Contextual Inter-modal Attention for Multi-modal Sentiment Analysis

Deepanway Ghosal[†], **Md Shad Akhtar**[†], **Dushyant Chauhan**[†], **Soujanya Poria**[⋆], **Asif Ekbal**[†] **and Pushpak Bhattacharyya**[†]

[†] *Department of Computer Science & Engineering, Indian Institute of Technology Patna, India*
{deepanway.me14,shad.pcs15,1821CS18,asif,pb}@iitp.ac.in

[⋆] *School of Computer Science and Engineering, Nanyang Technological University, Singapore*
sporia@ntu.edu.sg

Abstract

Multi-modal sentiment analysis offers various challenges, one being the effective combination of different input modalities, namely *text*, *visual* and *acoustic*. In this paper, we propose a recurrent neural network based multi-modal attention framework that leverages the contextual information for utterance-level sentiment prediction. The proposed approach applies attention on multi-modal multi-utterance representations and tries to learn the contributing features amongst them. We evaluate our proposed approach on two multi-modal sentiment analysis benchmark datasets, *viz.* CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus and the recently released CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) corpus. Evaluation results show the effectiveness of our proposed approach with the accuracies of 82.31% and 79.80% for the MOSI and MOSEI datasets, respectively. These are approximately 2 and 1 points performance improvement over the state-of-the-art models for the datasets.

1 Introduction

Traditionally, sentiment analysis (Pang and Lee, 2005, 2008) has been applied to a wide variety of texts (Hu and Liu, 2004; Liu, 2012; Turney, 2002; Akhtar et al., 2016, 2017; Mohammad et al., 2013). In contrast, multi-modal sentiment analysis has recently gained attention due to the tremendous growth of many social media platforms such as YouTube, Instagram, Twitter, Facebook (Chen et al., 2017; Poria et al., 2016, 2017d,b; Zadeh et al., 2017, 2016) etc. It depends on the information that can be obtained from more than one modality (e.g. *text*, *visual* and *acoustic*) for the analysis. The motivation is to leverage the varieties of (often distinct) information from multiple sources for building an efficient system. For example, it is a non-trivial task to detect the sentiment of a sarcastic sentence "*My neighbours are home!! it is good to wake up at 3am in the morning.*" as *negative* considering only the textual information. However, if the system has access to some other sources of information, e.g. *visual*, it can easily detect the unpleasant gestures of the speaker and would classify it with the *negative* sentiment polarity. Similarly, for some instances acoustic features such as *intensity*, *pitch*, *pause* etc. have important roles to play in the correctness of the system. However, combining these information in an effective manner is a non-trivial task that researchers often have to face (Zadeh et al., 2017; Chen et al., 2017).

A video provides a good source for extracting multi-modal information. In addition to the *visual frames*, it also provides information such as *acoustic* and *textual* representation of *spoken language*. Additionally, a speaker can utter multiple utterances in a single video and these utterances can have different sentiments. The sentiment information of an utterance often has inter-dependence on other contextual utterances. Classifying such an utterance in an independent manner poses many challenges to the underlying algorithm.

In this paper, we propose a novel method that employs a recurrent neural network based multi-modal multi-utterance attention framework for sentiment prediction.We hypothesize that applying attention to contributing neighboring utterances and/or multi-modal representations may assist the network to learn in a better way. The main challenge in multi-modal sentiment analysis lies in the proper utilization of the information extracted from multiple modalities. Although it is often argued that incorporation of all the available modalities is always beneficial for enhanced performance, it must be noted that not all the modalities play equal role. Another concern in multi-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3454–3466
Brussels, Belgium, October 31 - November 4, 2018.

modal framework is that the presence of noise in one modality can affect the overall performance. To better address these concerns we propose a novel fusion method by focusing on inter-modality relations computed between the target utterance and its context. We argue that in multi-modal sentiment classification, not only the relation among two modalities of the same utterance is important, but also relatedness with the modalities across its context are important.

Think of an utterance U_t that constitutes of three modalities, say A_t (i.e. audio), V_t (i.e. visual) and T_t (i.e. text). Let us also assume U_k being a member of the contextual utterances consisting of the modalities - A_k, V_k and T_k. In this case, our model computes the relatedness among the modalities (for e.g., V_t and T_k) of U_t and U_k in order to produce a richer multi-modal representation for final classification. The attention mechanism is then used to attend to the important contextual utterances having higher relatedness or similarity (computed using inter-modality correlations) with the target utterance.

Unlike previous approaches that simply apply attentions over the contextual utterance for classification, we attend over the contextual utterances by computing correlations among the modalities of the target utterance and the context utterances. This explicitly helps us to distinguish which modalities of the relevant contextual utterances are more important for sentiment prediction of the target utterance. The model facilitates this modality selection by attending over the contextual utterances and thus generates better multi-modal feature representation when these modalities from the context are combined with the modalities of the target utterance. We evaluate our proposed approach on two recent benchmark datasets, i.e. CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018c), with one being the largest (CMU-MOSEI) available dataset for multi-modal sentiment analysis (c.f. Section 4.1). Evaluation shows that the proposed attention framework attains better performance than the state-of-the-art systems for various combinations of input modalities (i.e. *text*, *visual* & *acoustic*).

The main contributions of our proposed work are three-fold: **a)** we propose a novel technique for multi-modal sentiment analysis; **b)** we propose an effective attention framework that leverages contributing features across multiple modalities and neighboring utterances for sentiment analysis; and **c)** we present the state-of-the-art systems for sentiment analysis in two different benchmark datasets.

2 Related Work

A survey of the literature suggests that multi-modal sentiment prediction is relatively a new area as compared to textual based sentiment prediction (Morency et al., 2011; Mihalcea, 2012; Poria et al., 2016, 2017b; Zadeh et al., 2018a). A good review covering the literature from uni-modal analysis to multi-modal analysis is presented in (Poria et al., 2017a). An application of multi-kernel learning based fusion technique was proposed in (Poria et al., 2016), where they employed deep convolutional neural networks for extracting the textual features and fused it with other (*visual* & *acoustic*) modalities for prediction.

