Shapes of Emotions: Multimodal Emotion Recognition in Conversations via Emotion Shifts

Harsh Agarwal[*]   **Keshav Bansal**[*]   **Abhinav Joshi**   **Ashutosh Modi**

Indian Institute of Technology Kanpur (IIT-K)

{harshagarwal0194,keshav22bansal}@gmail.com
{ajoshi,ashutoshm}@cse.iitk.ac.in

Abstract

Emotion Recognition in Conversations (ERC) is an important and active research area. Recent work has shown the benefits of using multiple modalities (e.g., text, audio, and video) for the ERC task. In a conversation, participants tend to maintain a particular emotional state unless some stimuli evokes a change. There is a continuous ebb and flow of emotions in a conversation. Inspired by this observation, we propose a multimodal ERC model and augment it with an emotion-shift component that improves performance. The proposed emotion-shift component is modular and can be added to any existing multimodal ERC model (with a few modifications). We experiment with different variants of the model, and results show that the inclusion of emotion shift signal helps the model to outperform existing models for ERC on MOSEI and IEMOCAP datasets.

1 Introduction

Humans are complex social beings, and emotions are indicative of not just their inner state and feelings but also their internal thinking process (Minsky, 2007). To fully understand a person, one needs to understand their inherent emotions. Recent research has witnessed colossal interest in including artificially intelligent machines as conversable companions for humans, e.g., personal digital assistants. However, communication with AI systems is quite limited. AI systems do not understand the inherent emotions expressed implicitly by humans making them unable to comprehend the underlying thought processes and respond appropriately. Consequently, a wide variety of approaches have been proposed for developing emotion understanding and generation systems (Sharma and Dhall, 2021; Witon et al., 2018; Singh et al., 2021a; Goswamy et al., 2020; Colombo et al., 2019; Singh et al., 2021b; Joshi et al., 2022).
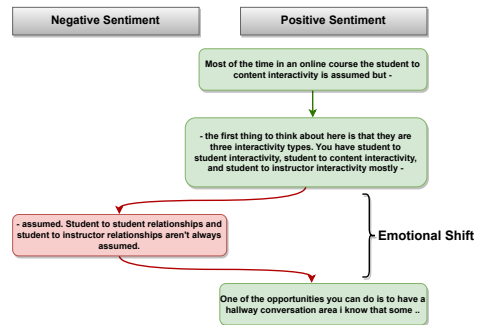

Figure 1: Emotional shift on the dialogue "m7SJs73SF8w" from CMU-MOSEI dataset

During an interaction, humans express different emotions and fluctuate between multiple emotional states. It is often the case that participants in a conversation tend to maintain a particular emotional state unless some stimuli evokes a change. This observation is closely related to *Shapes of Stories* proposed by renowned writer Kurt Vonnegut (Vonnegut, 1995), who posits that every story has a shape plotted by the ups and downs experienced by the characters of the story, and this, in turn, defines an *Emotional Arc* of a story. This phenomenon has also been empirically verified by Reagan et al. (2016), who analyzed around 1300 stories to come up with common emotional arc patterns across various stories. Moreover, apart from these flows, there exists a sudden shift of emotions from positive to negative sentiments. Consider an example shown in Figure 1, where the sentiment of the third utterance shifts from positive to negative and back again to positive in the fourth utterance. Current state-of-the-art methods are often oblivious to the presence of such emotion shifts and tend to fail in cases where there is a sudden change in the emotional state (Poria et al., 2019). To address this issue, we propose incorporating a novel module that explicitly tracks such emotional shifts in conversations. Humans express their emotions via various modalities, such as language, modulations in voice, facial expressions, and body gestures. In this paper, to

---

[*]Equal Contributions

44

fully and correctly recognize human emotions, we propose a multimodal emotion recognition system that utilizes language, audio, and video modalities. We propose a multimodal ERC model based on GRUs that fuses information from different modalities. An independent emotional shift component captures the emotion shift signal between consecutive utterances, allowing the model to forget past information in case of an emotional shift. We make the following contributions:

- We propose a new deep learning based multimodal emotion recognition model that captures information from text, audio, and video modalities. We release the model implementation and experiments code in the supplement.

- We propose a novel emotion shift network (modeled via a Siamese network) that guides the main emotion recognition system by providing information about possible emotion shifts or transitions. The proposed component is modular, it can be pretrained and added to any existing multi-modal ERC (with a few modifications) to improve emotion prediction.

