-	-	-	66.5	-	-	66.3	-	-	-	37.3		**61.6**	53.6
	Sd	*F1*	-	-	60.8	-	-	-	-	-	66.9	-	-	-	51.6		67.3	**72.4**
		W-Acc	-	-	-	-	-	58.9	-	-	60.4	-	-	-	49.8		**65.4**	61.4
	Sr	*F1*	-	-	85.4	-	-	-	-	-	85.5	-	-	-	83.0		**86.5**	86.0
		W-Acc	-	-	53.3	-	-	52.2	-	-	53.7	-	-	-	49.8		53.0	**60.6**
	Avg	*F1*	-	-	-	-	-	-	-	-	76.3	-	-	-	66.3		76.2	**78.6**
		W-Acc	-	-	-	-	-	-	-	-	62.3	-	57.6	-	47.7		61.3	**62.8**
E_I	An	PEAR	-	-	-	-	-	-	-	-	-	0.08	-	−0.13	−0.004		0.34	**0.40**
		MAE	-	-	-	-	-	-	-	-	-	**0.10**	-	0.18	1.35		0.17	0.18
	Dg	PEAR	-	-	-	-	-	-	-	-	-	0.06	-	−0.10	0.004		0.39	**0.41**
		MAE	-	-	-	-	-	-	-	-	-	**0.05**	-	0.13	1.39		0.12	0.14
	Fr	PEAR	-	-	-	-	-	-	-	-	-	0.01	-	−0.03	−0.002		0.07	**0.10**
		MAE	-	-	-	-	-	-	-	-	-	**0.05**	-	0.06	1.46		**0.05**	0.06
	Hp	PEAR	-	-	-	-	-	-	-	-	-	0.55	-	0.58	0.02		0.58	**0.60**
		MAE	-	-	-	-	-	-	-	-	-	0.40	-	0.38	1.10		**0.36**	0.37
	Sd	PEAR	-	-	-	-	-	-	-	-	-	−0.06	-	−0.16	−0.013		0.28	**0.29**
		MAE	-	-	-	-	-	-	-	-	-	**0.11**	-	0.17	1.36		0.18	0.19
	Sr	PEAR	-	-	-	-	-	-	-	-	-	−0.03	-	−0.007	−0.013		0.10	**0.13**
		MAE	-	-	-	-	-	-	-	-	-	**0.03**	-	0.06	1.46		0.05	0.07
	Avg	PEAR	-	-	-	-	-	-	-	-	-	-	-	0.27	0.02		0.29	**0.32**
		MAE	-	-	-	-	-	-	-	-	-	**0.12**	0.87	0.19	1.35		0.16	0.16

⋆Values are taken from system **I** [56]. Columns (A–M) are the existing systems defined in Section 4.6.

modality using another DNN. These confidence scores (C_i), along with their complement scores ($1 - C_i$), act as inputs to the fusion network for the prediction.

E. **Rajagopalan et al. [39]** proposed a Multi-View LSTM (MV-LSTM) to model the view-specific and cross-view interactions over time.
F. **Zadeh et al. [54]** introduced a multi-modal TFN framework to learn both intra and inter-dynamics of the input modalities (i.e., text, acoustic, and visual).
G. **Zadeh et al. [58]** proposed a dictionary-based approach for multi-modal sentiment analysis. The multi-modal dictionary was compiled utilizing the verbal and gesture features for each word in the dataset.
H. **Zadeh et al. [57]** introduced a multi-attention block (MAB) based framework for sentiment classification to capture the inter-modality information across modalities.
I. **Zadeh et al. [56]** proposed a DFG for the fusion of tri-modal inputs. The authors extended the MFN [55] by incorporating the DFG (called as Graph-MFN) for the fusion.
J. **Williams et al. [50]** proposed an input-level fusion technique followed by a deep neural network layer (i.e., CNN, LSTM and GRU) to combine the three modalities for the emotion intensity prediction.
K. **Beard et al. [5]** proposed a recursive multi-attention architecture that exploits the shared external memory for emotion recognition.
L. **Poria et al. [34]** exploited the contextual dependencies among the utterances in an LSTM architecture for the sentiment classification.
M. **Poria et al. [35]** proposed an attention-based recurrent model that incorporates both context learning and dynamic feature fusion for sentiment analysis.

