Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal and Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Patna
Patna, Bihar, India-801106
{1821CS17, shad.pcs15, asif, pb}@iitp.ac.in

Abstract

In recent times, multi-modal analysis has been an emerging and highly sought-after field at the intersection of natural language processing, computer vision, and speech processing. The prime objective of such studies is to leverage the diversified information, (e.g., textual, acoustic and visual), for learning a model. The effective interaction among these modalities often leads to a better system in terms of performance. In this paper, we introduce a recurrent neural network based approach for the multi-modal sentiment and emotion analysis. The proposed model learns the inter-modal interaction among the participating modalities through an auto-encoder mechanism. We employ a context-aware attention module to exploit the correspondence among the neighboring utterances. We evaluate our proposed approach for five standard multi-modal affect analysis datasets. Experimental results suggest the efficacy of the proposed model for both sentiment and emotion analysis over various existing state-of-the-art systems.

1 Introduction

In recent past, the world has witnessed tremendous growth of various social media platforms, e.g., YouTube, Instagram, Twitter, Facebook, etc. People treat these platforms as a communication medium and freely express themselves with the help of a diverse set of input sources, e.g. *videos*, *images*, *audio*, *text* etc. The amount of information produced daily through these mediums are enormous, and hence, the research on multi-modal information processing has attracted attention to the researchers and developers. A video is a multi-modal input which provides visual, acoustic, and textual information.

The motivation of multi-modal sentiment and emotion analysis lies in fact to leverage the varieties of (often distinct) information from multiple sources for building more efficient systems. For some cases, text can provide a better clue for the prediction, whereas for the others, acoustic or visual sources can be more informative. Similarly, in some situations, a combination of two or more information sources together ensures better and unambiguous classification decision. For example, only text *"shut up"* can not decide the mood of a person but *acoustic (tone of a person)* and *visual (expression of a person)* can reveal the exact mood. Similarly, for some instances visual features such as *gesture*, *postures*, *facial expression* etc. have important roles to play in determining the correctness of the system. However, effectively combining this information is a non-trivial task that researchers often have to face (Poria et al., 2016; Ranganathan et al., 2016; Lee et al., 2018).

Traditionally, '*text*' has been the key factor in any Natural Language Processing (NLP) tasks, including sentiment and emotion analysis. However, with the recent emergence of social media platforms, an interdisciplinary study involving *text*, *visual* and *acoustic* features have drawn a great interest among the research community. Expressing the feelings and emotions through a video is much convenient than the text for a user, and it is the best source to extract all multi-modal information. Not only the *visual*, it also provides other information such as *acoustic* and *textual* representation of *spoken language*. Additionally, a single video can have multiple utterances based on a speaker's pause (speech bounded by breaths) with different sentiments and emotions. The sentiments and emotions of an utterance often have inter-dependence on the other contextual utterances. Independently classifying such an utterance poses several challenges to the underlying problem. In contrast, multi-modal sentiment and emotion analysis take inputs from more than one sources e.g.

text, *visual*, *acoustic* for the analysis. Effectively fusing this diverse information is non-trivial and poses several challenges to the underlying problem.

In our current work, we propose an end-to-end Context-aware Interactive Attention (CIA) based recurrent neural network for sentiment and emotion analysis. We aim to leverage the interaction between the modalities to increase the confidence of individual task in prediction. The main contributions of our current research are as follows: **(1)** *We propose an Inter-modal Interactive Module (IIM) that aims to learn the interaction among the diverse and distinct features of the input modalities, i.e., text, acoustic and visual*; **(2)** *We employ a Context-aware Attention Module (CAM) that identifies and assigns the weights to the neighboring utterances based on their contributing features. It exploits the interactive representations of pairwise modalities to learn the attention weights*, and **(3)** *We present new state-of-the-arts for five benchmark datasets for both sentiment and emotion predictions*.

2 Related Work

Different reviews in (Arevalo et al., 2017; Poria et al., 2016, 2017b; Ghosal et al., 2018; Morency et al., 2011a; Zadeh et al., 2018a; Mihalcea, 2012; Lee et al., 2018; Tsai et al., 2018) suggest that multi-modal sentiment and emotion analysis are relatively new areas as compared to uni-modal analysis. Feature selection (fusion) is a challenging and important task for any multi-modal analysis. Poria et al. (2016) proposed a multi-kernel learning based feature selection method for multimodal sentiment and emotion recognition. A convolutional deep belief network (CDBN) is proposed in (Ranganathan et al., 2016) to learn salient multi-modal features of low-intensity expressions of emotions, whereas Lee et al. (2018) introduced a convolutional attention network to learn multimodal feature representation between speech and text data for multi-modal emotion recognition.

A feature level fusion vector was built, and then a Support Vector Machine (SVM) classifier was used to detect the emotional duality and mixed emotional experience in (Patwardhan, 2017). Similar work on feature-level fusion based on self- attention mechanism is reported in (Hazarika et al., 2018). Fu et al. (2017) introduced an enhanced sparse local discriminative canonical correlation analysis approach (En-SLDCCA) to learn the multi-modal shared feature representation. Tzirakis et al. (2017) introduced a Long Short Term Memory (LSTM) based end-to-end multi-modal emotion recognition system in which convolutional neural network (CNN) and a deep residual network are used to capture the emotional content for various styles of speaking, robust features.