- The proposed model is experimented on the two widely known multimodal emotion recognition datasets (MOSEI and IEMOCAP), and results show that emotional shift component helps to outperform some of the existing models. We perform detailed analysis and ablation studies of the model and show the contribution of different components. We analyse the performance of our model in the classification of utterances having a shift in emotion and compare this with previous models and report an improvement due to the use of emotion-shift information. We further examine how the internal GRU gates behave during emotion shifts.

2   Related Work

Emotion recognition using multiple modalities is an active area of research leading to the development of widely popular benchmark datasets, e.g., CMU-MOSEI (Bagher Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). Recent works have highlighted the crucial aspects of self, and interpersonal dependencies in the emotional dynamics of the conversations (Poria et al., 2019). Another essential feature is the role of the local and global context for emotion recognition systems. Some notable works like Dialogue RNN

(Majumder et al., 2018b) try to capture these properties by modeling each speaker with a party state and the emotion of each utterance by an emotional state. Furthermore, a context state is maintained to model the global conversation context. Another work Multilogue-Net (Shenoy and Sardana, 2020) highlights the limitation of the fusion mechanism used in Dialogue RNN (Majumder et al., 2018b) and tries to solve it using a party, context, and emotion GRUs for each modality. It uses a pairwise attention mechanism proposed by (Ghosal et al., 2018) to fuse the emotion states for all the modalities effectively. However, DialogueRNN highlights the poor performance in predicting the emotions with the utterances where the emotion shifts from positive to negative sentiments. Our work considers the emotion shifts present in the dialogues and tries to leverage them for improving emotion recognition. Another line of work includes Transformer-Based Joint-Encoding (TBJE) (Delbrouck et al., 2020) that achieves the state-of-the-art results on the sentiment task for the MOSEI dataset using a multimodal transformer-based model for combining multiple modalities. However, in the emotion task, TBJE is outperformed by the Multilogue-Net model. The possible reason highlighted by the paper is the lack of context-awareness in the architecture, as TBJE neither uses the previous nor next utterance to predict the emotion for the current utterance. Some of the other works in multimodal emotion recognition include the Memory Fusion network (MFN) (Zadeh et al., 2018), which aligns multimodal sequences using multi-view gated memory, Graph-MFN (Bagher Zadeh et al., 2018) which uses Dynamic Fusion Graph (DFG) and learns to model the n-modal interactions dynamically, and bc-LSTM (Poria et al., 2017) which uses an LSTM-based model to capture contextual information. CESTa (Wang et al., 2020) captures the emotional consistency in the utterances using CRF model (Lafferty et al., 2001) for boosting the performance of emotion classification and comes close to our idea of leveraging emotion shifts.

3   Task and Corpus

Problem Definition: Consider a conversation having utterances $u_1, \ldots, u_N$. The task of Emotion Recognition in Conversation (ERC) is to predict the emotion (or sentiment) of each utterance u_t. We define an utterance to be a coherent piece of information (single or multiple sentences) con-
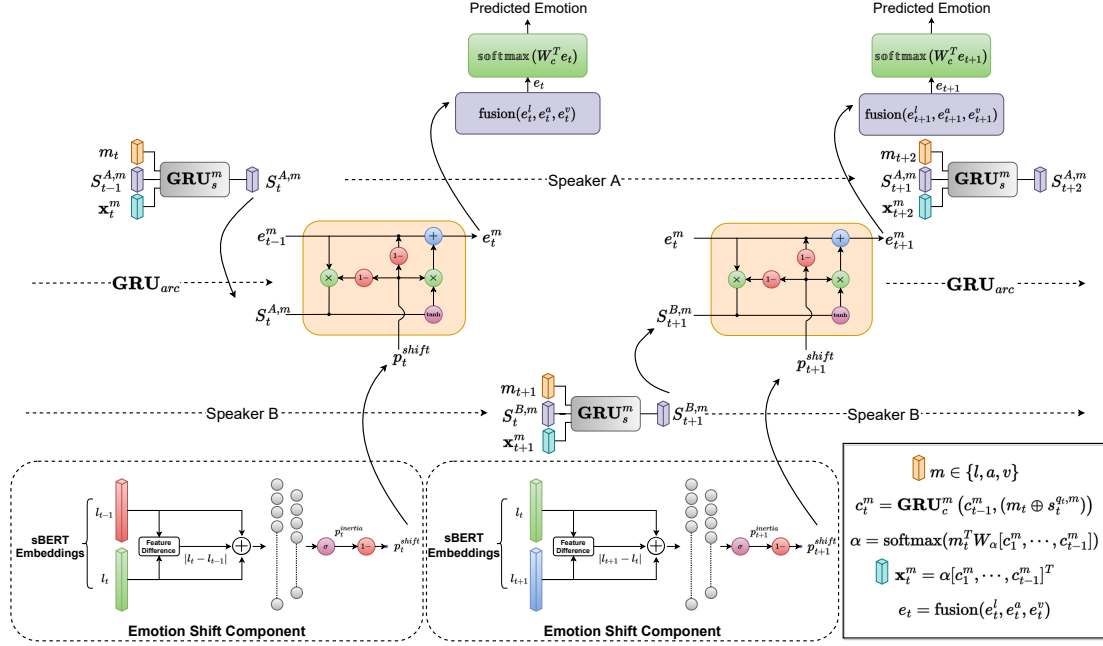
Figure 2: The model architecture for a conversation between two speakers, A and B, at time t and $t + 1$. The upper part highlights the Emotion Classification Component, and the lower part highlights the Emotion Shift Component.