Zadeh et al. (2016) introduced the multi-modal dictionary to better understand the interaction between facial gestures and spoken words when expressing the sentiment. Authors introduced the MOSI dataset, the first of its kind to enable the studies of multi-modal sentiment intensity analysis. Zadeh et al. (2017) proposed a Tensor Fusion Network (TFN) model to learn the intra-modality and inter-modality dynamics of the three modalities (i.e. text, visual and acoustic). They reported the improved accuracy using multi-modality on the CMU-MOSI dataset. An application to leverage on the gated multi-modal embedded Long Short Term Memory (LSTM) with temporal attention (GME-LSTM(A)) for the word-level fusion of multi-modality inputs is proposed in (Chen et al., 2017). The Gated Multi-modal Embedding (GME) alleviates the difficulties of fusion while the LSTM with Temporal Attention (LSTM(A)) performs word-level fusion.

The works mentioned above did not take contextual information into account. Poria et al. (2017b) proposed a LSTM based framework that leverages the contextual information to capture the inter-dependencies between the utterances. In another work, Poria et al. (2017d) proposed an user opinion based framework to combine the three modality inputs (i.e. *text, visual & acoustic*) by applying a multi-kernel learning based method. Zadeh et al. (2018a) proposed multi-attention blocks (MAB) to capture information across three modalities (*text, visual & acoustic*). They reported improved accuracies in the range of

2-3% over the state-of-the-art models for the different datasets.

The fundamental difference between our proposed method and the existing works is that our framework applies focus on the neighboring utterances to leverage contextual information for utterance-level sentiment prediction. To the best of our knowledge, our current work is the very first of its kind that attempts to employ multi-modal attention block (exploiting neighboring utterances) for sentiment prediction. We use *multi-modal attention* framework that leverages contributing features across *multiple modalities* and the *neighboring utterances* for sentiment analysis.

3 Proposed Methodology

In our proposed framework, we aim to leverage the multi-modal and contextual information for predicting the sentiment of an utterance. Utterances of a particular speaker in a video represent the time series information and it is logical that the sentiment of a particular utterance would affect the sentiments of the other neighboring utterances. To model the relationship with the neighboring utterances and multi-modality, we propose a recurrent neural network based multi-modal attention framework. The proposed framework takes multi-modal information (i.e. *text*, *visual* & *acoustic*) for a sequence of utterances and feeds it into three separate bi-directional Gated Recurrent Unit (GRU) (Cho et al., 2014). This is followed by a dense (fully-connected) operation which is shared across the time-steps or utterances (one each for text, visual & acoustic). We then apply multi-modal attention on the outputs of the dense layers. The objective is to learn the joint-association between the multiple modalities & utterances, and to emphasize on the contributing features by putting more attention to these. In particular, we employ bi-modal attention framework, where an attention function is applied to the representations of *pairwise* modalities i.e. *visual-text*, *text-acoustic* and *acoustic-visual*. Finally, the outputs of pairwise attentions along with the representations are concatenated and passed to the *softmax* layer for classification. We call our proposed architecture ***Multi-Modal Multi-Utterance - Bi-Modal Attention (MMMU-BA)*** framework. An overall architecture of the proposed *MMMU-BA* framework is illustrated in Figure 1. Please refer to Figure 3 in appendix for illustration of attention computation.

For comparison, we also experiment with two other variants of the proposed *MMMU-BA* framework i.e. a). *Multi-Modal Uni-Utterance-Self Attention (MMUU-SA)* framework and b). *Multi-Utterance-Self Attention (MU-SA)* framework. The architecture of these variants differ with respect to the attention computation module and the naming conventions "*MMMU*", "*MMUU*" or "*MU*" signify the information that participates in the attention computation. For example, in *MMMU-BA*, we compute attention over the multi-modal and multi-utterance inputs, whereas in *MMUU-SA*, the attention is computed over the mutli-modal but uni-utterance inputs. In contrast, we compute attention over only multi-utterance inputs in *MU-SA*. Rest of the components for all the three variants remain same.

3.1 Multi-modal Multi-utterance - Bi-modal Attention (*MMMU-BA*) Framework

Assuming a particular video has *'u'* utterances, the raw *utterance level* multi-modal features are represented as $T_R \in \mathbb{R}^{u \times 300}$ (*raw text*), $V_R \in \mathbb{R}^{u \times 35}$ (*raw visual*) and $A_R \in \mathbb{R}^{u \times 74}$ (*raw acoustic*). Three separate Bi-GRU layers with forward & backward state concatenation are first applied on the raw data followed by the fully-connected dense layers, resulting in $T \in \mathbb{R}^{u \times d}$ (*text*), $V \in \mathbb{R}^{u \times d}$ (*visual*) and $A \in \mathbb{R}^{u \times d}$ (*acoustic*), where *'d'* is the number of neurons in the dense layer. Finally, *pairwise-attentions* are computed on various combinations of three modalities- (*V*, *T*), (*T*, *A*) & (*A*, *V*). In particular the attention between *V* and *T* is computed as follows:

• ***Bi-modal Attention:*** Modality representations of *V* & *T* are obtained from the Bi-GRU network, and hence contain the contextual information of the utterances for each modality. At first, we compute a pair of matching matrices $M_1, M_2 \in \mathbb{R}^{u \times u}$ over two representations that account for the cross-modality information.

$$M_1 = V.T^T \quad \& \quad M_2 = T.V^T$$

• ***Multi-Utterance Attention:*** As mentioned earlier, in the proposed model we aim to leverage the contextual information of each utterance for the prediction. We compute the probability distribution scores ($N_1 \in \mathbb{R}^{u \times u}$ & $N_2 \in \mathbb{R}^{u \times u}$) over each utterance of bi-modal attention matrices M_1 & M_2 using a softmax function. This essentially computes the attention weights for the contextual