Poria et al. (2017a) presented a literature survey on various affect dimensions e.g., sentiment analysis, emotion analysis, etc., for the multi-modal analysis. A multi-modal fusion-based approach is proposed in (Blanchard et al., 2018) for sentiment classification. The author used exclusively high-level fusion of visual and acoustic features to classify the sentiment. Zadeh et al. (2016) presented the multi-modal dictionary-based technique to capture the interaction between spoken words and facial expression better when expressing the sentiment. In another work, Zadeh et al. (2017) proposed a Tensor Fusion Network (TFN) to capture the inter-modality and intra-modality dynamics between the multi-modalities (i.e., text, visual, and acoustic).

These works did not take contextual information into account. Poria et al. (2017b) introduced an Long Short Term Memory (LSTM) based framework for sentiment classification which uses contextual information to capture interrelationships between the utterances. In another work, Poria et al. (2017c) proposed a user opinion based framework to combine all the multi-modal inputs (i.e., visual, acoustic, and textual) by applying a multi-kernel learning-based approach. Contextual inter-modal attention mechanism was not explored in much details until recently. Zadeh et al. (2018a) introduced a multi-attention blocks based model for multi-modal sentiment classification but did not account for contextual information, whereas Ghosal et al. (2018) proposed a contextual inter-modal attention based framework for multi-modal sentiment classification. Recently, Zadeh et al. (2018c) introduced the largest multimodal dataset namely *CMU-MOSEI* for sentiment and emotion analysis. Author effectively fused the multi-modality inputs *i.e., text, visual, and acoustic* through a dynamic fusion graph and reported competitive performance *w.r.t.* various state-of-the-art systems for both sentiment and emotion analysis. Very recently, Akhtar et al. (2019) in-

troduced an attention based multi-task learning framework for sentiment and emotion classification on the *CMU-MOSEI* dataset.

In comparison to the existing systems, our proposed approach aims to exploits the interaction between the input modalities through an auto-encoder based *inter-modal interactive module*. The interactive module learns the joint representation for the participating modalities, which are further utilized to capture the contributing contextual utterances in a *context-aware attention module*.

3 Context-aware Interactive Attention (CIA) Affect Analysis

In this section, we describe our proposed approach for the effective fusion of multi-modal input sources. We propose an end-to-end Context-aware Interactive Attention (CIA) based recurrent neural network for sentiment and emotion analysis. As discussed earlier, one of the main challenges for multi-modal information analysis is to exploit the interaction among the input modalities. Therefore, we introduce an Inter-modal Interactive Module (IIM) that aims to learn the interaction between any two modalities through an auto-encoder like structure. For the *text-acoustic* pair of modalities, we aim to decode the *acoustic* representation through the encoded *textual* representation. After training of IIM, we extract the encoded representation for further processing. We argue that the encoded representation learns the interaction between the *text* and *acoustic* modalities. Similarly, we compute the interaction among all the other pairs (i.e., *acoustic-text*, *text-visual*, *visual-text*, *acoustic-visual*, and *visual-acoustic*). Next, we extract the sequential pattern of the utterances through a Bi-directional Gated Recurrent Unit (Bi-GRU) (Cho et al., 2014)). For each pair of modalities, the two representations denoting the interactions between them are combined through a mean operation. For an instance, we compute the mean of the *text-acoustic* and *acoustic-text* representations for text and acoustic modalities. The mean operation ensures that the network utilizes the two distinct representations by keeping the minimal dimension.

In our network, we, additionally, learn the interaction among the modalities through a feed-forward network. At first, all the three modalities are passed through a separate Bi-GRU. Then, pair-

Algorithm 1 Inter-modal Interactive Module for Multi-modal Sentiment and Emotion Recognition (IIM-MMSE)

procedure IIM-MMSE(t, v, a)
  for $i \in 1, ..., K$ **do** ▷ $K = \#modalities$
    for $j \in 1, ..., K$ **do**
      ▷$\forall x, y \in [T, V, A]$, $x \neq y$ and $i \leq j$
      $C_{x_i y_j} \leftarrow IIM(x_i, y_j)$
      $C_{x_i y_j} \leftarrow biGRU(C_{x_i y_j})$
      $C_{x_i} \leftarrow biGRU(x_i)$

  for $i, j \in 1, ..., K$ **do**
    ▷$\forall x, y \in [T, V, A]$, and $x \neq y$
    $M_{x_i,y_j} \leftarrow Mean(C_{x_i y_j}, C_{y_i x_j})$
    $cat_{x_i,y_j} \leftarrow Concatenate(C_{x_i}, C_{y_j})$
    $BI_{x_i,y_j} \leftarrow FullyConnected(cat_{x_i,y_j})$
    $A_{x_i,y_j} \leftarrow CAM(M_{x_i,y_j}, BI_{x_i,y_j})$
  $Rep \leftarrow [A_{TV}, A_{TA}, A_{AV}]$
  $polarity \leftarrow Sent(Rep)/Emo(Rep)$
  return $polarity$