veyed by a single participant at a given time. We model an utterance in terms of different modalities: $u_t = \{l_t, a_t, v_t\}$. An utterance (u_t) at time-step t is represented via features from textual transcript (l_t), audio (a_t), and visuals (v_t) of the speaker. We denote the speaker of utterance u_t as q_t.

3.1 Corpus Details

3.1.1 CMU-MOSEI:

The CMU Multimodal Opinion Sentiment and Emotion Intensity (Bagher Zadeh et al., 2018) is an English language dataset containing more than 65 hours of annotated video from more than 10000 speakers and 250 topics. Each sentence is annotated for a sentiment on a $[-3, 3]$ Likert scale. However, in this work, we project these labels to a two-class classification setup with values ≥ 0 signifies positive sentiments and values < 0 convey negative sentiments. Dataset also contains six emotion labels, namely angry, happy, sad, surprise, fear and disgust for each utterance. Note that in case of emotions labels the utterances are multi-label. Which means a single utterance can have more than one emotion label. We have shown results for both sentiment and emotion prediction tasks.

3.1.2 IEMOCAP:

The IEMOCAP benchmark (Busso et al., 2008) consists of a conversation between ten distinct speakers. The dataset contains two-way conver-

sations in videos where every video clip contains a single dyadic English dialogue. Further, each dialogue segments into utterances with an emotion label from six emotion labels, i.e., happy, sad, neutral, angry, excited, and frustrated. The dataset incorporates an acted setting where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expression.

4 The Proposed Model

During a conversation, speakers tend to maintain an emotional flow of affective states. These states majorly rely on the context of the entire conversation; for example, if the overall gist of the speaker about the topic is positive, the emotions like happiness, joy, and surprise can be seen more often than negative emotions like anger and sadness. Moreover, a speaker's emotions are often affected by the past emotions present in the conversation. Hence, an emotion prediction model should not only take into account the context but should also be able to maintain the speaker-level information along with the emotions present in the past utterances. Considering these assumptions, we propose the primary component of our emotion recognition model: the *Emotion Classification Component*. The emotions classification component predicts an utterance's emotion label using information from the current speaker, emotions of the previous utterances, and

the entire conversation context. Another significant insight about the emotions in a conversation is the sudden shift of emotional states. Many of the existing state-of-the-art approaches highlight this in their error analysis, where the model fails to capture the sudden shifts in emotional states leading to a misclassified emotion prediction. To incorporate the effect of a sudden shift in the emotion, we introduce a separately trained component called the *Emotion Shift Component*. The emotion shift component explicitly models the probability (p_t^{shift}) of a shift in emotion between the utterances u_{t-1} and u_t. This shift in emotion can be expressed as moving from a positive (e.g., happy) to negative emotion (e.g., sad) or vice-versa between consecutive utterances. The emotion shift component being independent of the primary architecture is pretrained, and helps control the information flow from past to future during a sudden change. The signal from the pretrained emotional shift component is added to the emotion classification component to control the flow of emotions from past to future. Figure 2 shows detailed architecture of the proposed model.

