Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis

Md Shad Akhtar[†], **Dushyant Singh Chauhan**[†], **Deepanway Ghosal**[†], **Soujanya Poria**[+], **Asif Ekbal**[†] **and Pushpak Bhattacharyya**[†]
[†] Department of Computer Science & Engineering
Indian Institute of Technology Patna, India
{shad.pcs15, 1821CS17, deepanway.ee14, asif, pb}@iitp.ac.in
[+] School of Computer Science and Engineering, Nanyang Technological University, Singapore
sporia@ntu.edu.sg

Abstract

Related tasks often have inter-dependence on each other and perform better when solved in a joint framework. In this paper, we present a deep multi-task learning framework that jointly performs sentiment and emotion analysis both. The multi-modal inputs (i.e., *text*, *acoustic* and *visual frames*) of a video convey diverse and distinctive information, and usually do not have equal contribution in the decision making. We propose a context-level inter-modal attention framework for simultaneously predicting the sentiment and expressed emotions of an utterance. We evaluate our proposed approach on CMU-MOSEI dataset for multi-modal sentiment and emotion analysis. Evaluation results suggest that multi-task learning framework offers improvement over the single-task framework. The proposed approach reports new state-of-the-art performance for both sentiment analysis and emotion analysis.

1 Introduction

With the rapid growth of social media video platforms such as Youtube, Vimeo, users now tend to upload videos on these platforms. Such video platforms offer users an opportunity to express their opinions on any topic. Videos usually consist of *audio* and *visual* modalities, and thus can be considered as a source of multi-modal information. Although videos contain more information than text, fusing multiple modalities is a major challenge. A common practice in sentiment analysis and emotion recognition or affective computing, in general, is to analyze textual opinions. However, in recent days multi-modal affect analysis has gained a major attention (Poria et al., 2017b, 2016). In these works, in addition to the *visual frames*, other sources of information such as *acoustic* and *textual* (transcript) representation of the spoken languages are also incorporated in the analysis. Multi-modal analysis (e.g. sentiment analysis Zadeh et al. 2018c, emotion recognition Poria et al. 2016, question-answering Teney et al. 2017 etc.) is an emerging field of study, that utilizes multiple information sources for solving a problem. These sources (e.g., text, visual, acoustic, etc.) offer a diverse and often distinct piece of information that a system can leverage on. For example, *'text'* carries semantic information of the spoken sentence, whereas '*acoustic*' information reveals the emphasis (pitch, voice quality) on each word. In contrast, the '*visual*' information (image or video frame) extracts the gesture and posture of the speaker.

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 and their available multi-modal contents, an interdisciplinary study involving *text*, *acoustic* and *visual* features have drawn significant interest among the research community. Effectively fusing this diverse information is non-trivial and poses several challenges to the underlying problem.

In our current work, we propose a multi-task model to extract both sentiment (i.e. ***positive*** or ***negative***) and emotion (i.e. ***anger***, ***disgust***, ***fear***, ***happy***, ***sad*** or ***surprise***) of a speaker in a video. In multi-task framework, we aim to leverage the inter-dependence of these two tasks to increase the confidence of individual task in prediction. For e.g., information about *anger* emotion can help in prediction of *negative* sentiment and vice-versa.

A speaker can utter multiple utterances (a unit of speech bounded by breathes or pauses) in a single video and these utterances can have different sentiments and emotions. We hypothesize that the sentiment (or, emotion) of an utterance often has inter-dependence on other contextual utterances i.e. the knowledge of sentiment (or, emo-

Proceedings of NAACL-HLT 2019, pages 370–379
Minneapolis, Minnesota, June 2 - June 7, 2019.

tion) for an utterance can assist in classifying its neighbor utterances. We utilize all three modalities (i.e. *text*, *acoustic* and *visual*) for the analysis. Although all these sources of information are crucial, they are not equally beneficial for each individual instance. Few examples are presented in Table 1. In the first example, *visual frames* provide important clues than *textual* information for finding the sentiment of a sarcastic sentence "*Thanks for putting me on hold! I've all the time in the world.*". Similarly, the *textual* representation of second example "*I'm fine.*' does not reveal the exact emotion of a *sad* person. For this particular case, *acoustic* or *visual* information such as low tone voice, facial expression etc. have bigger role to play for the classification.

Utterance	Feeling	**T**	**A**	**V**
Thanks for putting me on hold! I've all the time in the world.	Sentiment (Negative)	-	-	✓
I'm fine.	Emotion (Sad)	-	✓	✓

Table 1: Contributing modalities for different scenario. Tick represents the most contributing information.