Algorithm 2 Inter-Modal Interactive Module (IIM)

procedure IIM(X, Y)
  $C_{XY} \leftarrow IIM_{Encoder}(X, Y)$
  $\widetilde{Y} \leftarrow IIM_{Decoder}(C_{XY})$
  $loss \leftarrow$ cross_entropy$(\widetilde{Y}, Y)$
  Backpropagation to update the weights
  return C_{XY}

Algorithm 3 Context-aware Attention Module (CAM)

procedure CAM(M, BI)
  $P \leftarrow M.BI^T$ ▷ Cross product
  for $i, j \in 1, ..., u$ **do** ▷ $u =$ #utterances
    $N(i,j) \leftarrow \frac{e^{P(i,j)}}{\sum_{k=1}^{u} e^{P(i,k)}}$
  $O \leftarrow N.BI$
  $A \leftarrow O \odot M$ ▷ Multiplicative gating.
  return A

wise concatenation is performed over the output of Bi-GRU and passed through a fully-connected layer to extract the bi-modal interaction (BI). Further, we employ a Context-aware Attention Module (CAM) to exploit the correspondence among the neighboring utterances. The inputs to the CAM are the two representations for each pair of modalities, e.g., mean representation M_{TA} and bi-modal interaction BI_{TA} for the text-acoustic

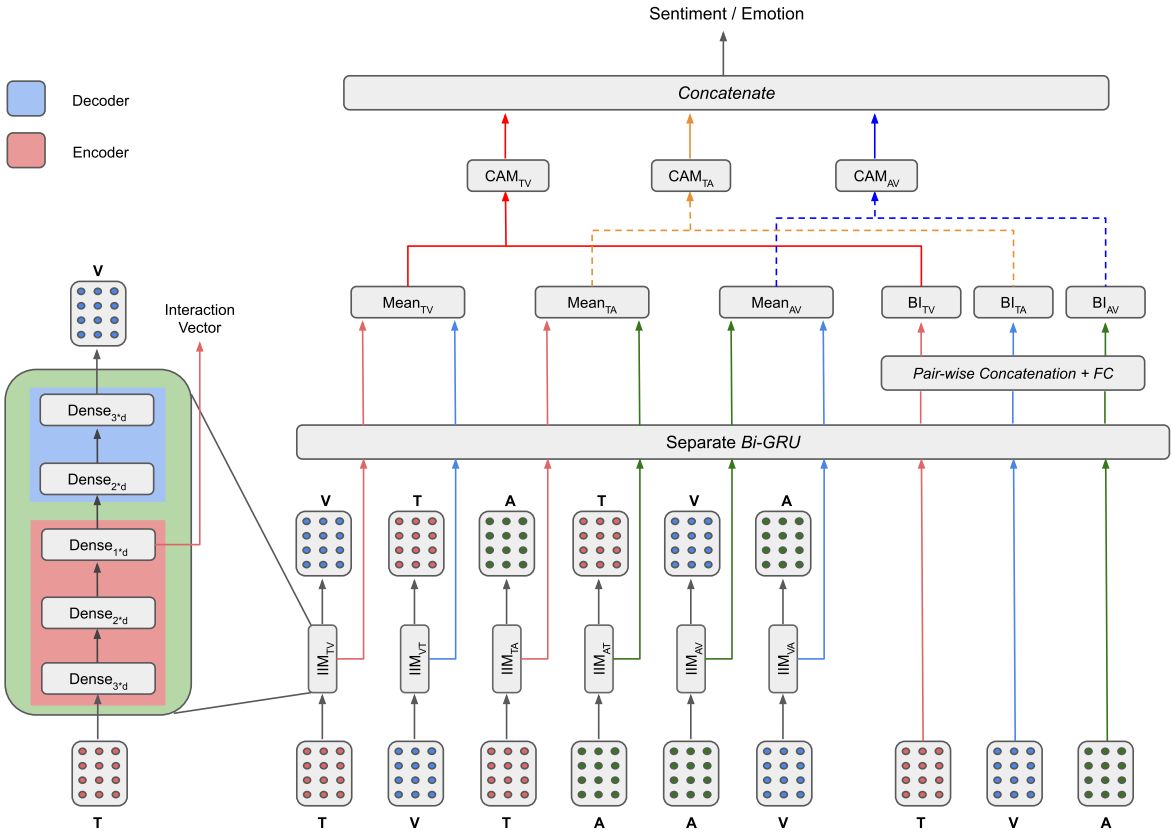


Figure 1: Overall architecture of the proposed Context-aware Interactive Attention framework.

pair. The attention module assists the network in attending the contributing features by putting weights to the current and the neighboring utterances in a video. In the end, the pair-wise (i.e., *text-acoustic*, *text-visual*, and *acoustic-visual*) attended representations are concatenated and fed to an output layer for the prediction.

We depict and summarize the proposed approach in Figure 1 and Algorithm 1, 2, and 3. The source code is available at `http://www.iitp.ac.in/~ai-nlp-ml/resources.html`.

3.1 Context-aware Attention Module (CAM)

Since the utterances in a video are the split units of the break/pause of the speech, their emotions (or sentiments) often have relations with their neighboring utterances. Therefore, knowledge of the emotions (or, sentiments) of the neighboring utterances is an important piece of information and has the capability to derive the prediction of an utterance, if the available inputs are insufficient for the correct prediction.