Emotion Classification Component: For modeling the underlying emotions in a conversation, we maintain a *party state*, *emotion state* and *context state*. The party state is maintained for each speaker and helps to keep track of the participant specific aspect in a conversation. The context state is global (common across each participant) and helps to encode the entire conversation context, thereby capturing inter-utterance dependencies. Akin to the context state, the emotion state is also global and helps to leverage the emotion information flow between utterances. Moreover, the emotion shift signal between the current and previous utterance is used to update the global emotion state. The emotion label for each utterance is then predicted by decoding the emotion state. In our model each of the party, context and emotion states are modality specific and are updated using a modality specific GRU (Chung et al., 2014) network for each modality $m \in \{l, a, v\}$ (indicated by the superscript m). We employ late fusion to combine the emotion states from different modalities. Next, we explain different GRU networks used in the model.

$$s_t^{q_t,m} = \mathbf{GRU}_s^m \left(s_{t-1}^{q_t,m}, (m_t \oplus \mathbf{x}_t^m) \right) \quad (1)$$

$$c_t^m = \mathbf{GRU}_c^m \left(c_{t-1}^m, (m_t \oplus s_t^{q_t,m}) \right) \quad (2)$$

$$e_t^m = \mathbf{GRU}_{arc}^m \left(e_{t-1}^m, s_t^{q_t,m}, p_t^{\text{shift}} \right) \quad (3)$$

$$e_t = \text{fusion}(e_t^l, e_t^a, e_t^v) \quad (4)$$

Party State Update (GRU$_s$): The state of each participant is modeled by the party state update GRU$_s$. For each modality $m \in \{l, a, v\}$, q_t's party state $s_{t-1}^{q_t,m}$ is updated to $s_t^{q_t,m}$ using an attention vector $\mathbf{x}_t^m$ and modality specific feature m_t (Eq. 1), $\oplus$ denotes concatenation operation. Here $\mathbf{x}_t^m$ is calculated using a simple dot product attention mechanism over the context states (c_t^m). Note that for all speakers other than q_t, the party state at $t-1$ and t remains the same.

Context State Update (GRU$_c$): Global conversation context is modeled using the context state update GRU$_c$. For each modality $m \in \{l, a, v\}$, the global context state c_{t-1}^m is updated to c_t^m (Eq. 2) using the q_t's party state $s_t^{q_t,m}$ and the corresponding modality feature m_t. Context states ($c_1^m, \cdots, c_{t-1}^m$) are used for calculating the attention vector x_t^m for each modality $m \in \{l, a, v\}$ as follows:

$$\alpha = \text{softmax} \left(m_t^T W_\alpha \left[c_1^m, \cdots, c_{t-1}^m \right] \right) \quad (5)$$

$$x_t^m = \alpha \left[c_1^m, \cdots, c_{t-1}^m \right]^T \quad (6)$$

Emotion State Update (GRU$_{arc}$): For each modality $m \in \{l, a, v\}$, the global emotion state e_{t-1}^m is updated to e_t^m (Eq. 3) using the current party state $s_t^{q_t,m}$ and modulated by the emotion shift component (p_t^{shift}). The emotion states for all the three modalities are fused together (Eq. 4) to create e_t using a pairwise attention mechanism (Shenoy and Sardana, 2020). e_t is later used to decode the emotion class for an utterance.

The emotion classification component is a context-aware model similar to that of previous works like Multilogue-net (Shenoy and Sardana, 2020) but with a few key differences. Firstly, instead of modelling an emotion state for each participant, we introduce global emotion state for each conversation. This is done to make use of the flow of emotion between utterances. Secondly, the emotion shift signal between the current and previous utterance (p_t^{shift}) is used to update the global emotion state using a GRU$_{arc}$ which aims to model the emotion arc in the conversation.