Some of the above existing works reported the performance[7] on the MOSEI dataset, while others evaluated their systems on different datasets, for which we have executed their models[8] and obtained the values. Further, we take the results of some systems, as reported in Zadeh et al. [56] (System I). For sentiment classification (S_C), the existing state-of-the-art (i.e., System I [56]) reported the F1-score and accuracy values as 77.0% and 76.9%. In comparison, our proposed MTL frameworks yield 78.8% F1-score and 80.5% accuracy value with an increment of 1.8 and 3.6 points, respectively, for the sentiment classification. In the sentiment intensity prediction (S_I) task, our MTL framework reports 0.56 Pearson score against 0.54 Pearson score of the state-of-the-art system. Similarly, we obtain lesser error (0.69 MAE) in comparison to the existing system (0.71 MAE).

For emotion classification (E_C), we report both the overall and class-wise performance of the system in terms of obtained F1-scores and weighted-accuracies. On average, our MTL framework reports 79.2% F1-score and 63.5% weighted-accuracy, whereas the best existing system yields 76.3% and 62.3% F1-score and weighted-accuracy, respectively. Furthermore, our system also performs better for most of the individual emotion classes (except F1-score for *happy* and *fear*). In the emotion intensity prediction (E_I) task, we observe contrasting behavior among two evaluation metrics. While our proposed MTL framework yields a better Pearson score for all the cases, the existing system J [50] obtains lesser error in terms of mean-absolute-error. We obtained the overall Pearson score of 0.32 and MAE of 0.16, while the existing systems obtained the best Pearson score of 0.27 (i.e., System L [34]) and MAE of 0.12 (i.e., System J [50]).

During analysis, we make an important observation. Small improvements in performance do not reveal the exact improvement in the number of instances. Since there are more than 4.6K test samples, even the improvement by one point (in the classification tasks) reflects that the system improves its predictions for 46 samples.

[7]Some of these systems do not report their performances for all the four tasks.
[8]We have experimented with their published codes.

We also perform a significance test (*T-test*) to compute the *p*-values for all four tasks and observe that the obtained results are significant (*p*-values > 0.5) for most of the cases (except MAE for the *emotion intensity*).

In our current work, we have considered only two tasks at a time in an MTL framework; however, we would like to emphasize that the MTL framework can be scaled for n number of related-tasks. For each task, we can add task-specific layers for the prediction on top of the shared representation, i.e., the shared representation can be fed to each task-specific branch of the framework for the respective prediction.

Please note that to provide an unbiased comparison between the proposed MTL and STL frameworks, we kept the same network configuration for both, except the task-specific layers in the MTL framework. Therefore, in comparison with any STL framework, the only addition in the MTL framework is the task-specific layer for $n-1$ tasks[9] (in our case, the number of tasks $n = 2$). We also kept the network hyperparameters (e.g., loss, optimizer, dropout, batch size and epoch) intact, as mentioned in Table 2.

Let us assume that the whole network is divided into the following two modules, i.e., a base network B that is common in both MTL and STL frameworks (starting from the input layer to the concatenations of the CIM representation) and a task-specific network T_i for the i^{th} task. Therefore, the time complexity to train the complete network for n tasks is

$$\text{STL} = \sum_{i}^{n} O(e * \alpha(B + T_i)) \tag{1}$$

$$\text{MTL} = O\left(e * \left[\alpha(B) + \sum_{i}^{n} \alpha(T_i)\right]\right) \tag{2}$$

where $\alpha()$ is the function that computes the time required for one forward-backward pass of a network, and e is the number of epochs. We observe from Equations (1) and (2) that during the training of MTL framework for n tasks, the base network B has been trained only once; however, for the STL framework, the base network B needs to be trained n times for each task. Therefore, the overall complexity of the STL framework for n tasks is higher than the MTL framework. Further, assuming that the task-specific network T_i is small enough (say, just the output layer), we can achieve an improvement in the overall complexity by an approximate factor of n.

4.7 Analysis of Attention Mechanism

In this section, we present our analysis of the proposed contextual inter-modal attention mechanism. For the case study, we select a representative video from the dataset, as depicted in Figure 4.

There are seven utterances in the video, and for each case, we depict three representational visual-frames along with their textual representations. The person in the video is sharing his experience, and at times he feels anger/sadness remembering the bitter events. However, in the present, he seems confident and satisfied with all the things that he learned. In Table 7, we list the predictions for the seven utterances in our three multi-task setups. Further, the heatmaps of the attention weights (i.e., N_1 & N_2 of Algorithm 1) are depicted in Figure 5.