Our proposed context-aware attention module leverages the contextual information. For each utterance in a video, we compute the attention weights of all the neighboring utterances based on their contributions in predicting the current utterance. It ensures that the network properly utilizes the local contextual information of an utterance as well as the global contextual information of a video together. The aim is to compute the interactive attention weights utilizing a *softmax* activation for each utterance in the video. Next, we apply a multiplicative gating mechanism following the work of Dhingra et al. (2016). The attentive representation is, then, forwarded to the upper layers for further processing. We summarize the process of *CAM* in Algorithm3.

3.2 Inter-modal Interactive Module (IIM)

One of the key objectives of the multi-modal analysis is to fuse the available input modalities effectively. In general, different modalities represent distinct features despite serving a common goal. For example, in multi-modal sentiment analysis all the three modalities, i.e., *text*, *acoustic*, and *visual*, aim to predict the expressed polarity of an utterance. The distinctive features in isolation

might create an ambiguous scenario for a network to learn effectively. Therefore, we introduce an auto-encoder based inter-modal interactive module whose objective is to learn the interaction between two distinct modalities to serve a common goal. The IIM encodes the feature representation of one modality (say, text), and aims to decode it into the feature representation of another modality (say, acoustic). Similar to an auto-encoder where the input and output are conceptually the same (or closely related), in our case the input and output feature representations of two modalities also intuitively serve a common goal. After training of IIM, the encoded vector signifies a joint representation of the two modalities, which can be further utilized in the network.

As the proposed architecture in Figure 1 depicts, our proposed model is an *end-to-end* system, which takes multi-modal raw features for each utterance in a video and predicts an output. We also train our proposed IIM in the combined framework. For any pair of modalities, e.g., *text-visual*, the encoded vector in IIM receives two gradients of errors, i.e., one error from the IIM output (visual) l_1 and another from the task-specific label l_2. We aggregate the errors ($l_1 + l_2$) at the encoded vector and backpropagate it to the input (text). Thus, the weights in the encoder part will adjust according to the desired task-specific label as well. However, in contrast, the decoder part does not have such information. Therefore, we employ another IIM to capture the interaction between the *visual-text*. This time, the *visual* features are aware of the desired label during the interaction with *textual* features. A conceptual diagram, depicting the gradient flow in IIM for the text and visual modalities, is shown in Figure 2.

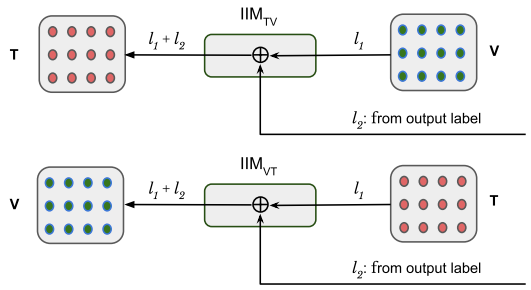


Figure 2: Difference between the interaction modules of *text-visual* and *visual-text* pairs. l_1 & l_2 are the losses from the IIM and output labels, respectively.

4 Datasets, Experiments and Analysis

In this section, we present our experimental results along with necessary analysis. We also compare our obtained results with several state-of-the-art systems.

4.1 Datasets

For the evaluation of our proposed approach, we employ five multi-modal benchmark datasets[1] covering two affect analysis tasks, i.e., sentiment and emotion.

1. **YouTube** (Morency et al., 2011b): The YouTube opinion dataset contains 269 product reviews utterances across 47 videos. There are 169, 41, and 59 utterances in training, validation, and test set, respectively.

2. **MOUD** (Pérez-Rosas et al., 2013): The dataset consists of 79 product review videos in Spanish. Each video consists of multiple utterances labeled with either positive, negative, or neutral sentiment. There are 243, 37, 106 utterances in training, validation, and test set, respectively.

3. **ICT-MMMO** (Wöllmer et al., 2013): It is an extension of YouTube opinion dataset that extends the number of videos from 47 to 340. Each online social review video is annotated for the sentiment. There are 220, 40, and 80 videos in training, validation, and test set, respectively.

4. **CMU-MOSI** (Zadeh et al., 2016): It is a collection of 2199 opinion utterances annotated with the sentiment class. There are 1284 utterances in the training set, 229 utterances in the validation set, and 686 utterances in the test set.

5. **CMU-MOSEI** (Zadeh et al., 2018c): CMU-MOSEI is the largest dataset among all the above. It has 3,229 videos which comprise of approximately 23,000 utterances. Each utterance is associated with one sentiment value and six emotions (i.e., *anger* (4903 utterances), *disgust* (4028 utterance), *fear* (1850 utterance), *happy* (12135 utterance),

[1]These datasets can be accessed through https://github.com/A2Zadeh/CMU-MultimodalSDK.