2 Problem Definition

Multi-task learning paradigm provides an efficient platform for achieving generalization. Multiple tasks can exploit the inter-relatedness for improving individual performance through a shared representation. Overall, it provides three basic advantages over the single-task learning paradigm a). it helps in achieving generalization for multiple tasks; b). each task improves its performance in association with the other participating tasks; and c). offers reduced complexity because a single system can handle multiple problems/tasks at the same time.

Sentiments (Pang et al., 2005) and emotions (Ekman, 1999) are closely related. Most of the emotional states have clear distinction of being a positive or negative situation. Emotional states e.g. '*anger*', '*fear*', '*disgust*', '*sad*' etc. belong to negative situations, whereas '*happy*' and '*surprise*' reflect the positive situations. Motivated by the association of sentiment & emotion and the advantages of the multi-task learning paradigm, we present a multi-task framework that jointly learns and classifies the sentiments and emotions in a video. As stated earlier, contextual-utterances and/or multi-modal information provide important cues for the classification. Our proposed approach applies attention over both of these sources of information simultaneously (i.e., contextual utterance and inter-modal information), and aims to reveal the most contributing features for the classification. We hypothesize that applying attention to contributing neighboring utterances and/or multimodal representations may assist the network to learn in a better way.

Our proposed architecture employs a recurrent neural network based contextual inter-modal attention framework. In our case, unlike the previous approaches, that simply apply attention over the contextual utterance for classification, we take a different approach. Specifically, we attend over the contextual utterances by computing correlations among the modalities of the target utterance and the context utterances. This particularly helps us to distinguish which modalities of the relevant contextual utterances are more important for the classification of the target utterance. The model facilitates this modality selection process 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 the recent benchmark dataset of CMU-MOSEI (Zadeh et al., 2018c). It is the largest available dataset (approx. 23K utterances) for multi-modal sentiment and emotion analysis (c.f. Dataset Section). The evaluation shows that contextual inter-modal attention framework attains better performance than the state-of-the-art systems for various combinations of input modalities.

The main contributions of our proposed work are three-fold: **a)** *we leverage the inter-dependence of two related tasks (i.e. sentiment and emotion) in improving each others performance using an effective multi-modal framework*; **b)** *we propose contextual inter-modal attention mechanism that facilitates the model to assign weightage to the contributing contextual utterances and/or to different modalities simultaneously*. Suppose, to classify an utterance '*u1*' of 5 utterances video, visual features of '*u2*' & '*u4*', acoustic features of '*u3*' and textual features of '*u1*', '*u3*' & '*u5*' are more important than others. Our attention model is capable of highlighting such diverse contributing features; and **c)** *we present the state-of-the-arts for both sentiment and emotion predictions*.

3 Related Work

A survey of the literature suggests that multi-modal sentiment prediction is a relatively new area as compared to textual based sentiment prediction (Morency et al., 2011; Poria et al., 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).

Zadeh et al. (2016) introduced the multi-modal dictionary to understand the interaction between facial gestures and spoken words better when expressing sentiment. In another work, 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). Authors reported improved accuracy using multi-modality on the CMU-MOSI dataset. These works did not take contextual information into account. Poria et al. (2017b) proposed a Long Short Term Memory (LSTM) based framework for sentiment classification that leverages the contextual information to capture the inter-dependencies between the utterances. Zadeh et al. (2018a) proposed multi-attention blocks (MAB) to capture the information across the three modalities (text, visual and acoustic) for predicting the sentiments. Authors evaluated their approach on the different datasets and reported improved accuracies in the range of 2-3% over the state-of-the-art models. Blanchard et al. (2018) proposed a multi-modal fusion model that exclusively uses high-level visual and acoustic features for sentiment classification.

An application of multi-kernel learning based fusion technique was proposed in (Poria et al., 2016), where the authors employed deep convolutional neural network (CNN) for extracting the textual features and fused it with other modalities (*visual & acoustic*) for emotion prediction. Ranganathan et al. (2016) proposed a convolutional deep belief network (CDBN) models for multi-modal emotion recognition. The author used CDBN to learn salient multi-modal (acoustic and visual) features of low-intensity expressions of emotions. Hazarika et al. (2018) introduced a self- attention mechanism for multi-modal emotion detection by feature level fusion of text and speech. Recently, Zadeh et al. (2018c) introduced the CMU-MOSEI dataset for multi-modal sentiment analysis and emotion recognition. They effectively fused the tri-modal inputs through a dynamic fusion graph and also reported competitive performance *w.r.t.* various state-of-the-arts on MOSEI dataset for both sentiment and emotion classification.