For each multi-task setup, there are three attention weight matrices of $2 \times (7 \times 7)$ dimension, corresponding to the *text-visual*, *acoustic-visual*, and *text-acoustic* pairs. The solid black line, at the center, defines the boundary of participating modalities, i.e., the left and right sides represent the *textual* (N_1)[10] and *visual* (N_2)[10] representations, respectively, for the *text-visual* attention pair.

[9]Task-specific layers for one task will also be there in any STL framework.

[10]Probability distribution matrix as computed in the CIM attention framework (c.f. Section 3.1 and Algorithm 1).

(a) u_1: *they have doubt they have fear they lose in a ferrari race or they lose in a race and then they just give up* [Neg; Sad]

(b) u_2: *but the hardest thing in life to learn is to lose easy* [Pos; Disgust, Happy, Sad]

(c) u_3: *you win you're happy you get a high five your friends are happy* [Pos; Happy]

(d) u_4: *it's your ability to take a loss get up the next day dust yourself off and keep going* [Pos; Happy]

(e) u_5: *i have lost at everything you can imagine seriously* [Pos; Happy]

(f) u_6: *i have been beat up and put down and everything that can go wrong has gone wrong for me* [Neg; Sad]

(g) u_7: *but i just kept going* [Pos; No class]

Fig. 4. We take a video from MOSEI dataset which has 7 utterances $[u_1, u_2, u_3, u_4, u_5, u_6 \ \& \ u_7]$. We depict three representative visual frames for each utterance with their actual sentiments and emotions.

Table 7. Prediction for the Utterances of a Video in Test Set for Various Multi-task Frameworks

		S_C	E_C
u_1	Actual	Neg	Sd
	MTL	*Neg*	*An, Dg, Hp, Sd*
u_2	Actual	Pos	Dg, Hp, Sd
	MTL	*Pos*	*An, Hp, Sd*
u_3	Actual	Pos	Hp
	MTL	*Pos*	*Hp, Sd*
u_4	Actual	Pos	Hp
	MTL	*Pos*	*Hp*
u_5	Actual	Pos	Hp
	MTL	*Pos*	*Hp*
u_6	Actual	Neg	Sd
	MTL	*Neg*	*An, Hp, Sd*
u_7	Actual	Pos	No class
	MTL	*Pos*	*Hp, Sd*

(a) S_C, E_C

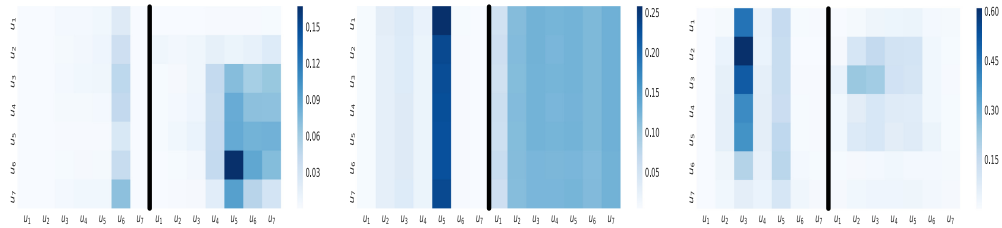
S_C	S_I
Neg	−0.33
Neg	*−0.18*
Pos	0.66
Pos	*0.21*
Pos	1.32
Pos	*0.33*
Pos	0.66
Pos	*2.40*
Pos	0
Pos	*1.80*
Neg	−1.98
Neg	*−1.20*
Pos	0
Pos	*0.48*