	MOSEI								MOSI					ICT-MMMO		YouTube		MOUD	
	Emotion		Sentiment						Sentiment					Sentiment		Sentiment		Sentiment	
	F1	*W-Acc*	*F1*	A^2	A^5	A^7	*MAE*	*r*	*F1*	A^2	A^7	*MAE*	*r*	*F1*	A^2	*F1*	A^3	*F1*	A^2
T	77.66	60.13	77.22	79.45	48.50	48.87	0.695	0.572	77.43	77.69	37.75	0.996	0.658	71.80	77.50	48.27	49.15	72.24	73.58
A	76.14	57.68	74.64	77.27	45.02	44.97	0.782	0.433	59.30	59.32	23.17	1.423	0.165	77.90	78.34	43.12	44.06	45.46	60.37
V	75.03	57.59	69.13	75.04	42.22	42.51	0.804	0.317	50.48	51.31	23.03	1.465	0.122	78.28	78.75	49.79	50.84	62.47	62.76
T+V	78.01	57.10	77.31	79.51	48.98	49.84	0.685	0.592	77.70	78.13	38.19	0.946	0.673	80.10	81.25	53.11	53.91	78.40	79.24
T+A	77.53	60.69	77.80	80.06	48.66	49.08	0.694	0.579	78.98	79.15	39.35	0.952	0.671	78.99	80.01	48.82	49.37	72.93	74.52
A+V	76.60	59.47	74.76	77.38	45.32	45.87	0.775	0.437	49.30	53.49	24.19	1.464	0.189	79.12	81.25	51.93	52.54	64.41	65.09
T+A+V	**79.02**	**62.97**	**78.23**	**80.37**	**49.15**	**50.14**	**0.683**	**0.594**	**79.54**	**79.88**	**38.92**	**0.914**	**0.689**	**81.47**	**82.75**	**55.13**	**55.93**	**82.07**	**82.41**

Table 1: Results of sentiment and emotion analysis for the proposed approach. **T:** Text, **V:** Visual, **A:** *Acoustic*. Weighted accuracy as a metric is chosen due to unbalanced samples across various emotions and it is also in line with the other existing works (Zadeh et al., 2018c).

Multi-label	**No**	**One**	**Two**	**Three**	**Four**	**Five**	**Six**
Count	3372	11050	5526	2084	553	84	8

Table 2: Statistics of multi-label emotions in CMU-MOSEI: *Emotions-per-utterance*.

sad (5856 utterance), and *surprise* (2262 utterance). The emotion class *happy* is approximately 52% of the total utterances, while emotions *fear* and *surprise* are approximately 8-9% in the dataset. Further, many utterances have more than one emotions representing the case of multi-label classification problem. The utterances for which all the emotions are absent, we classify them into a *no-emotion* class. We depict the statistics of multi-label emotions in Table 2. The training, validation and test set distributions are approximately 16K, 2K, and 4.6K, respectively.

4.2 Setups

The above datasets offer different dimension of sentiment analysis. We define the following setups for our experiments.

- **Two-class** (*pos* and *neg*) classification: MOSEI, MOSI, ICT-MMMO, and MOUD.
- **Three-class** (*pos*, *neu*, and *neg*) classification: YouTube.
- **Five-class** (*strong pos*, *weak pos*, *neu*, *weak neg*, and *strong neg*) classification: MOSEI.
- **Seven-class** (*strong pos*, *moderate pos*, *weak pos*, *neu*, *weak neg*, *moderate neg*, and *strong neg*) classification: MOSEI and MOSI.
- **Intensity prediction:** MOSEI and MOSI.

4.3 Experiments

We implement our proposed model on the Python-based Keras deep learning library. As the evaluation metric, we employ accuracy (weighted accuracy (Tong et al., 2017)) and F1-score for the classification problems, while for the intensity prediction task, we compute Pearson correlation scores and mean-absolute-error (MAE).

We evaluate our proposed *CIA* model on five benchmark datasets *i.e., MOUD, MOSI, YouTube, ICT-MMMO, and MOSEI*. For all the datasets, we perform *grid search* to find the optimal hyper-parameters (c.f. Table 4). Though we push for a generic hyper-parameter configuration for all datasets, in some cases, a different choice of the parameter has a significant effect. Therefore, we choose different parameters for different datasets for our experiments. Details of hyper-parameters for different datasets are depicted in Table 4.

We use different activation functions for the various modules in our model. We use *tanh* as the activation function for the inter-modal interactive module (IIM), while we employ *ReLu* for the context-aware attention module. For each dataset, we use *Adam* as optimizer.

In this paper, we address three multi-modal affective analysis problems, namely *i.e., sentiment classification (S_C), sentiment intensity (S_I) and emotion classification (E_C)*. We use *softmax* as a classifier for sentiment classification, while optimizing the *categorical cross-entropy* as a loss function. In comparison, we use *sigmoid* for prediction and *binary cross-entropy* as the loss function for the emotion classification. As the emotions in the dataset are multi-labeled, we apply a threshold over the predicted *sigmoid* outputs for each emotion and consider all the emotions as present whose respective values are above the threshold. We cross-validate and optimize both

Modality	MOSEI								MOSI					ICT-MMMO		YouTube		MOUD	
	Emotion		Sentiment						Sentiment					Sentiment		Sentiment		Sentiment	
	F1	*W-Acc*	*F1*	A^2	A^5	A^7	*MAE*	*r*	*F1*	A^2	A^7	*MAE*	*r*	*F1*	A^2	*F1*	A^3	*F1*	A^2
CIA	**79.02**	**62.97**	**78.23**	**80.37**	**49.15**	**50.14**	**0.683**	**0.594**	**79.54**	**79.88**	**38.92**	**0.914**	**0.689**	**81.47**	**82.75**	**55.13**	**55.93**	**82.07**	**82.41**
CIA – IIM	77.81	61.86	77.69	79.53	48.02	49.16	0.714	0.566	39.32	55.24	34.40	0.941	0.652	80.10	81.25	50.02	52.54	78.14	78.30

Table 3: Ablation results for IIM module.

the evaluation metrics, i.e., F1-score and weighted accuracy, and set the threshold as 0.4 and 0.18, respectively.