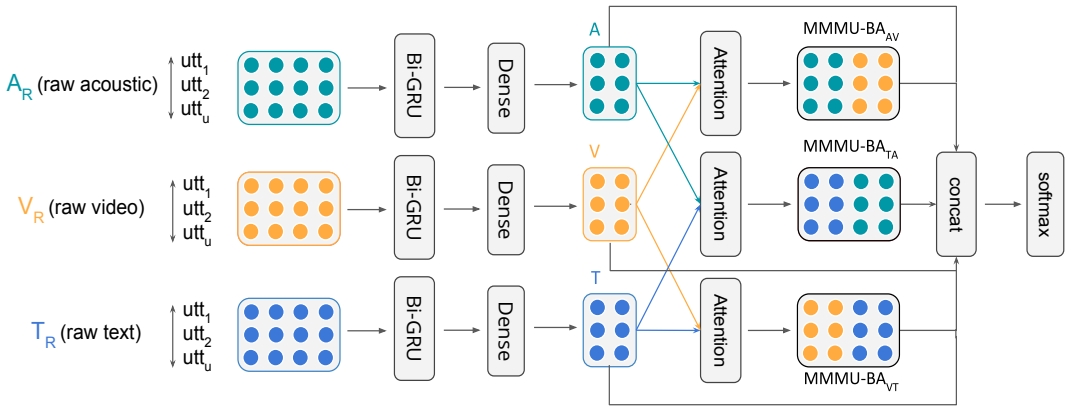


Figure 1: Overall architecture of the proposed MMMU-BA framework.

utterances. Finally, soft attention is applied over the multi-modal multi-utterance attention matrices to compute the modality-wise attentive representations (i.e. O_1 & O_2).

$$N_1(i,j) = \frac{e^{M_1(i,j)}}{\sum_{k=1}^{u} e^{M_1(i,k)}} \quad \text{for} \quad i,j = 1,..,u$$

$$N_2(i,j) = \frac{e^{M_2(i,j)}}{\sum_{k=1}^{u} e^{M_2(i,k)}} \quad \text{for} \quad i,j = 1,..,u.$$

$$O_1 = N_1.T \quad \& \quad O_2 = N_2.V$$

• ***Multiplicative Gating & Concatenation:*** Finally, a multiplicative gating function following (Dhingra et al., 2016) is computed between the multi-modal utterance specific representations of each individual modality and the other modalities. This element-wise matrix multiplication assists in attending to the important components of multiple modalities and utterances.

$$A_1 = O_1 \odot V \quad \& \quad A_2 = O_2 \odot T$$

Attention matrices A_1 & A_2 are then concatenated to obtain the *MMMU-BA*$_{VT}$ $\in \mathbb{R}^{u \times 2d}$ between *V* and *T*.

$$\textit{MMMU-BA}_{VT} = concat[A_1, A_2]$$

***MMMU-BA*$_{AV}$ & *MMMU-BA*$_{TA}$ computations:** Similar to *MMMU-BA*$_{VT}$, we follow the same procedure to compute *MMMU-BA*$_{AV}$ & *MMMU-BA*$_{TA}$. For a data source comprising of raw visual (V_R), acoustic (A_R) & text (T_R) modalities, at first, we compute the bi-modal attention pairs for each combination i.e. *MMMU-BA*$_{VT}$, *MMMU-BA*$_{AV}$ & *MMMU-BA*$_{TA}$. Finally, motivated by the residual skip connection network (He et al., 2016), we concatenate the bi-modal attention pairs with individual modalities (i.e. *V*, *A* & *T*) to boost the gradient flow to the lower layers. This concatenated feature is then used for final classification.

3.2 Multi-Modal Uni-Utterance - Self Attention (*MMUU-SA*) Framework

MMUU-SA framework does not account for information from the other utterances at the attention level, rather it utilizes multi-modal information of single utterance for predicting the sentiment. For a video having *'q'* utterances, *'q'* separate attention blocks are needed, where each block computes the self-attention over multi-modal information of a single utterance. Let $X_{u_p} \in \mathbb{R}^{3 \times d}$ is the information matrix of the p^{th} utterance where the three *'d'* dimensional rows are the outputs of the dense layers for the three modalities.

The attention matrix $A_{u_p} \in \mathbb{R}^{3 \times d}$ is computed separately for, p = 1^{st}, 2^{nd}, ... q^{th} utterances. Finally, for each utterance p, A_{u_p} and X_{u_p} are concatenated and passed to the output layer for classification. Please refer to the appendix for more details.

3.3 Multi-Utterance - Self Attention (*MU-SA*) Framework

In *MU-SA* framework, we apply self attention on the utterances of each modality separately, and use these for classification. In contrast to *MMUU-SA* framework, *MU-SA* utilizes the contextual information of the utterances at the attention level. Let, $T \in \mathbb{R}^{u \times d}$ (*text*), $V \in \mathbb{R}^{u \times d}$ (*visual*) and $A \in \mathbb{R}^{u \times d}$ (*acoustic*) are the outputs of the dense layers. For the three modalities, three separate attention blocks are required, where each block takes

multi-utterance information of a single modality and computes the self attention matrix. Attention matrices A_t, A_v and A_a are computed for text, visual and acoustic, respectively. Finally A_v, A_t, A_a, V, T & A are concatenated and passed to the output layer for classification.

4 Datasets, Experiments and Analysis

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

4.1 Datasets

We evaluate our proposed approach on two benchmark datasets, namely CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus (Zadeh et al., 2016) and the recently published CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Zadeh et al., 2018c). CMU-MOSI dataset consists of 93 videos spanning over 2199 utterances. Each utterance has a sentiment label associated with it. It has 52, 10 & 31 videos in training, validation & test set accounting for 1151, 296 & 752 utterances.

CMU-MOSEI has 3229 videos with 22676 utterances from more than 1000 online YouTube speakers. The training, validation & test set comprise of 16216, 1835 & 4625 utterances, respectively. More details about these datasets are presented in the appendix.