(b) S_C, S_I

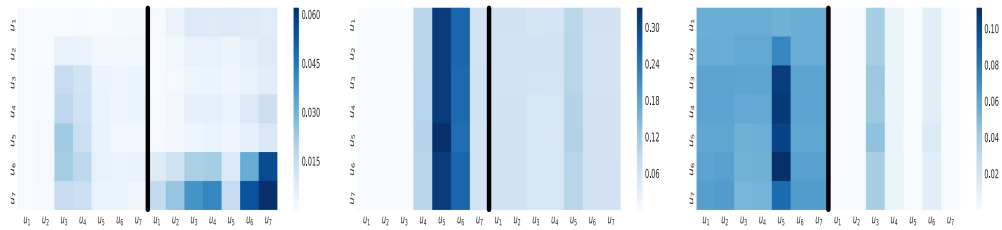
E_C	E_I					
	An	**Dg**	**Fr**	**Hp**	**Sd**	**Sr**
Sd	0	0	0	0	0.33	0
An, Dg, Hp, Sd	*0.15*	*0.09*	*0.06*	*0.21*	*0.36*	*0.06*
Dg, Hp, Sd	0	0.33	0	0.66	0.66	0
Hp, Sd	*0.09*	*0.33*	*0.06*	*0.42*	*0.51*	*0.06*
Hp	0	0	0	1.66	0	0
Hp	*0.09*	*0.03*	*0.03*	*0.45*	*0.12*	*0.03*
Hp	0	0	0	1	0	0
Hp	*0.04*	*0.09*	*0.06*	*0.51*	*0.21*	*0.09*
Hp	0	0	0	1	0	0
Hp, Sd	*0.15*	*0.09*	*0.06*	*0.67*	*0.18*	*0.15*
Sd	0	0	0	0	1.66	0
Hp, Sd	*0.21*	*0.12*	*0.06*	*0.69*	*0.78*	*0.24*
No class	0	0	0	0	0	0
Hp	*0.15*	*0.09*	*0.03*	*0.63*	*0.09*	*0.15*

(c) E_C, E_I

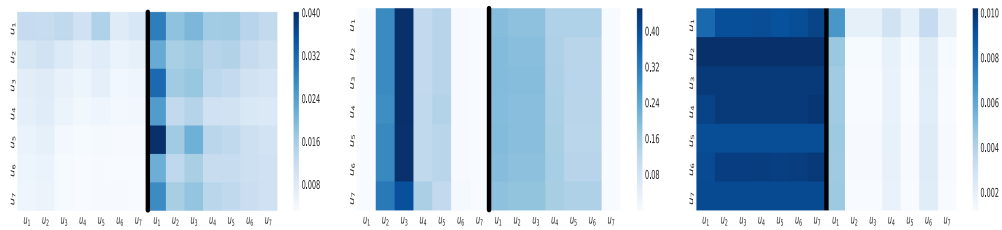
In Figure 5, we present the heatmaps of the attention weights for our CIM module.



(a) *Sentiment and Emotion Classification* (S_C, E_C).



(b) *Sentiment Classification and Sentiment Intensity Prediction* (S_C, S_I).



(c) *Emotion Classification and Emotion Intensity Prediction* (E_C, E_I).

Fig. 5. Pair-wise softmax attention weights N_1 and N_2 of *text-visual*, *acoustic-visual* and *text-acoustic* for a video (c.f. Figure 4 and Table 7) in three MTL setups. Solid (black) line at the center represents boundary of N_1 and N_2 matrices. The heatmaps represent attention weights of a particular utterance with respect to other utterances in N_1 and N_2. Each cell (i, j) of the heatmap signifies the weights of utterance "j" for the classification of utterance "i" of the pair-wise modality matrices, hence, assists in predicting the labels more concisely by incorporating contextual inter-modal information. Depending on the set of tasks in the MTL framework, the CIM module learns different set of attention weights for each case, hence, it suggests that different set of features contributes differently.

Each matrix depicts the associations of an utterance with all the other utterances in the video, with each cell (i, j) signifies the degree of association of utterance "j" for the classification of utterance "i" in the pair-wise modality matrices. The darker shade represents the higher association. As an instance, for the utterance "u_6" in Figure 5(a), our model ignores the *textual* features of other utterances, and at the same time, it puts more attention on the visual features of u_5, u_6 and u_7 utterances for the *textual-visual* attention pair. We observe that the model predicts the sentiment of u_6 correctly, which we can relate to the *textual* representation. Similarly, for the emotion classification, we can argue that the model focuses on the textual features for the emotions *sad* (correct) and

anger (incorrect); however, the misclassified class *happy* was predicted due to the *visual* representations of the utterance u_5, u_6 or possibly u_7 as well. Therefore, in light of the above discussion, it might not be unfair to say that the attention weights assist in predicting the labels more concisely by incorporating contextual inter-modal information. Furthermore, depending on the set of tasks in the MTL framework, the CIM module learns a different set of attention weights for each case; hence, it suggests that different set of features contributes differently towards the prediction of different outputs.