The main difference between the proposed and existing methods is contextual inter-modal attention. Systems (Poria et al., 2016; Zadeh et al., 2016, 2017; Blanchard et al., 2018) do not consider context for the prediction. System (Poria et al., 2017b) uses contextual information for the prediction but without any attention mechanism. In contrast, (Zadeh et al., 2018a) uses multi-attention blocks but did not account for contextual information. Our proposed model is novel in the sense that our approach applies attention over multi-modal information of the contextual utterances in a single step. Thus, it ensures to reveal the contributing features across *multiple modalities* and *contextual utterances* simultaneously for sentiment and emotion analysis. Further, to the best of our knowledge, this is the first attempt at solving the problems of multi-modal sentiment and emotion analysis together in a multi-task framework.

The contextual inter-modal attention mechanism is not much explored in NLP domains as such. We found one work that accounts for bi-modal attention for visual question-answering (VQA) (Teney et al., 2017). However, its attention mechanism differs from our proposed approach in the following manner: a) VQA proposed question guided image-attention, but our attention mechanism attends multi-modalities; b) attention is applied over different positions of the image, whereas our proposed approach applies attention over multiple utterances and two-modalities at a time; c). our proposed attention mechanism attends a sequence of utterances (text, acoustic or visual), whereas VQA applies attention in the spatial domain. In another work, Ghosal et al. (2018) proposed an inter-modal attention framework for the multi-modal sentiment analysis. However, the key differences with our current work are as follows: a) Ghosal et al. (2018) addressed only sentiment analysis, whereas, in our current work, we address both the sentiment and emotion analysis; b) Ghosal et al. (2018) handles only sentiment analysis in single task learning framework, whereas our proposed approach is based on multi-task learning framework, where we solve two tasks, i.e., sentiment analysis and emotion analysis, together in a single network; c) we perform detailed com-

parative analysis over the single-task vs. multi-task learning; and d) we present state-of-the-art for both sentiment and emotion analysis.

4 Multi-task Multi-modal Emotion Recognition and Sentiment Analysis

In our proposed framework, we aim to leverage multi-modal and contextual information for predicting sentiment and emotion of an utterance simultaneously in a multi-task learning framework. As stated earlier, a video consists of a sequence of utterances and their semantics often have inter-dependencies on each other. We employ three bi-directional Gated Recurrent Unit (bi-GRU) network for capturing the contextual information (i.e., one for each modality). Subsequently, we introduce pair-wise inter-modal attention mechanism (i.e. *visual-text*, *text-acoustic* and *acoustic-visual*) to learn the joint-association between the multiple modalities & utterances. The objective is to emphasize on the contributing features by putting more attention to the respective utterance and neighboring utterances. Motivated by the residual skip connection (He et al., 2016) the outputs of pair-wise attentions along with the representations of individual modalities are concatenated. Finally, the concatenated representation is shared across the two branches of our proposed network- corresponding to two tasks, i.e., sentiment and emotion classification for prediction (one for each task in the multi-task framework). Sentiment classification branch contains a *softmax* layer for final classification (i.e. *positive* & *negative*), whereas for emotion classification we use sigmoid layer. The shared representation will receive gradients of error from both the branches (sentiment & emotion) and accordingly adjust the weights of the models. Thus, the shared representations will not be biased to any particular task, and it will assist the model in achieving generalization for the multiple tasks. Empirical evidences support our hypothesis (c.f. Table 4).

4.1 Contextual Inter-modal (CIM) Attention Framework

Our contextual inter-modal attention framework works on a pair of modalities. At first, we capture the cross-modality information by computing a pair of matching matrices $M_1, M_2 \in \mathbb{R}^{u \times u}$, where *'u'* is the number of utterances in the video. Further, to capture the contextual dependencies, we compute the probability distribution scores ($N_1, N_2 \in \mathbb{R}^{u \times u}$) over each utterance of cross-modality matrices M_1, M_2 using a softmax function. This essentially computes the attention weights for contextual utterances. Subsequently, we apply soft attention over the contextual inter-modal matrices to compute the modalitiy-wise attentive representations ($O_1 \& O_2$). Finally, a multiplicative gating mechanism (Dhingra et al., 2016) ($A_1 \& A_2$) is introduced to attend the important components of multiple modalities and utterances. The concatenated attention matrix of $A_1 \& A_2$ then acts as the output of our contextual inter-modal attention framework. The entire process is repeated for each pair-wise modalities i.e. *text-visual*, *acoustic-visual* and *text-acoustic*. We illustrate and summarize the proposed methodology in Figure 1 and Algorithm 1, respectively.