Each utterance in CMU-MOSI dataset has been annotated as either positive or negative, whereas in CMU-MOSEI dataset labels are in the continuous range of -3 to +3. However, in this work we project the instances of CMU-MOSEI in a *two-class classification* setup with $values \geq 0$ signify positive sentiments and $values < 0$ signify negative sentiments. We adopt such a strategy to be consistent with the previous published works on CMU-MOSI datasets (Poria et al., 2017b; Chen et al., 2017).

4.2 Feature extraction

We use the CMU-Multi-modal Data SDK[1] (Zadeh et al., 2018a) for feature extraction. For MOSEI dataset, *word-level* features were provided where *text* features were extracted by *GloVe embeddings*, *visual* features by *Facets*[2] & *acoustic* features by *CovaRep* (Degottex et al., 2014). Thereafter, we compute the average of *word-level* features in an utterance to obtain the *utterance-level* features. For each word, the dimension of the feature vector is set to 300 (*text*), 35 (*visual*) & 74 (*acoustic*).

In contrast, for MOSI dataset we use *utterance-level* features[3] provided in (Poria et al., 2017b). These *utterance-level* features represent the outputs of a convolutional neural network (Karpathy et al., 2014), 3D convolutional neural network (Ji et al., 2013) & openSMILE (Eyben et al., 2010) for *text*, *visual* & *acoustic* modalities, respectively. Dimensions of *utterance-level* features are 100, 100 & 73 for *text*, *visual* & *acoustic*, respectively.

4.3 Experiments

We evaluate our proposed approach for CMU-MOSI (test data) & CMU-MOSEI (dev data) [4]. Accuracy score is used as the evaluation metric.

We use Bi-directional GRUs having 300 neurons, each followed by a dense layer consisting of 100 neurons. Utilizing the dense layer, we project the input features of all the three modalities to the same dimensions. We set *dropout=0.5 (MOSI) & 0.3 (MOSEI)* as a measure of regularization. In addition, we also use *dropout=0.4 (MOSI) & 0.3 (MOSEI)* for the Bi-GRU layers. We employ *ReLu* activation function in the dense layers, and *softmax* activation in the final classification layer. For training the network we set the batch size=32, use *Adam* optimizer with *cross-entropy* loss function and train for 50 epochs. We report the average result of 5 runs for all our experiments.

We experiment with all the valid combinations of *uni-modal* (where only one modality is taken at a time), *bi-modal* (any two modalities are taken at a time) and *tri-modal* (all three modalities are taken at a time) inputs for *text*, *visual* and *acoustic*. In multi-modal attention frameworks i.e. *MMMU-BA & MMUU-SA*, the attention is computed over at least two modalities, hence, these two frameworks are not-applicable (*NA*) for *uni-modal* experiments in Table 1).

For MOSEI dataset, we obtain better performance with *text*. Subsequently, we take two modalities at a time for constructing bi-modal inputs and feed it to the network. For *text-acoustic* input pairs, we obtain the highest accuracies with 79.74%, 79.60% and 79.32% for *MMMU-BA*,

[1] https://github.com/A2Zadeh/CMU-MultimodalDataSDK

[2] https://pair-code.github.io/facets/

[3] https://github.com/SenticNet/contextual-sentiment-analysis

[4] Gold annotation of CMU-MOSEI test data wasn't released at the time of paper submission.

Modality	T	V	A	CMU-MOSEI			CMU-MOSI		
				MMMU-BA	*MMUU-SA*	*MU-SA*	*MMMU-BA*	*MMUU-SA*	*MU-SA*
Uni-modal	✓	-	-	*NA*	*NA*	78.23	*NA*	*NA*	80.18
	-	✓	-	*NA*	*NA*	74.84	*NA*	*NA*	63.70
	-	-	✓	*NA*	*NA*	75.88	*NA*	*NA*	62.10
Bi-modal	✓	✓	-	**79.40**	79.02	79.26	**81.51**	80.85	80.45
	✓	-	✓	**79.74**	79.60	79.32	**80.58**	80.31	79.78
	-	✓	✓	**76.66**	76.46	76.43	**65.16**	64.22	63.22
Tri-modal	✓	✓	✓	**79.80**	79.76	79.63	**82.31**	79.52	80.58

Table 1: Experimental results on CMU-MOSEI and CMU-MOSI datasets. *MMMU-BA* & *MMUU-SA* frameworks require atleast two modalities to compute the attentions, hence, these two frameworks are *not-applicable (NA)* for *uni-modal* inputs. All results are average of 5 runs with different random seeds. **T:** *Text*, **V:** *Visual*, **A:** *Acoustic*. Results are reported in accuracy.

MMUU-SA and *MU-SA* frameworks, respectively. The results that we obtain from the bi-modal combinations suggest that the *text-acoustic* combination is a better choice than the others as it improves the overall performance. Finally, we experiment with tri-modal inputs and observe an improved performance of 79.80%, 79.76% and 79.63% for *MMMU-BA*, *MMUU-SA* and *MU-SA* frameworks, respectively. This improvement entails that combination of all the three modalities is a better choice. The performance improvement was also found to be statistically significant (*T-test*) than the bi-modality and uni-modality inputs. Further, we observe that the *MMMU-BA* framework reports the best accuracy of 79.80% for the MOSEI dataset, thus supporting our claim that multi-modal attention framework (i.e. *MMMU-BA*) captures more information than the self-attention frameworks (i.e. *MMUU-SA* & *MU-SA*).