Emotion Shift Component: To capture the emotional arc across the conversation, we explicitly model probability of emotion shift (p_t^{shift}) between successive utterances (u_{t-1} and u_t). We use a Siamese network (Bromley et al., 1993) to model the emotional shift present across utterances. A Siamese network generally consists of two or more identical subnetworks having the same configuration with shared parameters and weights. The pro-

posed emotion shift architecture takes the textual features of the current (l_t) and previous (l_{t-1}) utterances and outputs the probability of maintaining emotional inertia ($p_t^{inertia}$). The architecture of the emotion shift prediction network is shown in lower half of Figure 2. We use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) embeddings as textual features. SBERT is a modification of the pre-trained BERT (Devlin et al., 2019) network that uses the Siamese network to derive semantically meaningful sentence embeddings for transfer learning tasks. The emotion shift prediction network makes use of only the text modality for two reasons. Firstly, it has often been found empirically that among the text, audio, and video modalities, text modality carries more information for ERC tasks (Poria et al., 2018). Secondly, early fusion techniques to combine the three modalities can suffer in a Siamese-type architecture due to difficulty in mapping the fused modality vector to a vector space in which similar vectors are close. We also experimentally verify this (§6).

The emotion shift prediction network (between u_t and u_{t-1}) takes in text features corresponding to utterances (l_t and l_{t-1}) and their element wise differences to output the probability of a shift as given by (Eq. 7). Here, $p_t^{inertia}$ is calculated using Siamese network (Eq. 8, 9). Here, $\mathcal{H}_t$ is the Siamese hidden state, $\mathcal{W}$ $(\in \mathbb{R}^{3d_l})$ is the model parameter. For the Siamese network, we use Binary Cross Entropy loss ($\mathcal{L}_s$) over the distribution p_t^{shift}. The emotion shift component modulates the Emotion State GRU_{arc} via p_t^{shift} and hence controls the flow of information during the conversation. The Emotion Shift component captures the emotional consistency in the utterances and can act as an independent modular component that can be pre-trained and added to any existing multi-modal ERC framework with a few modifications for improving emotion recognition in conversations.

$$p_t^{\text{shift}} = 1 - p_t^{inertia} \tag{7}$$

$$p_t^{inertia} = \sigma\left(\mathcal{H}_t\right) \tag{8}$$

$$\mathcal{H}_t = \mathcal{W}^T(l_{t-1} \oplus l_t \oplus |l_t - l_{t-1}|) \tag{9}$$

Overall Architecture: The motivation for the proposed architecture follows from the intuition that we need to weigh down the contribution of the previous emotion state in case of an emotion shift. In other words, we need to reduce the influence of e_{t-1}^m in the calculation of e_t^m when there is a high p_t^{shift}. To do so, we modify the reset and update

gates in the GRU modelling the emotional arc of the conversation i.e. GRU_{arc}. A GRU has gating units (reset and update gates) that modulate the flow of information inside the unit. Ravanelli et al. (2017) mention the usefulness of reset gate in scenarios where significant discontinuities are present in the sequence, thereby indicating its crucial role to forget information. Their work also finds a redundancy in the activations of the reset and update gates when processing speech sequences. Motivated by this, and the intuition that we need to forget more information when there is a higher probability of an emotional shift, we directly use the value of $(1 - p_t^{shift})$ for both the reset and update gates. The updates for GRU_{arc} unit are given by Eqs. 10, 11. Eq. 10 calculates a candidate emotion state $\tilde{e}_t^m$ in which the prior emotion state's (e_{t-1}^m) is controlled by the emotion shift signal. The output e_t^m is a linear interpolation between $\tilde{e}_{t-1}^m$ and e_{t-1}^m. Again, p_t^{shift} controls the influence of e_{t-1}^m (Eq. 11). Therefore, a higher value of p_t^{shift} will limit the contribution of the previous emotion state. In the absence of the emotion shift component, the GRU gates are learned using only the classification data, much like the rest of the parameters in the model. If the total number of parameters in a model is huge (as is the case with most deep learning models), the gates might be unable to learn well. We verify that the modeling of the shift in emotion encourages better learning of these gates (§6).

$$\tilde{e}_t^m = \tanh\left(W s_t^{qt,m} + (1 - p_t^{shift}) \odot \left(U e_{t-1}^m\right)\right) \tag{10}$$

$$e_t^m = (1 - p_t^{shift}) \odot e_{t-1}^m + p_t^{shift} \odot \tilde{e}_t^m \tag{11}$$

For prediction at time t, the emotion vector e_t (formed from fusion of e_t^m as described in (Eq. 4)) is passed through a final classification layer W_c $(\in \mathbb{R}^{d_e \times K})$ where K is the number of emotion or sentiment classes. This is used to obtain probability distribution over emotion labels via the Softmax activation: $o = \text{softmax}(W_c^T e_t)$. We use the Cross-Entropy Loss over this distribution to train the weights of the emotion classification component.