Algorithm 1 Multi-task Multi-modal Emotion and Sentiment (MTMM-ES)

procedure MTMM-ES(t, v, a)
- $d \leftarrow 100$ ▷ GRU dimension
- $T \leftarrow biGRU_T(t, d)$
- $V \leftarrow biGRU_V(v, d)$
- $A \leftarrow biGRU_A(a, d)$
- $Atn_{TV} \leftarrow$ ***CIM-Attention***(T, V)
- $Atn_{AV} \leftarrow$ ***CIM-Attention***(A, V)
- $Atn_{TA} \leftarrow$ ***CIM-Attention***(T, A)
- $Rep \leftarrow [Atn_{TV}, Atn_{AV}, Atn_{TA}, T, V, A]$
- $polarity \leftarrow Sentiment(Rep)$
- $emotion \leftarrow Emotion(Rep)$
- **return** $polarity, emotion$

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

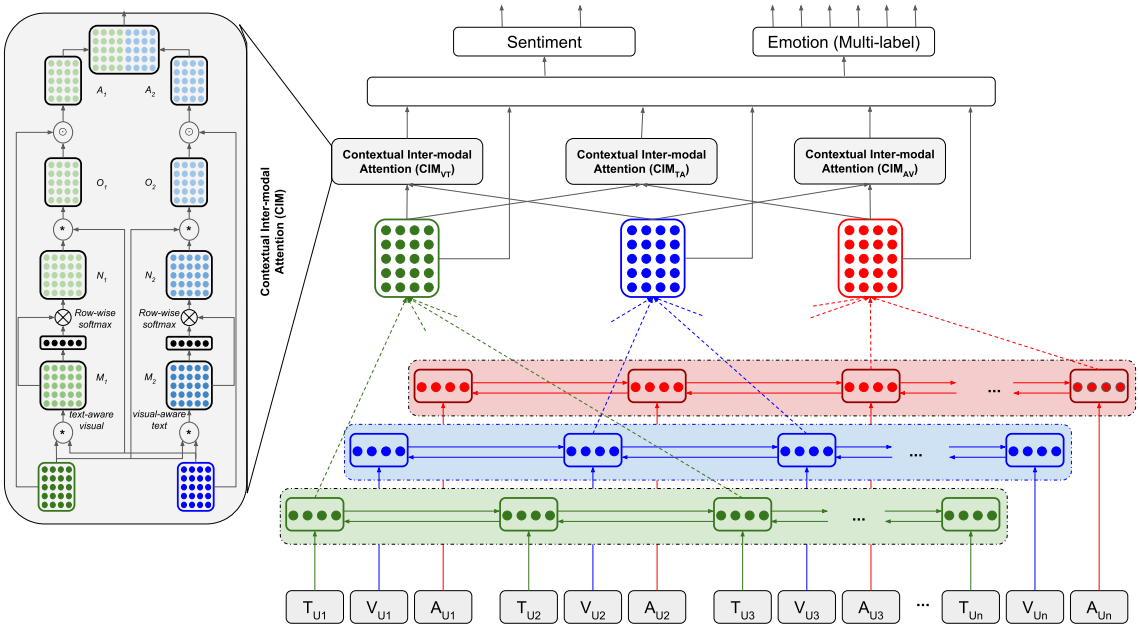


Figure 1: Overall architecture of the proposed framework. Contextual inter-modal (CIM) attention computation between *visual* and *text* modality.

5 Datasets, Experiments, and Analysis

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

5.1 Datasets

We evaluate our proposed approach on the benchmark datasets of sentiment and emotion analysis, namely CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Zadeh et al., 2018c). CMU-MOSEI dataset consists of 3,229 videos spanning over 23,000 utterances from more than 1,000 online YouTube speakers. The training, validation & test set comprises of 16216, 1835 & 4625 utterances, respectively.

Each utterance has six emotion values associated with it, representing the degree of emotion for *anger*, *disgust*, *fear*, *happy*, *sad* and *surprise*. Emotion labels for an utterance are identified as all non-zero intensity values, i.e. if an utterance has three emotions with non-zero values, we take all three emotions as multi-labels. Further, an utterance that has no emotion label represents the absence of emotion. For experiments, we adopt 7-classes (6 *emotions* + 1 *no emotion*) and pose it as multi-label classification problem, where we try to optimize the binary-cross entropy for each of the class. A brief statistics for multi-label emotions is presented in Table 3. In contrast, the sentiment values for each utterance are disjoint, i.e. *value* < 0 and *value* ≥ 0 represent the negative and positive sentiments, respectively. A detailed statistics of the CMU-MOSEI dataset is shown in Table 2.

Statistics	Train	Dev	Test
#Videos	2250	300	679
#Utterance	16216	1835	4625
#Positive	11499	1333	3281
#Negative	4717	502	1344
#Anger	3506	334	1063
#Disgust	2946	280	802
#Fear	1306	163	381
#Happy	8673	978	2484
#Sad	4233	511	1112
#Surprise	1631	194	437
#Speakers		1000	

Table 2: Dataset statistics for CMU-MOSEI. Each utterance contains multi-modal information.