4.4 Analysis of Attention Mechanism

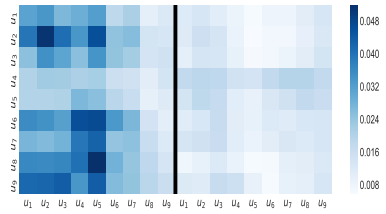
We analyze the attention values to understand the learning behavior of the proposed architecture. To illustrate, we take an example video from the CMU-MOSI test dataset. The transcript of the utterances for this particular video are presented in Table 2. The gold sentiments are positive for all the utterances except u_3 & u_4. We found that the proposed tri-modal *MMMU-BA* model predicts the labels of all the nine instances correctly, whereas other models make at least one misclassification. For the proposed tri-modal *MMMU-BA* model, the heatmaps of the pair-wise *MMMU-BA* softmax attention weights N_1 & N_2 of *visual-text*, *acoustic-visual* & *text-acoustic* are illustrated in Figure 2a, Figure 2b & Figure 2c, respectively. N_1 & N_2 are the softmax attention weights obtained from the pairwise matching matrices M_1 & M_2. Elements of the rows of N_1 & N_2 matrices signify different weights across multiple utterances. From the attention heatmaps, it is evident that by applying different weights across contextual utterances and modalities the model is able to predict labels of all the utterances correctly. All the heatmaps justify that the model learns to incorporate multi-modal & multi-utterance information and thus is able to correctly predict the labels of all the utterances. For example, heatmap of *MMMU-BA*$_{VT}$ (Figure 2a) signifies that elements of N_1 are weighted higher than N_2, and thus the model puts more attention on the *textual* part and relatively lesser on the *visual* part (as N_1 is multiplied with T & N_2 is multiplied with V). Also it can be concluded that textual features of the first few utterances are the most helpful compared to the rest of the textual features and visual features.

The softmax attention weights of *text* (N_t), *visual* (N_v) & *acoustic* (N_a) in tri-modal *MU-SA* model are illustrated in Figure 2d, Figure 2e & Figure 2f, respectively. The attention matrices are 9*9 dimensional. This model wrongly predicts the label of the utterance u_5. On the other hand, softmax attention weights in tri-modal *MMUU-SA* model are illustrated in Figure 2g. Nine separate attention weights (N_{u_1}, N_{u_2}, .., N_{u_9}) are computed for the nine utterances. This model wrongly predicts the labels of the utterances u_4 & u_5.

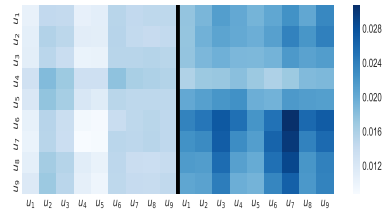
We further analyze our proposed architecture (i.e. *MMMU-BA*) *with* and *without* attention. In MOSI for tri-modal inputs, the *MMMU-BA* architecture reports a reduced accuracy of 80.89% without attention framework as compared to 82.31% with attention. We observe similar performance in the MOSEI dataset, where we obtain 79.02%

	Transcript	Gold Label	Predicted		
			MMMU-BA	*MMUU-SA*	*MU-SA*
u_1	*well he plays that well so he's good villain*	Positive	Positive	Positive	Positive
u_2	*he also has some really cool guns so that its like a desert eagle but it has to barrels to it*	Positive	Positive	Positive	Positive
u_3	*it's pretty pretty scary looking*	Negative	Negative	Negative	Negative
u_4	*i wouldn't want that pointed at me*	**Negative**	**Negative**	***Positive***	**Negative**
u_5	*and i who would i mean*	**Positive**	**Positive**	***Negative***	***Negative***
u_6	*um like i said i thought the movie was great*	Positive	Positive	Positive	Positive
u_7	*the action they do have is really well done*	Positive	Positive	Positive	Positive
u_8	*um they did a good job with car*	Positive	Positive	Positive	Positive
u_9	*they did a good job with fight scenes*	Positive	Positive	Positive	Positive

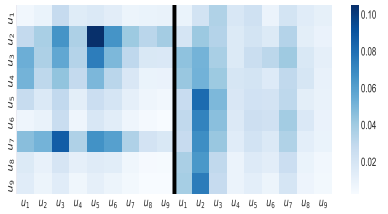
Table 2: Transcript, gold labels and predicted labels of a video in CMU-MOSI dataset having nine utterances. 7 utterances are labeled positive whereas 2 utterances are labeled negative. Predicted labels are for our different tri-modal models. Bolded labels are misclassified by at least one model.



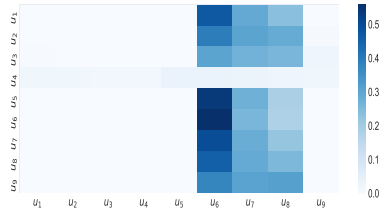
(a) *Softmax attention weights* N_1 *&* N_2 *for MMMU- BA*$_{VT}$.



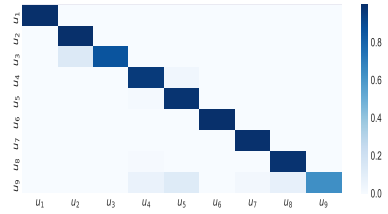
(b) *Softmax attention weights* N_1 *&* N_2 *for MMMU- BA*$_{AV}$.



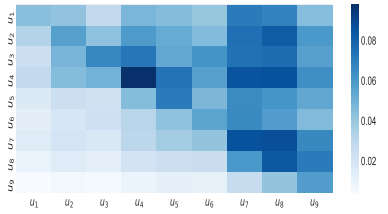
(c) *Softmax attention weights* N_1 *&* N_2 *for MMMU- BA*$_{TA}$.



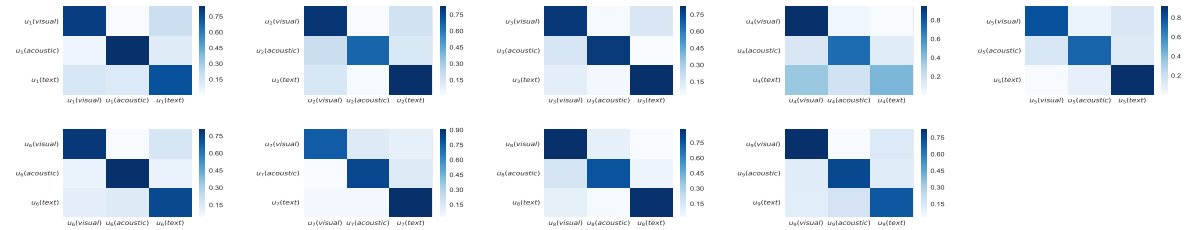
(d) *MU-SA*$_{text}$ matrix.