5 CONCLUSION AND FUTURE DIRECTION

In this article, we have presented a recurrent architecture based MTL framework for the multi-modal affect analysis. We employed a contextual inter-modal attention mechanism to learn the degree of association among neighboring utterances and the input modalities (i.e., textual, acoustic, and visual). We defined three multi-task setups for the four tasks of sentiment and emotion analysis, i.e., sentiment and emotion classification, sentiment classification and sentiment intensity prediction, and emotion classification and emotion intensity prediction. Our proposed multi-task models learn the inter-relatedness among the participating tasks and leverage them for the overall performance improvement. We evaluated our proposed approach on the benchmark dataset of multi-modal sentiment and emotion analysis, i.e., CMU-MOSEI. Experimental results suggest that, in all three setups, the MTL framework obtains better performance than that of a STL framework. Further, we compared our proposed model with various existing systems, and in comparison, we report state-of-the-art for three tasks, sentiment classification, sentiment intensity prediction and emotion classification, while for the emotion intensity prediction, we obtained comparable performance (better Pearson correlation, but slightly higher mean-absolute-error).

The significance of our obtained results are followings:

- Our MTL framework has effectively utilized the inter-relatedness among the participating tasks. Our system leverages the inter-relatedness information of each task, and improves the performance. The inferior performance of the STL framework verifies that the inter-relatedness information has a positive effect on the proposed MTL framework.
- The overall complexity to solve multiple tasks is reduced, as only a single system is required for all the participating tasks. Therefore, the obtained results provide benefits on both fronts of performance and complexity.
- Our proposed attention mechanism intelligently selects the contributing input modality based on its significance for the prediction. It enables us not to worry about the presence of noise in the input as the noise gets filtered in the attention module.

In our experiments for the multi-label emotion classification, we have manually fixed threshold value for the final prediction, i.e., predicted values higher than the threshold represent the presence of emotion, while values less than the threshold reflect the absence of respective emotion. There are two drawbacks of such an approach:

- The threshold has been set manually to maximize the performance.
- The optimized threshold value differs for the different objective functions (i.e., 0.4 for the F1-score and 0.2 for the Weighted-Accuracy in Emotion classification, c.f. Table 2).

A possible solution to the above limitations could be the application of a *multi-objective optimization technique* to optimize the threshold for various objective functions simultaneously. Such an approach would also ensure that the threshold is found automatically as a result of the optimization.

Our multi-task framework shares the concatenated representation up to the attention layer. The shared representation receives gradients of errors from all the branches of the multiple tasks (e.g., sentiment and emotion) and accordingly adjust the weights of the model. Thus, the shared representations do not pose any bias towards any particular task, and it will assist the model to achieve generalization for multiple tasks.

Since the shared representation aims to achieve the generalization, not all these attentive representations are equally important to both sentiment and emotion. In other words, some of these representations might be more significant than others for sentiment classification, whereas the same might be less significant. A potential extension of the current study is to filter the shared-representation to fulfill the task-specific requirements. In future, we would like to explore both the dimensions.

ACKNOWLEDGMENT

Authors duly acknowledge the partial supports of "Sevak-An Intelligent Indian Language Chatbot," Sponsored by SERB, Govt. of India (IMP/2018/002072), and Skymap Global Private Limited. Asif Ekbal gratefully acknowledges Young Faculty Research Fellowship (YFRF), 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).

REFERENCES

[1] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN*. 370–379.

[2] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore sound classification using image-based deep spectrum features. In *Proceedings of the 2017 Interspeech, Stockholm, Sweden*. 3512–3516.

[3] Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask learning for fine-grained twitter sentiment analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan*. 1005–1008.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY*. 1–10.

[5] Rory Beard, Ritwik Das, Raymond W. M. Ng, P. G. Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. 2018. Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium*. 251–259.

[6] Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer. 2018. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language, Melbourne, Australia*. 1–10.

[7] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (Oct 2001), 5–32.

[8] Erik Cambria, Soujanya Poria, and Amir Hussain. 2019. Speaker-independent multimodal sentiment analysis for big data. In *Multimodal Analytics for Next-Generation Big Data Technologies and Applications*. Springer, 13–43.

[9] Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. 2019. Embodied multimodal multitask learning. *CoRR* abs/1902.01385 (2019). arxiv:1902.01385

[10] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Hong Kong, China, 5651–5661.

[11] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK*. 163–171.

[12] S. Chowdhuri, T. Pankaj, and K. Zipser. 2019. MultiNet: Multi-Modal multi-task learning for autonomous driving. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV'19)*. 1496–1504.

[13] Nicholas Cummins, Shahin Amiriparian, Sandra Ottl, Maurice Gerczuk, Maximilian Schmitt, and Björn Schuller. 2018. Multimodal bag-of-words for cross domains sentiment analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, Canada*. 1–5.