Single emotion	11050	Two emotions	5526
Three emotions	2084	Four emotions	553
Five emotions	84	Six emotions	8
No emotion	3372		

Table 3: Statistics of multi-label emotions.

Tasks		F1-score							Acc (Sentiment) & Weighted-Acc (Emotion)						
		T	A	V	T+V	T+A	A+V	T+A+V	T	A	V	T+V	T+A	A+V	T+A+V
Sent	STL	75.1	67.9	66.3	77.0	76.5	69.6	77.6	78.2	74.8	75.8	79.4	79.7	76.6	79.8
	MTL	77.5	72.1	69.1	78.7	78.6	75.8	78.8	79.7	75.7	76.5	80.4	80.2	77.4	80.5
Emo	STL	75.9	72.3	73.6	77.5	76.8	76.0	77.7	58.0	56.7	53.7	60.1	59.6	58.0	60.8
	MTL	76.9	74.6	75.4	78.5	77.6	77.0	78.6	60.2	56.2	57.5	62.5	60.5	59.3	62.8

Table 4: Single-task learning (STL) and Multi-task (MTL) learning frameworks 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).

5.2 Feature extraction

We use the CMU-Multi-modal Data SDK[1] for downloading and feature extraction. The dataset was pre-tokenized and a feature vector was provided for each word in an utterance. The *textual*, *visual* and *acoustic* features were extracted by *GloVe* (Pennington et al., 2014), *Facets*[2] & *CovaRep* (Degottex et al., 2014), respectively. Thereafter, we compute the average of *word-level* features to obtain the *utterance-level* features.

5.3 Experiments

We evaluate our proposed approach on the datasets of CMU-MOSEI. We use the Python based Keras library for the implementation. We compute *F1-score* and *accuracy values* for sentiment classification and *F1-score* and *weighted accuracy* (Tong et al., 2017) for emotion classification. 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). To obtain multi-labels for emotion classification, we use 7 sigmoid neurons (corresponds to 6 emotions + 1 no-emotion) with binary cross-entropy loss function. Finally, we take all the emotions whose respective values are above a $threshold$. We optimize and cross-validate both the evaluation metrics (i.e. F1- score and weighted accuracy) and set the *threshold* as 0.4 & 0.2 for F1-score and weighted accuracy, respectively. We show our model configurations in Table 5.

As stated earlier, our proposed approach requires at least two modalities to compute bi-modal attention. Hence, we experiment with bi-modal and tri-modal input combinations for the proposed approach i.e. taking *text-visual*, *text-acoustic*, *acoustic-visual* and *text-visual-acoustic* at a time. For completeness (i.e., uni-modal inputs), we also experiment with a variant of the proposed approach where we apply self-attention on the utterances of each modality separately. The self-attention unit utilizes the contextual information of the utterances (i.e., it receives $u \times d$ hidden representations), applies attention and forward it to the output layer for classification. We report the experimental results of both single-task (STL) and multi-task (MTL) learning framework in Table 4. In the single-task framework, we build separate systems for sentiment and emotion analysis, whereas in multi-task framework a joint-model is learned for both of these problems. For sentiment classification, our single-task framework reports an F1-score of 77.67% and accuracy value of 79.8% for the tri-modal inputs. Similarly, we obtain 77.71% F1-score and 60.88% weighted accuracy for emotion classification.

Comparatively, when both the problems are learned and evaluated in a multi-task learning framework, we observe performance enhancement for both sentiments as well as emotion classification. MTL reports 78.86% F1-score and 80.47% accuracy value in comparison to 77.67% and 79.8% of STL with tri-modal inputs, respectively. For emotion classification, we also observe an improved F-score (78.6 (MTL) vs. 77.7 (STL)) and weighted accuracy (62.8 (MTL) vs. 60.8 (STL))