(e) *MU-SA*$_{visual}$ matrix.



(f) *MU-SA*$_{acoustic}$ matrix.



(g) Softmax attention weights (N_{u_1}, N_{u_2}, .., N_{u_9}) for *MMUU-SA* model.

Figure 2: (a), (b) & (c): Pair-wise softmax attention weights N_1 & N_2 of *visual-text*, *acoustic-visual* & *text-acoustic* in Tri-modal *MMMU-BA* model. Solid line at the center represents boundary of N_1 & N_2. The heatmaps represent attention weights of a particular utterance with respect to other utterances in N_1 & N_2. (d), (e) & (f) Softmax attention weights of *text* (N_t), *visual* (N_v) and *acoustic* (N_a) in Tri-modal MU-SA model. This model wrongly predicts the label of utterance u_5. (g) Softmax attention weights of the 9 utterances (N_{U_1}, N_{U_2}, .., N_{U_9}) in Tri-modal MMUU-SA model. This model wrongly predicts the label of utterance u_4 & u_5. The Tri-modal MMMU-BA model predicts all 9 instances correctly, whereas, the other two models makes at least one misclassification. Heatmap signifies that the model is able to predict labels of all the utterances correctly by incorporating multi-modal & multi-utterance information.

T	V	A	CMU-MOSEI		CMU-MOSI	
			w/ attention	*w/o attention*	*w/ attention*	*w/o attention*
✓	✓	-	**79.40**	78.27	**81.51**	80.71
✓	-	✓	**79.74**	78.12	**80.58**	80.18
-	✓	✓	**76.66**	76.32	**65.16**	63.69
✓	✓	✓	**79.80**	79.02	**82.31**	80.89

Table 3: Analysis of attention mechanism in *MMMU-BA* architecture. *w/ attention* $\rightarrow$ with multi-modal multi utterance attention mechanism and *w/o attention* $\rightarrow$ without attention mechanism.

accuracy without attention framework against 79.80% accuracy with attention framework. Statistical *T-test* shows these improvements to be significant. We also observed the similar trends for bi-modal inputs in both the datasets. All these experiments (c.f. Table 3) suggest that the attention framework is an important component in our proposed architecture, and in absence of this the network finds it more difficult for learning in all the cases (i.e. bi-modal & tri-modal input setups).

We successfully show that attention computation on pairwise combination of modalities (i.e. bi-modal attention framework) is more effective than the combination of self-attention on single modality. Further for the completeness of the proposed approach, we also experiment with tri-modal attention framework (attention is computed on three modalities at a time). Though the results that we obtain are convincing, it does not improve the performance over the bi-modal attention framework. We obtain the accuracies of 79.58% & 81.25% on MOSEI and MOSI, respectively, for the tri-modal attention framework.

4.5 Comparative Analysis

For MOSI datasets we compare the performance of our proposed approach with the the following state-of-the-art systems: i). Poria et al. (2017b)- LSTM-based sequence model to capture the contextual information of the utterances; ii). Poria et al. (2017c)- Tensor level fusion technique for combining all the three modalities; iii). Chen et al. (2017)-A gated multi-modal embedded LSTM with temporal attention (GME-LSTM(A)) for word-level fusion of multi-modality inputs. and iv). Zadeh et al. (2018a)- Multiple attention blocks for capturing the information across the three modalities.

In Table 4 we present the comparative performance between our proposed model and other state-of-the-art systems. In MOSI dataset, Poria et al. (2017b; 2017c) reported the accuracies of 80.3% & 81.3 %, respectively, utilizing tri-modal inputs. Zadeh et al. (2018a) obtained an accuracy of & 77.4%. Chen et al. (2017) reported accuracies of 75.7% (LSTM(A)) & 76.5% (GME-LSTM(A)) for two variants of their model. In contrast to the state-of-the-art systems, our proposed model attains an improved accuracy of 82.31% when we utilize all the three modalities, i.e. *text*, *visual* & *acoustic*. Our proposed system also obtains better performance as compared to the state-of-the-arts for bi-modal inputs.

For MOSEI dataset, we evaluate against the following systems: i) Poria et al. (2017b), ii) Zadeh et al. (2018a), and iii) Zadeh et al. (2018b), where authors proposed a memory fusion network for multi-view sequential learning. We evaluate the system of Poria et al. (2017b) on MOSEI dataset and obtain 77.64% accuracy with the tri-modal inputs. Authors in (Zadeh et al., 2018a) & (Zadeh et al., 2018b) reported the accuracy 76.0% and 76.4%, respectively, with the tri-modal inputs. In comparison, our proposed approach yields an accuracy of 79.80%. As reported in Table 4 the proposed approach also attains better performance for all the bi-modal and uni-modal input combinations when compared to Poria et al. (2017b).

As reported in Table 4, we observe that the performance achieved in our proposed approach is significantly better in comparison to the state-of-the-art systems with p-value< 0.05 (obtained using T-test). For further analysis, we also report results for *three-class classification* (positive, neutral & negative classes) problem setup for MOSEI dataset in Table 7. Note that this setup is not feasible in MOSI as labels are only positive or negative.

4.6 Error Analysis

We perform error analysis on the predictions of our proposed *MMMU-BA* model with all the three input sources. Confusion matrices for both the datasets are demonstrated in Table 5. For MOSEI dataset we observe that the precision and recall for positive class (84% *precision* & 88% *recall*; are quite encouraging. However, the same are comparatively on the lower side for the negative class (68% *precision* & 58% *recall*. In contrast, for the MOSI dataset - which is relatively balanced - we obtain quite similar performance for both the classes i.e. positive (86% *precision* & 85% *recall*) and negative (77% *precision* & 75% *recall*). Please