[14] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy*. 960–964.

[15] Didan Deng, Yuqian Zhou, Jimin Pi, and Bertram E. Shi. 2018. Multimodal utterance-level affect analysis using visual, audio and text features. *arXiv preprint arXiv:1805.00625* (2018).

[16] Jan Milan Deriu and Mark Cieliebak. 2016. Sentiment analysis using convolutional neural networks with multi-task training and distant supervision on Italian tweets. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, December 5–7, 2016, Napoli, Italy*.

[17] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1832–1846.

[18] Paul Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*. Wiley Online Library, 45–60.

[19] Jiamin Fu, Qirong Mao, Juanjuan Tu, and Yongzhao Zhan. 2017. Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. *Multimedia Systems* 25, 5 (2017), 451–461.

[20] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium*. 3454–3466.

[21] Devamanyu Hazarika, Sruthi Gorantla, Soujanya Poria, and Roger Zimmermann. 2018. Self-attentive feature-level fusion for multimodal emotion detection. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval, Miami, FL*. 196–201.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV*. 770–778.

[23] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen. 2019. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5866–5870.

[24] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[25] Chan Woo Lee, Kyu Ye Song, Jihoon Jeong, and Woo Yong Choi. 2018. Convolutional attention networks for multimodal emotion recognition from speech and text data. *arXiv preprint arXiv:1805.06606* (2018).

[26] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain*. 169–176.

[27] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan*. 284–288.

[28] Mathieu Pagé Fortin and Brahim Chaib-draa. 2019. Multimodal multitask emotion recognition using images, texts and tags. In *Proceedings of the ACM Workshop on Crossmodal Learning and Application*. ACM, 3–10.

[29] Bo Pang, and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan*. 115–124.

[30] Amol S. Patwardhan. 2017. Multimodal mixed emotion detection. In *Proceedings of the 2nd International Conference on Communication and Electronics Systems, Coimbatore, India*. 139–143.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*. 1532–1543.

[32] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[33] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108 (2016), 42–49.

[34] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, Vancouver, Canada*. 873–883.

[35] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Proceedings of the 2017 IEEE International Conference on Data Mining, New Orleans, LA*. 1033–1038.

[36] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Proceedings of 2016 IEEE 16th International Conference on Data Mining*. Barcelona, Spain, 439–448.

[37] Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. 2017. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* 261 (2017), 217–230.

[38] Farhad Rahdari, Esmat Rashedi, and Mahdi Eftekhari. 2019. A multimodal emotion recognition system using facial landmark analysis. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 43, 1 (2019), 171–189.

[39] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas BaltrusÌaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *Proceedings of the 2016 European Conference on Computer Vision, Amsterdam, Netherland*. 338–353.

[40] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY*. 1–9.

[41] Saurav Sahay, Shachi H. Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. 2018. Multimodal relational tensor network for sentiment and emotion classification. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language, Melbourne, Australia*. 20–27.

[42] Suyash Sangwan, Dushyant Singh Chauhan, Md. Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis. In *Neural Information Processing*, Tom Gedeon, Kok Wai Wong, and Minho Lee (Eds.). Springer International Publishing, Cham, 662–669.

[43] Maximilian Schmitt and Björn Schuller. 2017. OpenXBOW: Introducing the passau open-source crossmodal bag-of-words toolkit. *The Journal of Machine Learning Research* 18, 1 (2017), 3370–3374.

[44] Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2018. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language, Melbourne, Australia*. 35–39.

[45] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15), San Diego, CA, USA, May 7-9, 2015*.

[46] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT*. 4223–4232.

[47] Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada*. 1547–1556.

[48] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[49] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo, Hong Kong*. 949–954.

[50] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language, Melbourne, Australia*. 11–19.

[51] R. Xia and Y. Liu. 2017. A multi-task learning framework for emotion recognition using 2D continuous space. *IEEE Transactions on Affective Computing* 8, 1 (Jan 2017), 3–14.

[52] Nan Xu and Wenji Mao. 2017. MultiSentiNet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore*. 2399–2402.

[53] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining, San Francisco, CA*. 13–22.

[54] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*. 1103–1114.

[55] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA*. 5634–5641.

[56] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia*. 2236–2246.

[57] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA*. 5642–5649.

[58] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (Nov 2016), 82–88.

[59] Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. 2018. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science* 752 (2018), 21–40.

Received May 2019; revised November 2019; accepted January 2020