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

Table 5: Model configurations

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

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

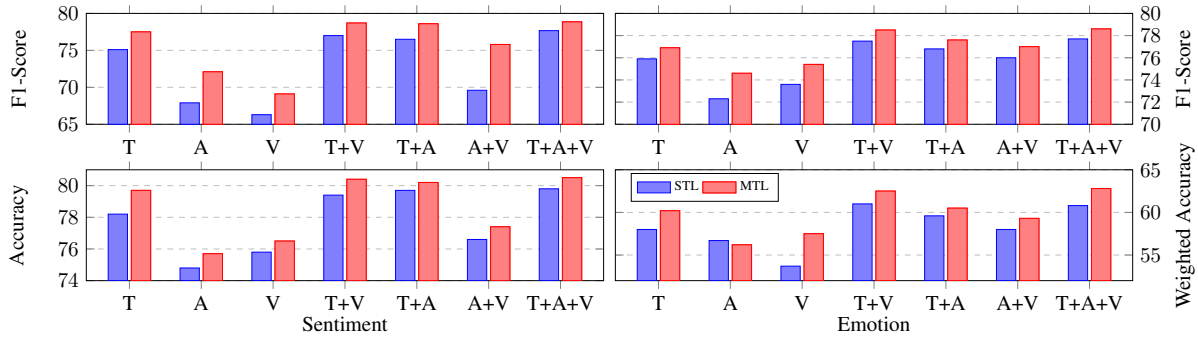
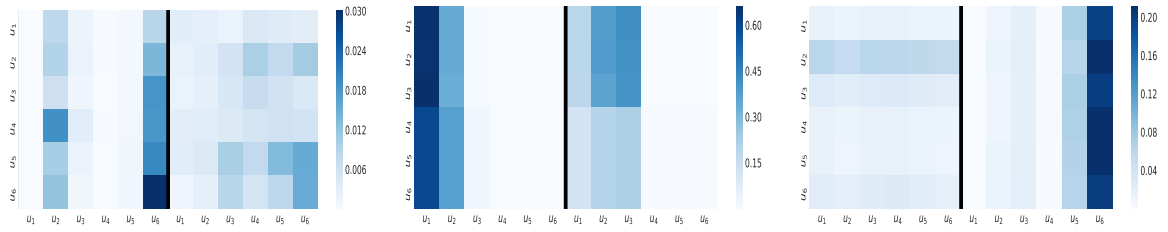


Figure 2: Single-task learning (STL) and Multi-task (MTL) learning frameworks for the proposed approach.

	Utterances	Sentiment		Emotion	
		Actual	MTL	Actual	MTL
1	*line hello my name is sarah and i will be doing my video opinion on the movie shall we dance uhh starring jennifer*	Pos	Pos	Anger	Anger
2	*richard gere and susan umm you i really didn't enjoy this movie at all it kinda boring for*	Neg	Neg	Anger, Disgust	Anger, Disgust, Happy, Sad
3	*for umm it kinda felt as if there were parts in there they*	Pos	Neg	No class	Anger, Disgust, Happy, Sad
4	*they just put in there to kinda pass the time on basically the movie is about umm richard character and him being a*	Pos	Pos	Happy	Anger, Disgust, Happy, Sad
5	*umm looking for some some stutter extra sizzle to add into his life he meets up with a dance instructor who is played by jennifer lopez and basically she convinces him to sign up for some ballroom he gets into it he enjoys it a lot but still a secret from his family*	Pos	Pos	Anger	Anger Disgust, Happy, Sad
6	*family he is trying to cope with having this this stutter*	Neg	Neg	Anger, Disgust	Anger, Disgust, Happy, Sad

Table 6: Example video for heatmap analysis of the contextual inter-modal (CIM) attention mechanism of the proposed MTMM-ES framework. Figure 3 depicts the heatmaps for the above video.



(a) *Softmax attention weights* N_1 & N_2 *for TV.* (b) *Softmax attention weights* N_1 & N_2 *for AV.* (c) *Softmax attention weights* N_1 & N_2 *for TA.*

Figure 3: (a), (b) & (c): Pair-wise softmax attention weights N_1 & N_2 of *visual-text*, *acoustic-visual* & *text-acoustic* for multi-task learning framework. 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. Each cell (i, j) of the heatmap signifies the weights of utterance 'j' for the classification of utterance 'i' of the pair-wise modality matrices, hence, assists in predicting the labels concisely by incorporating contextual inter-modal information.

in the multi-task framework. It is evident from Figure 2 that multi-task learning framework successfully leverages the inter-dependence of both the tasks in improving the overall performance in comparison to single-task learning. The improvements of MTL over STL framework is also statistically significant with p-value < 0.05 (c.f. Table 7).

We also present attention heatmaps of the multi-task learning framework in Figure 3. For illustration, we take the video of the first utterance of Table 6. It has total six utterances. We depict three pair-wise attention matrices of $2 \times (6 \times 6)$ dimension-one each for *text-visual*, *text-acoustics* and *acoustics-visual*. Solid lines in between represent the boundary of the two modalities, e.g. left