5 Experiments and Results

Multimodal Emotion Corpora: We evaluate our model using two benchmark English ERC datasets - CMU Multimodal Opinion Sentiment and Emotion

Dataset	#utterances		Emotion shift (in %)	
	Train	Test	Train	Test
CMU-MOSEI	18191	4655	33.61	34.62
IEMOCAP	5810	1623	12.89	12.75

Table 1: Statistics for number of utterances and emotion shift percentage in various datasets

Model	F1	Accuracy
Graph-MFN	77.00	76.90
DialogueRNN	79.82	79.98
Multilogue-Net	80.01	82.10
TBJE	-	82.4
Our Model	**83.07**	**82.66**

Table 2: Performance comparison on the sentiment task of CMU-MOSEI dataset (all numbers in %)

Emotion	Multilogue-Net		TBJE		Our Model	
	A	F1	A	F1	A	F1
Happiness	**70.05**	**70.03**	66.00	65.50	68.51	68.61
Sadness	71.04	70.42	73.90	67.90	**74.20**	**71.74**
Anger	74.78	74.31	**81.90**	76.00	75.17	**76.10**
Disgust	77.98	79.20	**86.50**	**84.50**	83.67	82.79
Fear	69.04	75.50	**89.20**	**87.20**	87.11	85.90
Surprise	88.89	85.98	**90.60**	**86.10**	78.99	81.62

Table 3: Performance comparison on the emotion task of CMU-MOSEI dataset (all numbers in %)

Intensity (CMU-MOSEI) dataset and the Interactive Emotional Dyadic Motion Capture (IEMO-CAP) dataset. Details of these corpora are discussed in §3.1. In a nutshell, each of the two corpora has language, audio, and video modalities. MOSEI has both sentiment and six emotion labels, IEMOCAP has video recordings of dyadic conversations and is labeled with six emotion labels.

Emotion shift in Dataset: We define an emotion shift between consecutive utterances if there is a shift from a positive to a negative emotion or vice-versa. CMU-MOSEI dataset provides annotated (positive/negative) sentiment label for each utterance. This is not the case for the IEMOCAP dataset, therefore we divide the emotion classes into a positive and negative category. Happiness and surprise are taken into the positive category while disgust, angry and sad are considered as the negative category. Note that IEMOCAP also has a neutral emotion, but a shift is only counted if it is from a positive to negative emotion or vice-versa. Table 1 shows the percentage of emotion shift observed in the datasets. Since CMU-MOSEI shows a larger amount of emotion shift, we were motivated to perform experiments on CMU-MOSEI first.

5.1 Results

We evaluate our approaches using standard F1 score and Accuracy evaluation metrics (App. A). We train and report the performance of our model for four sub-tasks, 2-way sentiment classification and binary emotion classification on CMU-MOSEI, and four-class and six-class emotion classification task for IEMOCAP. The focus of our work is multimodal ERC and consequently, as is done in pre-

vious work, we compare only with previous multimodal approaches, since comparison with unimodal (e.g., text) only approaches does not make sense. Moreover, SOTA unimodal approaches (such as text based) use additional information such external knowledge sources (e.g., (Ghosal et al., 2020)) which makes the comparison with multimodal approach unfair specially given that such knowledge may not be available for other modalities. Nevertheless, it is possible to incorporate the emotion shift component into existing emotion prediction architectures (unimodal or multimodal) and we leave this exploration for future.

Results on CMU-MOSEI: Table 2 shows comparison of our best performing model on CMU-MOSEI sentiment labels, with current state of the art models: TBJE (Delbrouck et al., 2020), Multilogue-Net (Shenoy and Sardana, 2020), Dialogue RNN (Majumder et al., 2018b), and Graph-MFN (Poria et al., 2017). As evident from the results, we are able to significantly outperform the previous SOTA Multilogue-Net model with an increase of 3% in F1 score. We further compare our model on the emotion classification task with TBJE and Multilogue-Net (Table 3). As shown in the table, our model outperforms for some of the emotion classes. We speculate that poor performance is due to the multilabel setting in the CMU-MOSEI dataset. As the emotion labels are multilabel, the emotion shift component is not able to play a meaningful role in providing a performance boost to the emotion classification component. We consider multilabel settings as another line of future work where the emotion shift modeling takes into account the multilabel property.