Parameters	MOUD	MOSI	YouTube	MMMO	MOSEI
Bi-GRU	50N			200N	100N
	0.3D				0.5D
FC	50N		200N	100N	
	0.5D				
Activations	*ReLu* as activation for our model				
	tanh as activation in IIM				
Output	Softmax (S_C), tanh (S_I) & Sigmoid (E_C)				
Optimizer	Adam (lr=0.001)				
IIM Loss	Mean Square Error (MSE)				
Model Loss	Cross-entropy (Classification) & MSE (Intensity)				
Threshold	0.4 (F1) & 0.18 (W-Acc) for multi label in E_C				
Batch	16				
Epochs	50				

Table 4: Hyper-parameters for our experiments where *N, D*, S_C, S_I and E_C stands for #neurons, dropout, sentiment classification, sentiment intensity and emotion classification respectively.

We evaluate our proposed approach for all the possible input combinations *i.e., uni-modal (T, A, V), bi-modal (T+V, T+A, A+V) and tri-modal (T+V+A)*. We depict our obtained results in Table 1. For MOSEI dataset, with tri-modal inputs, our proposed system reports 79.02% F1-score and 62.97% weighted-accuracy for emotion classification. For sentiment classification, we obtain 78.23%, 80.37%, 49.15% and 50.14% as F1-score for two-class, five-class and seven-class, respectively. For sentiment intensity prediction task, our proposed system yields MAE and Pearson score of 0.683 and 0.594, respectively. We also observe that the proposed approach yields better performance for the tri-modal inputs than the bi-modal and uni-modal input combinations. This improvement implies that our proposed CIA architecture utilizes the interaction among the input modalities very effectively. Furthermore, for the other datasets, i.e., MOSI, ICT-MMMO, YouTube, and MOUD, we also observe a similar phenomenon as well (c.f. Table 1).

To show that our proposed IIM module, indeed, learns the interaction among the distinct modalities, we also perform an ablation study of the proposed CIA architecture. Consequently, we omit the IIM module from our architecture and compute the self-attention on the pair-wise fully-connected representations for the prediction. We observe that, for all the datasets, the performance of this modified architecture (i.e., CIA - IIM) is constantly inferior (with 1% to 7% F-score points) to the proposed CIA architecture. This performance degradation suggests that the IIM module is, indeed, an important component of our proposed architecture. In Table 3, we depict the evaluation results for both- with and without IIM.

4.4 Comparative Analysis

In this section, we present our comparative studies against several existing and recent state-of-the-art systems. For each dataset, we report three best systems for the comparisons[2]. In particular, we compare with the following systems: **Bag of Feature - Multimodal Sentiment Analysis (BoF-MSA)** (Blanchard et al., 2018), **Memory Fusion Network (MFN)** (Zadeh et al., 2018b), **Deep Fusion - Deep Neural Network (DF-DNN)** (Nojavanasghari et al., 2016), **Multi View - LSTM (MV-LSTM)** (Rajagopalan et al., 2016), **Early Fusion - LSTM (EF-LSTM)** (Zadeh et al., 2018c), **Tensor Fusion Network (TFN)** (Zadeh et al., 2017), **Random Forest (RF)** (Breiman, 2001), **Support Vector Machine** (Zadeh et al., 2016), **Multi-Attention Recurrent Network (MARN)** (Zadeh et al., 2018a), **Dynamic Fusion Graph (DFG)** (Zadeh et al., 2018c), **Multi Modal Multi Utterance-Bimodal Attention (MMMU-BA)** (Ghosal et al., 2018), **Bi-directional Contextual LSTM (BC-LSTM)** (Poria et al., 2017b) and **Multimodal Factorization Model (MFM)** (Tsai et al., 2018).

We show the comparative results in Table 5a and Table 5b for emotion and sentiment analysis, respectively. We observe that the proposed CIA framework yields better performance against the state-of-the-art for all the cases. For emotion classification, our proposed approach achieves approximately 3 and 0.6 percentage higher F1-

[2]Please note that we report all the results, which are available for comparison

System	MOSEI													
	Anger		*Disgust*		*Fear*		*Happy*		*Sad*		*Surprise*		*Average*	
	F1	*W-Acc*	*F1*	*W-Acc*	*F1*	*W-Acc*	*F1*	*W-Acc*	*F1*	*W-Acc*	*F1*	*W-Acc*	*F1*	*W-Acc*
MFN⋆	-	-	71.4	65.2	**89.9**	-	-	-	60.8	-	85.4	53.3	-	-
DF⋆	71.4	-	-	67.0	-	-	-	-	-	-	-	-	-	-
MV-LSTM⋆	-	56.0	-	-	-	-	-	-	-	-	-	-	-	-
EF-LSTM⋆	-	-	-	-	-	56.7	-	57.8	-	59.2	-	-	-	-
TFN⋆	-	60.5	-	-	-	-	66.6	66.5	-	58.9	-	52.2	-	-
RF⋆	72.0	-	73.2	-	**89.9**	-	-	-	61.8	-	85.4	-	-	-
SVM⋆	-	-	-	-	-	60.0	-	-	-	-	-	-	-	-
MARN⋆	-	-	-	-	-	-	71.0	-	-	-	-	-	-	-
DFG	72.8	62.6	76.6	69.1	**89.9**	62.0	66.3	66.3	66.9	60.4	85.5	53.7	76.3	62.3
CIA	**74.7**	**67.4**	**81.8**	**74.1**	87.8	**63.9**	**71.3**	51.9	**72.6**	**61.8**	**86.0**	**58.2**	**79.0**	**62.9**
T-test	-	-	-	-	-	-	-	-	-	-	-	-	***0.0002***	***0.0057***