System	Emotion														Sentiment	
	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*	*F1*	*Acc*
Blanchard et al. (2018)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*63.2*	*60.0*
Zadeh et al. (2018b)⋆	-	-	71.4	65.2	**89.9**	-	-	-	60.8	-	85.4	53.3	-	-	76.0	76.0
Nojavanasghari et al. (2016)⋆	71.4	-	-	67.0	-	-	-	-	-	-	-	-	-	-	-	-
Rajagopalan et al. (2016)⋆	-	56.0	-	-	-	-	-	-	-	-	-	-	-	-	76.4	76.4
EF-LSTM (Zadeh et al., 2018c)⋆	-	-	-	-	-	56.7	-	57.8	-	59.2	-	-	-	-	-	-
TFN (Zadeh et al., 2017)⋆	-	60.5	-	-	-	-	66.6	66.5	-	58.9	-	52.2	-	-	-	-
Random Forest (Breiman, 2001)⋆	72.0	-	73.2	-	**89.9**	-	-	-	61.8	-	85.4	-	-	-	-	-
SVM (Zadeh et al., 2016)⋆	-	-	-	-	-	60.0	-	-	-	-	-	-	-	-	-	-
Zadeh et al. (2018a)⋆	-	-	-	-	-	-	**71.0**	-	-	-	-	-	-	-	-	-
Zadeh et al. (2018c)	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	77.0	76.9
Proposed (Single-task learning)	75.6	64.5	81.0	72.2	87.7	51.5	59.3	**61.6**	67.3	**65.4**	**86.5**	53.0	76.2	61.3	77.6	79.8
Proposed (Multi-task learning)	**75.9**	**66.8**	**81.9**	**72.7**	87.9	**62.2**	67.0	53.6	**72.4**	61.4	86.0	**60.6**	**78.6**	**62.8**	**78.8**	**80.5**
Significance *T*-test *w.r.t.* SOTA	**-**	**-**	**-**	**-**	-	-	**-**	**-**	**-**	**-**	-	**-**	***0.0240***	***0.0420***	***0.0012***	***0.0046***
Significance *T*-test *w.r.t.* STL	**-**	**-**	**-**	**-**	-	-	**-**	**-**	**-**	**-**	-	**-**	***0.0171***	***0.0312***	***0.0015***	***0.0278***

Table 7: Comparative results: Proposed multi-task framework attains better performance as compared to the state-of-the-art (SOTA) systems in both the tasks i.e. emotion recognition (average) and sentiment analysis. ⋆Values are taken from Zadeh et al. (2018c). †Six-class average. Significance *T*-test (< 0.05). STL: Single-task learning.

side of Figure 3a represents *text* modality and right side represents the *visual* modality. The heatmaps represent the contributing features for the classification of utterances. Each cell (*i,j*) of Figure 3 signifies the weights of utterance '*j*' for the classification of utterance '*i*' of the pair-wise modality matrices. For example, for the classification of utterance '*u4*' in Figure 3a, model puts more focus on the textual features of '*u2*' and '*u6*' than others and more-or-less equal focus on the visual features of all the utterances.

5.4 Comparative Analysis

We compare our proposed approach against various existing systems (Nojavanasghari et al., 2016; Rajagopalan et al., 2016; Zadeh et al., 2017, 2018a,b,c; Blanchard et al., 2018) that made use of the same datasets. A comparative study is shown in Table 7. We report the results of the top three existing systems (as reported in Zadeh et al. 2018c) for each case. In emotion classification, the proposed multi-task learning framework reports the best F1-score of 78.6% as compared to the 76.3% and Weighted Accuracy of 62.8% as compared to the 62.3% of the state-of-the-art. Similarly, for sentiment classification, the state-of-the-art system reports 77.0% F1-score and 76.9% accuracy value in the multi-task framework. In comparison, we obtain the best F1-score and accuracy value of 78.8% and 80.4%, respectively, i.e., an improvement of 1.8% and 3.5% over the state-of-the-art systems.

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

We also perform test-of-significance (*T*-test) and observe that the obtained results are statistically significant *w.r.t.* the state-of-the-art and proposed single-task results with p-values< 0.05.

5.5 STL v/s MTL framework

In this section, we present our analysis *w.r.t.* single-task and multi-task frameworks. Table 8 lists a few example cases where the proposed multi-task learning framework shows how it yields better performance for both sentiment and emotion, while the single-task framework finds it non-trivial for the classification. For example, first utterance has gold sentiment label as negative which was misclassified by STL framework. However, the MTL framework improves this by correctly predicting '*positive*'. Similarly, in emotion analysis STL predicts three emotions i.e. *disgust*, *happy* and *sad*, out of which only one emotion (*disgust*) matches the gold emotions of *anger* and *disgust*. In comparison, MTL predicts four emotions (i.e. *anger*, *disgust*, *happy* and *sad*) for the same utterance. The precision (2/4) and recall (2/2) for MTL framework is better than the precision (1/3) and recall (1/2) for the STL framework. These analyses suggest that the MTL framework, indeed, captures better evidences than the STL framework.