Results on IEMOCAP: Previous works on Multimodal-IEMOCAP have shown performance only on angry, happy, sad, and neutral emotions. We compare our model performance on these four classes with state-of-the-art models CHFusion (Majumder et al., 2018a) and bc-LSTM (Poria et al.,

Emotion	bc-LSTM		CHFusion		Our model	
	A	F1	A	F1	A	F1
Happy	**79.31**	-	74.30	**81.40**	68.75	72.79
Sad	**78.30**	-	75.60	77.00	76.73	**81.21**
Neutral	69.92	-	78.40	71.20	**81.51**	78.25
Angry	77.98	-	79.60	77.60	**82.35**	79.77
Avg(w)	75.20	-	76.50	76.80	**78.47**	**78.46**

Table 4: Performance comparison on the IEMOCAP dataset for four emotion labels (all numbers in %)

Emotion	Hap	Sad	Neu	Ang	Exc	Fru	Avg(w)
Acc	54.17	65.31	62.50	62.94	67.89	64.04	63.59
F1	50.81	70.48	60.23	63.69	70.73	62.72	63.82

Table 5: Performance of our model on IEMOCAP dataset for 6 labels (all numbers in %)

Datasets		Accuracy	F1
CMU-MOSEI		72.65	67.32
IEMOCAP	4-label	80.50	79.68
	6-label	85.63	84.28

Table 6: Performance of Siamese Model on MOSEI and IEMOCAP

Input Features	Classification		Emotion Shift	
	Accuracy	F1	Accuracy	F1
G(L), O(A), F(V)	80.85	80.31	63.38	62.15
B(L), O(A), O(V)	81.98	80.91	64.01	63.30
B(L), O(A), O(V)	**82.66**	**83.07**	**72.65**	**67.32**

Table 7: Effect of different feature combinations for MOSEI. The classification columns are results on 2 class sentiment prediction task and emotion shift columns are results on 2 class emotion shifts classes. Here G(L): Glove, B(L): BERT, F(V): Facet, O(A): OpenSmile, O(V): OpenFace2.0

2017) (Table 4). Our model significantly outperforms both of these on average weighted F1 and Accuracy. Also, emotion classes neutral and angry show improved performance. We also provide results on six emotion classes - happy, sad, neutral, angry, excited, and frustrated (Table 5). For these experiments, we use BERT features for text (§6), OpenSmile features for audio and 3D-CNN features (Majumder et al., 2018b) for video. We did not come across any existing work on 6-class multimodal IEMOCAP for the comparison.

Performance of emotion shift component: The results describing the capability of the emotion shift component to predict the shift for CMU-MOSEI and IEMOCAP dataset are shown in Table 6. It is to be noted that predicting the shift accurately is not our primary objective. Our objective is to be able to improve the emotion prediction by using the signal (p_t^{shift}) received from the emotion shift component.

6 Analysis and Ablation Studies

Due to a wide variety of components, it becomes vital to perform a detailed analysis of the architecture to understand the importance of various choices.

Feature and Design choices: For understanding the importance of features used for different modalities, we choose two different sets of features for text and visual modalities. In one setting, we use averaged GloVe embeddings (Pennington et al., 2014) for text, OpenSmile features (Eyben et al., 2010) for Audio and Facet features (Stöckli et al., 2017) for Video. Whereas in another setting, for text modality, we make use of a pre-trained BERT (Devlin et al., 2019) model's output layer. We calculate the average of the output layer to get a fixed-sized

vector. For visual modality, we use the features provided by OpenFace2.0 (Baltrusaitis et al., 2018), which are useful in performing facial analysis tasks such as facial landmark detection, head-pose tracking, and eye-gaze tracking. The results in Table 7 (first two rows) highlight the advantage of features used in second setting.

Importance of Pretraining the Emotion Shift Component: To review the significance of the pretraining emotion shift component, we compare it with the two settings described above. We argue that jointly optimizing the emotion classification and emotion shift component from scratch might degrade the model classification performance. At the onset of training, the Siamese component does not provide a helpful signal to the classification component due to the random initialization of its weights, hampering the learning of the classification component. To prevent this, we pre-train the emotion shift component on the emotion shift labels separately before the joint training task, which helps provide the classification component with a better emotion shift signal at the start of training, making learning more accessible. The results in Table 7 (third row) shows an increase of approximately 2% in the F1 score when compared to the same features setting without pretraining (second row). Moreover, the Siamese network, when pretrained, also achieves an F1 score of 67.32%, the highest among all the experiments depicting the hindrance caused by joint training from scratch.