(a) Emotion Analysis

System	MOSEI						MOSI					ICT-MMMO		YouTube		MOUD	
	F1	A^2	A^5	A^7	*MAE*	*r*	*F1*	A^2	A^7	*MAE*	*r*	*F1*	A^2	*F1*	A^3	*F1*	A^2
MARN†	-	-	-	-	-	-	77.0	77.1	34.7	0.968	0.625	-	-	-	48.3	81.2	81.1
MFN†	76.0	76.0	-	-	-	-	77.3	77.4	34.1	0.965	0.632	73.1	73.8	51.6	51.7	80.4	81.1
TFN†	-	-	-	-	-	-	-	-	-	-	-	72.6	72.5	-	-	-	-
BC-LSTM†	-	-	-	-	-	-	-	-	-	-	-	-	-	45.1	-	-	-
MFM	-	-	-	-	-	-	78.1	78.1	36.2	0.951	0.662	79.2	81.3	52.4	53.3	81.7	82.1
BoF-MSA⋆	63.2	60.0	-	-	0.91	0.30	-	-	-	-	-	-	-	-	-	-	-
MV-LSTM⋆	76.4	76.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DFG	77.0	76.9	45.1	45.0	0.71	0.54	-	-	-	-	-	-	-	-	-	-	-
MMMU-BA	77.6	79.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CIA	**78.2**	**80.4**	**49.2**	**50.1**	**0.68**	**0.59**	**79.54**	**79.88**	**38.92**	**0.914**	**0.689**	**81.47**	**82.75**	**55.13**	**55.93**	**82.07**	**82.41**
T-test	*0.038*	*0.038*	*0.00003*	*0.00001*	*0.0002*	*0.00001*	*0.0049*	*0.0022*	*0.0003*	*0.0018*	*0.0003*	*0.0005*	*0.0041*	*0.00006*	*0.00003*	*0.031*	*0.040*

(b) Sentiment Analysis

Table 5: **Comparative results.** ⋆Values are taken from (Zadeh et al., 2018c). †Values are taken from (Tsai et al., 2018). Significance *T*-test (< 0.05) signifies that the obtained results are statistically significant over the existing systems with 95% confidence score.

score and weighted accuracy, respectively, than the state-of-the-art DFG (Zadeh et al., 2018c) system. Furthermore, we also see improvements for most of the individual emotion classes as well.

In sentiment analysis (c.f. Table 5b), for all the five datasets and different experimental setups, the proposed CIA framework obtains the improved accuracies for the classification tasks. For intensity prediction, our proposed framework yields lesser mean-absolute-error with high Pearson correlation scores. On average, we observe 1 to 5% improvement in accuracy values in comparison to the next best systems. Similarly, for the intensity prediction task, we report approximately 0.03 and 0.04 points improvement in mean-absolute-error and Pearson score, respectively.

We perform statistical significance test (*paired T-test*) on the obtained results and observe that performance improvement in the proposed model over the state-of-the-art is significant with 95% confidence (i.e., *p*-value< 0.05).

4.5 Error Analysis

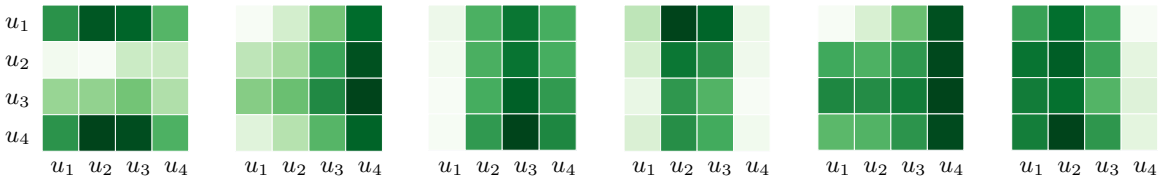
We analyze our proposed *CIA* model to understand the importance of the baseline framework *CIA-IIM*. We study the predictions of both the models and observe that the proposed *CIA* framework improves the predictions of the baseline *CIA–IIM* model. It indicates that the CIA framework, indeed, learns the interaction among the input modalities, and the model effectively exploits this interaction for better judgment. In Table 6, we list the utterances of a CMU-MOSEI video along with their correct and predicted labels for both the proposed and baseline systems.

The video in Table 6 has 4 utterances, out of which the correct sentiments of three utterances (i.e., u_1, u_3, and u_4) are *positive*, while one utterance (i.e., u_2) is *negative*. We observe that our proposed *CIA* model predicts all the 4 utterances correctly, while the *CIA-IIM* mis-classify the sentiments of the utterances, u_2 and u_3.