Modality	T	V	A	CMU-MOSEI				CMU-MOSI					
				Poria et al. (2017b)	Zadeh et al. (2018a)	Zadeh et al. (2018b)	Proposed	Poria et al. (2017b)	Poria et al. (2017c)	Chen et al. (2017) LSTM(A)	Chen et al. (2017) GME-LSTM(A)	Zadeh et al. (2018a)	Proposed
Uni-modal	✓	-	-	76.75	-	-	**78.23**	78.1	-	71.3	-	-	**80.18**
	-	✓	-	71.84	-	-	**74.84**	60.3	-	52.3	-	-	63.70
	-	-	✓	70.94	-	-	**75.88**	55.8	-	55.4	-	-	62.10
Bi-modal	✓	✓	-	77.03	-	-	**79.40**	79.3	79.9	74.3	-	-	**81.51**
	✓	-	✓	76.89	-	-	**79.74**	80.2	80.1	73.5	-	-	**80.58**
	-	✓	✓	72.74	-	-	**76.66**	62.1	62.9	-	-	-	**65.16**
Tri-modal	✓	✓	✓	77.64	76.0	76.4	**79.80**	80.3	81.3	75.7	76.5	77.4	**82.31**
T-test (p-values)				-	-	-	*0.0025*	-	-	-	-	-	*0.0006*

Table 4: Comparative analysis of the proposed approach with recent state-of-the-art systems. Significance T-test p-values < 0.05

MOSEI

102	234
1230	269

Positive Negative

63	215
404	70

Positive Negative

MOSI

Table 5: Confusion matrix for tri-modal *MMMU-BA*.

	Text	Actual	Predicted	Possible Reason
MOSI	*At first I thought the movie would appeal more to younger audience.*	negative	positive	Implicit sentiment.
	Its really non-stop from beginning to end.	negative	positive	
	But its action isn't particularly memorable.	negative	positive	Negation & strong word.
	I mean I don't regret seeing it.	positive	negative	
	Um I was really looking forward to it.	negative	positive	Sarcastic sentence.
MOSEI	*And when I was going to school it was really difficult for me to find avenues and resources to be able to reach higher education.*	negative	positive	Implicit sentiment.
	We could have a decision from the court on the stay any day now.	positive	negative	
	Holidays never really happen in online courses I guess.	negative	positive	Negation & strong word.
	Young people dropping out of the labour market are actually not counted anymore as unemployed as they are inactive.	positive	negative	
	Thank you for your efforts and consideration.	negative	positive	Sarcastic sentence.

Table 6: **Error Analysis:** Frequent error cases and their possible reasons of failure for the tri-modal *MMMU-BA* framework.

Metric	CMU-MOSEI	
	Poria et al. (2017b)	MMMU-BA (Tri-modal)
Accuracy	61.89	**63.30**
F1 Score	61.60	**63.07**

Table 7: Three class (positive, negative, neutral) classification results in MOSEI dataset.

refer to the appendix for PR curves of different input combinations.

We further analyze our outputs qualitatively and list a few frequently occurring error categories with examples in Table 6.

5 Conclusion

In this paper, we have proposed a recurrent neural network based multi-modal attention framework that leverages the contextual information for utterance-level sentiment prediction. The network learns on top of three modalities, *viz.* text, visual and acoustic, considering sequence of utterances in a video. Through evaluation results on two benchmark datasets (one being the popular & commonly used (MOSI) and other being the most recent & largest (MOSEI) dataset for multi-modal sentiment analysis), we successfully showed that the proposed attention based framework performs better than various state-of-the-art systems.

In future, we would like to investigate new techniques, and explore the ways to handle implicit sentiment and sarcasm. Future direction of work also include adding more dimensions, e.g. emotion analysis & intensity prediction.

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, Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, 125:116 – 135.

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A Hybrid Deep Learning Architecture for Sentiment Analysis. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers, December 11-16, 2016*, pages 482–493, Osaka, Japan.

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*, pages 163–171. ACM.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. 2014. Covarep - a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

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

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

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*, pages 770–778.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th KDD*, pages 168–177, Seattle, WAs.

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):221–231.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA. Association for Computational Linguistics.

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 Interaction, ICMI 2011, Alicante, Spain, November 14-18, 2011*, pages 169–176.

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 on Association for Computational Linguistics*, ACL '05, pages 115–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

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, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017c. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 1033–1038. IEEE.

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. 2017d. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 261:217–230.

P. D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Association for Computational Linguistics (ACL)*, pages 417–424, Philadelphia, USA.

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. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Memory Fusion Network for Multi-view Sequential Learning. *arXiv preprint arXiv:1802.00927*.

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.

7 Appendix

7.1 Multi-Modal Uni-Utterance - Self Attention (*MMUU-SA*) Framework

Let $X_p \in \mathbb{R}^{3\times d}$ is the information matrix of the p^{th} utterance where the three *'d'* dimensional rows are the outputs of the time-distributed dense layer for the three modalities. Computation in the p^{th} attention block proceeds as follows:

$$
\begin{aligned}
M_{u_p} &= X_{u_p}.X_{u_p}^T \\
N_{u_p}(i,j) &= \frac{e^{M_{up}(i,j)}}{\sum_{k=1}^{3} e^{M_{up}(i,k)}} \quad \text{for} \quad i,j = 1,2,3; \\
O_{u_p} &= N_{u_p}.X_{u_p} \\
A_{u_p} &= O_{u_p} \odot X_{u_p}
\end{aligned}
$$

The attention matrix $A_{u_p} \in \mathbb{R}^{3\times d}$ is computed separately for, p = 1^{st}, 2^{nd}, ... q^{th} utterances. Finally, for each utterance p, A_{u_p} and X_{u_p} are concatenated and passed to the output layer for classification.