In the second example, knowledge of sentiment helps in identifying the correct emotion label in the MTL framework. For the gold sentiment (*positive*) and emotion (*happy* and *sad*) labels, STL correctly classifies one emotion (i.e. *sad*), but fails to predict the other emotion (i.e. *happy*). In addition, it misclassifies another emotion (i.e. *anger*). Since, gold label *happy* corresponds to the *posi-*

	Utterances	Sentiment			Emotion		
		Actual	STL	MTL	Actual	STL	MTL
1	*richard gere and susan umm you i really didn't enjoy this movie at all it kinda boring for*	Neg	Pos	Neg	Anger, Disgust	Disgust, Happy, Sad	Anger, Disgust, Happy, Sad
2	*we look forward to cooperating with the new government as it works to make progress on a wide range of issues including further democratic reforms promotion of human rights economic development and national reconciliation*	Pos	Pos	Pos	Happy, Sad	Anger, Sad	Happy, Sad
3	*laughter and applause still there though..*	Pos	Neg	Pos	Happy	Happy, Surprise	Happy
4	*is in love with some other person so you know the story*	Neg	Pos	Neg	Anger, Disgust, Sad	Disgust, Happy, Sad	Anger, Disgust, Sad
5	*i can say unfortunately i don't think it's a serious program*	Neg	Pos	Neg	Disgust, Sad, Surprise	Anger, Happy, Sad	Anger, Disgust, Happy, Sad
6	*the last administration bought into just as much as this one does unfortunately*	Neg	Pos	Neg	Anger, Disgust, Sad	Anger, Happy	Anger, Disgust, Happy, Sad
7	*it's just too great of a risk and it is socially unacceptable*	Neg	Pos	Neg	Anger, Disgust, Happy	Anger, Happy	Anger, Disgust, Happy
8	*had a robot here at hopkins since the year longer than most institutions in this country and around the world we*	Pos	Pos	Pos	Happy	Happy, Sad, No class	Happy
9	*in total we spent hundreds of hours on the ground on site watching these leaders in action*	Pos	Pos	Pos	No class	Happy	No class

Table 8: Comparison with multi-task learning and single-task learning frameworks. Few error cases where multi-task learning framework performs better than the single-task framework. *First utterance*: Improved MTL (*Pre: 0.5, Rec: 1.0*) performance over STL (*Pre: 0.3, Rec: 0.5*). *Second utterance:* Sentiment (i.e. *Pos*) assists in emotion classification (i.e. *Happy*). Red color represents error in classification.

tive scenario and predicted label *anger* is related to negative, knowledge of sentiment is a crucial piece of information. Our MTL framework identifies this relation and leverage the predicted sentiment for the classification of emotion i.e. *positive* sentiment assists in predicting *happy* emotion. This is an example of inter-dependence between the two related tasks and the MTL framework successfully exploits it for the performance improvement.

We also observe that the system puts comparatively more focus on some classes in MTL framework than the STL framework. As an instance, MTL predicts '*anger*' class for 1173 utterances, whereas STL predicts it for 951 utterances (1063 anger utterances in the gold dataset). Further, we observe contrasting behavior for the '*sad*' class, where MTL predicts 1618 utterances as '*sad*' compared to the 2126 utterances of STL. For '*disgust*' and '*happy*' classes, both STL and MTL frameworks predict the approximately equal number of utterances.

Further, we observe that MTL performs poorly for the '*fear*' and '*surprise*' classes, where it could not predict a significant number of utterances. A possible reason would be the under-representation of these instances in the given dataset.

6 Conclusion

In this paper, we have proposed a deep multi-task framework that aims to leverage the inter-dependence of two related tasks, i.e., multi-modal sentiment and emotion analysis. Our proposed approach learns a joint-representation for both the tasks as an application of GRU based inter-modal attention framework. We have evaluated our proposed approach on the recently released benchmark dataset on multi-modal sentiment and emotion analysis (MOSEI). Experimental results suggest that sentiment and emotion assist each other when learned in a multitask framework. We have compared our proposed approach against the various existing systems and observed that multi-task framework attains higher performance for all the cases.

In the future, we would like to explore the other dimensions to our multi-task framework, e.g., Sentiment classification & intensity prediction, Emotion classification & intensity prediction and all the four tasks together.

Acknowledgement

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. 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

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, Melbourne, Australia.

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

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*, 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*.

Paul Ekman. 1999. *Basic Emotions*. The handbook of cognition and emotion.

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.

D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann. 2018. Self-attentive feature-level fusion for multimodal emotion detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval*, pages 196–201.

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.

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

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*, pages 284–288, New York, NY, USA.

Bo Pang, , and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

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.

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.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CoRR*, abs/1708.02711.

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*, pages 1547–1556, Vancouver, Canada.

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*, pages 5642 – 5649, New Orleans, LA, 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.

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*, pages 2236–2246, Melbourne, Australia.