We also analyze the *context-aware attention module (CAM)* with the help of heatmaps of the attention weights. The heatmaps, as depicted in Figure 3, represent the contributing utterances in the neighbourhood for the classification of each utterance. Figures 3a, 3b and 3c show the heatmaps of the pair-wise modality interaction of the proposed model *CIA*. In Figure 3a, each cell(*i,j*) of

	Utterances	Sentiment			Emotion		
		Actual	CIA−IIM	CIA	Actual	CIA−IIM	CIA
1	*these critics argue that the welfare state breeds dependence and incompetence among those who receive it they*	Pos	Pos	Pos	Happy, Sad	Happy	Happy, Sad
2	*argue that it creates social pathologies such as single parent families excess fertility and laziness*	Neg	Pos	Neg	Sad	Happy	Sad
3	*some argue that people who receive welfare benefits cannot spend their benefits rationally and*	Pos	Neg	Pos	Fear, Sad	Sad	Fear, Sad
4	*and then lastly some people on the moral side argue that nothing should be given to a person without requiring a reciprocal obligation from that person so*	Pos	Pos	Pos	Sad	Happy, Sad	Sad

Table 6: **Comparison between proposed *CIA* and *CIA−IIM* frameworks in MOSEI dataset.** Few cases where *CIA* framework performs better than the *CIA−IIM* framework. Red text signifies error in classification.



(a) $Sentiment_{TV}$ (b) $Sentiment_{TA}$ (c) $Sentiment_{AV}$ (d) $Emotion_{TV}$ (e) $Emotion_{TA}$ (f) $Emotion_{AV}$

Figure 3: **Heatmap analysis of MOSEI dataset** where **(a), (b) & (c)** represents the *Contextual Attention weights* for *TV, TA and AV* for sentiment and **(d), (e) & (f)** are *contextual attention weights* for *TV, TA and AV* for emotion.

the heatmap signifies the weights of utterance '*j*' for the classification of utterance '*i*'. For the utterance u_4, the model puts more attention weights on the u_2 and u_3 of the *text-visual* interactions, while for the *text-acoustic* interaction the model assigns higher weights to the u_4 utterance itself. Similarly, the model assigns the least weight to the u_1 utterance, whereas the utterance u_3 gets the highest weights. We argue that the proposed CAM module captures the diversity in the input modalities of the contextual utterances for the correct prediction.

For emotion prediction, the *CIA* model captures all the emotions correctly, while the *CIA-IIM* framework fails to predict the correct emotions of the utterances, u_2 and u_3. For the same video, we also show the attention heatmaps for emotion in Figure 3. For the utterance u_2, our proposed model (CIA) captures the emotion class '*sad*' as the CAM module assigns higher attention weights on the utterances u_2 and u_3 in Figure 3d, u_4 in Figure 3e, and u_2 in Figure 3f. Since the system finds the contributing neighbours as utterances u_2, u_3 and u_4 for various combinations, we argue that it utilizes the information of these utterances - which all express the '*sad*' emotion - for the correct prediction of utterance u_2 as '*sad*'.

5 Conclusion

In this paper, we have proposed a Context-aware Interactive Attention framework that aims to capture the interaction between the input modalities for the multi-modal sentiment and emotion prediction. We employed a contextual attention module to learn the contributing utterances in the neighborhood by exploiting the interaction among the input modalities. We evaluate our proposed approach on five standard multi-modal datasets. Experiments suggest the effectiveness of the proposed model over various existing systems, for both sentiment and emotion analysis, as we obtained new state-of-the-art for all five datasets.

In current work, we undertook the problem of sentiment and emotion analysis for a single-party utterances. In future, we would like to extend our work towards the multi-party dialogue.

6 Acknowledgment

The research reported here is partially supported by SkyMap Global India Private Limited. Asif Ekbal acknowledges the 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

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

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*, pages 1–10.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

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*, pages 1–11.

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*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Sruthi Gorantla, Soujanya Poria, and Roger Zimmermann. 2018. Self-attentive feature-level fusion for multimodal emotion detection. In *IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, Miami, FL, USA, April 10-12, 2018*, pages 196–201.

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*.

Rada Mihalcea. 2012. Multimodal sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2012, July 12, 2012, Jeju Island, Republic of Korea*, page 1.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011a. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011*, pages 169–176.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011b. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM.

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*, ICMI 2016, pages 284–288, New York, NY, USA. ACM.

Amol S Patwardhan. 2017. Multimodal mixed emotion detection. In *Communication and Electronics Systems (ICCES), 2017 2nd International Conference on*, pages 139–143. IEEE.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 973–982.

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

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. 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)*, volume 1, pages 873–883.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

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

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision (ECCV-2016)*, pages 338–353. Springer International Publishing.

Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE.

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)*, pages 1547–1556. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

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):1301–1309.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

A Zadeh, PP Liang, S Poria, P Vij, E Cambria, and LP Morency. 2018a. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-2018)*, pages 5642 – 5649, New Orleans, USA.

A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88.

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*, pages 1103–1114, Copenhagen, Denmark.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. 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)*, pages 2236–2246. Association for Computational Linguistics.