7.2 Multi-Utterance - Self Attention (*MU-SA*) Framework

In *MU-SA* framework, we apply self attention on the utterances of each modality separately, and use these for classification. In contrast to *MMUU-SA* framework, *MU-SA* utilizes the contextual information of the utterances at the attention level. Let, $T \in \mathbb{R}^{u\times d}$ (*text*), $V \in \mathbb{R}^{u\times d}$ (*visual*) and $A \in \mathbb{R}^{u\times d}$ (*acoustic*) are the outputs of the dense layers. For the three modalities, three separate attention blocks are required, where each block takes multi-utterance information of a single modality and computes the self attention matrix. Specifically, the *MU-SA* attention (A_v) on V (visual) will be computed as follows,

$$
\begin{aligned}
M_v &= V.V^T \\
N_v(i,j) &= \frac{e^{M_v(i,j)}}{\sum_{k=1}^{u} e^{M_v(i,k)}} \quad \text{for} \quad i,j = 1,..,u \\
O_v &= N_v.V \\
A_v &= O_v \odot V
\end{aligned}
$$

The attention matrix $A_p \in \mathbb{R}^{3\times d}$ is computed for $p = 1^{st}, 2^{nd}, ...u^{th}$ utterances. Finally, for each utterance u, A_p and X_p are concatenated and passed to the output layer with softmax activation for classification.

7.3 Dataset Statistics

Dataset statistics are presented in Table 8.

7.4 Attention Computation

MMMU-BA$_{VT}$ attention computation is illustrated in Figure 3.

Statistics	CMU-MOSI			CMU-MOSEI		
	Tr	Dv	Ts	Tr	Dv	Ts
#Videos	52	10	31	2250	300	679
#Utterance	1151	296	752	16216	1835	4625
#Utterance/Video - Min	9	9	10	1	1	1
#Utterance/Video - Max	63	34	43	98	37	52
#Utterance/Video - Avg	24.692	22.9	22.129	7.207	6.116	6.821
#Positive	556	153	467	11498	1332	3281
#Negative	595	143	285	4718	503	1344
#Words/Utter. - Min	1	1	1	1	1	1
#Words/Utter. - Max	99	44	108	515	224	549
#Words/Utter. - Avg	11.533	10.786	13.176	18.227	18.498	18.658
#Utter-Len/Video - Min	0.219s	0.648s	0.229s	0.089s	0.22s	0.15s
#Utter-Len/Video - Max	38.233s	13.599s	31.957s	208.27s	90.42s	188.22s
#Utter-Len/Video - Avg	3.635s	3.538s	4.536s	6.896s	6.960s	7.158s
#Speakers		89			1000	

(a) Data Statistics. ***Tr***→Train set; ***Dv***→Development set; ***Ts***→Test set;

Table 8: Dataset statistics for MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018c).

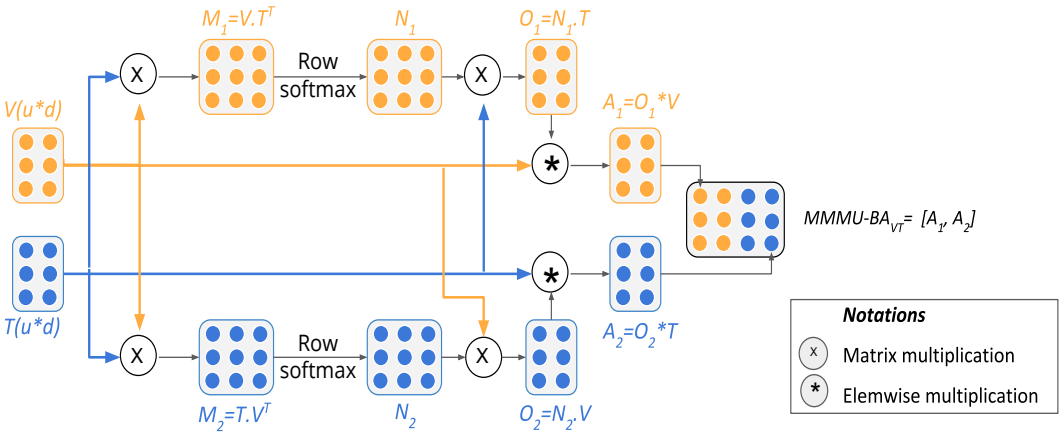


Figure 3: *MMMU-BA*$_{VT}$ attention computation.

7.5 Precision-Recall (PR) curve

We illustrate the precision, recall & f-measure for different input combinations in Figure 4 & Figure 5.

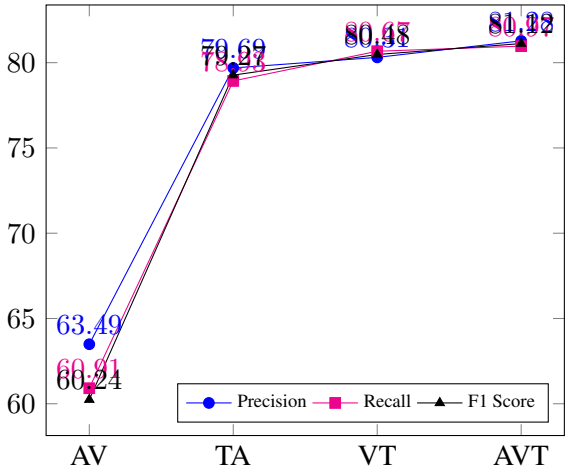


Figure 4: *Precision*, *Recall* & *F-measure* for different input combinations in *MMMU-BA* architecture of MOSI dataset.

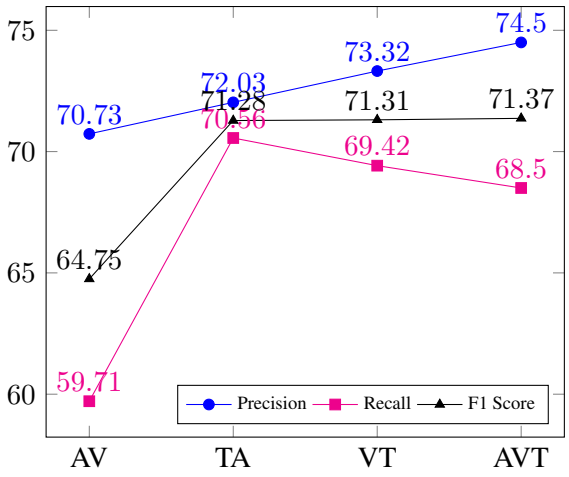


Figure 5: *Precision*, *Recall* & *F-measure* for different input combinations in *MMMU-BA* architecture of MOSEI dataset.