Performance over emotion shift utterances: To verify the effectiveness of the emotion shift com-

Emotion shift type	Multilogue-Net	Our Model
Positive - Negative	69.78	**73.83**
Negative - Positive	59.49	**80.35**

Table 8: Accuracy comparison with Multilogue-Net on MOSEI emotion shift utterances

Emotion shift type	4-label	6-label
Positive - Negative	63.04	53.22
Negative - Positive	69.77	70.68

Table 9: Accuracy of the proposed model on IEMOCAP emotion shift utterances

Datasets		Without		With	
		A	F1	A	F1
CMU-MOSEI		81.31	81.01	**82.66**	**83.07**
IEMOCAP	4-label	77.53	77.51	**78.47**	**78.46**
	6-label	61.37	61.65	**66.73**	**66.86**

Table 10: Performance with and without the emotion shift component (all numbers in %)

Datasets		L+A		L+V		A+V		L+A+V	
		A	F1	A	F1	A	F1	A	F1
CMU-MOSEI		82.49	82.81	81.65	82.16	72.22	71.28	**82.66**	**83.07**
IEMOCAP	4-label	**80.06**	**80.07**	77.52	77.56	78.26	78.20	78.47	78.46
	6-label	64.26	64.48	63.34	63.40	58.90	58.84	**66.73**	**66.86**

Table 11: Ablation study to observe the contribution of different modalities

ponent we consider cases where an emotion shift has occurred between a target utterance u_t and the prior utterance u_{t-1} if there is a switch from positive emotion in u_{t-1} to negative emotion in u_t, or vice versa (§3). We evaluate our emotion classification performance on such utterances u_t displaying an emotion shift. Popular architectures like CMN, ICON, IANN and DialogueRNN perform poorly on the utterances with an emotion shift (Poria et al., 2019). In particular, in cases where the emotion of the target utterance differs from the previous utterance, DialogueRNN could only correctly predict 47.5% instances, much lesser than the 69.2% success rate that it achieves at the regions of no emotional shift. In Table 8 we compare our results with another multimodal ERC SOTA: Multilogue-Net. The results show a significant increase in accuracy for both positive to negative and negative to positive emotion shifts on the CMU-MOSEI dataset depicting the importance of the independent emotion shift component introduced in our architecture. Even though the Siamese network can predict the presence of emotion shift with an accuracy of about 72.65% (Table 7), the signal received from it (in the form of reset and update gates of GRU) helps the emotion classification network to overcome the emotional inertia and predict the correct emotion. We also show the accuracy of our model on emotion shift utterances of the IEMOCAP dataset in Table 9. We could not calculate these numbers for CHFusion (SOTA on IEMOCAP) due unavailability of their code.

Effect of Modeling Emotion Shift: To further verify the significance of modeling emotion shift as a separate component, we compare two variants of our best model - one with and the other without the Emotion shift component. Table 10 shows a comparison on both the datasets. Across both the

datasets, we observe a substantial increase in performance while using the emotion shift component.

Contributions of the Modalities: To understand the importance of different modalities present in the datasets, we conduct experiments by choosing a combination of two out of the three modalities. As expected, models using all three modalities outperform models using only two modalities across most datasets (Table 11). On the IEMOCAP dataset with four classes, the text+audio model performs better than six classes. The text modality seems to be the most essential compared to other modalities highlighting the significance of context.

Using other modalities in the Emotion Shift Component: To observe the effectiveness of modalities other than text on the Emotion Shift Component, we empirically analyze the effect of using all three modalities for training this component. We make use of the early fusion technique where modalities l_t, a_t, v_t are concatenated ($l_t \oplus a_t \oplus v_t$) and then passed to the Siamese network. Observing the obtained results, we see that the use of the three modalities does not lead to an improvement (Table 12). A possible reason for this might be the importance of context (captured in the text modality) in predicting emotion shifts.

Analyzing reset gate updates in GRU_{arc}: We also verify the practical significance of the emotion shift component qualitatively. We compare the GRU reset gate activations obtained from the emotion shift component and the reset gate activations learned by GRU without explicit emotion shift information. We randomly pick an instance from the CMU-MOSEI test set and analyze the GRU unit using it. In Figure 3, we show these activations for the Video ID "m7SJs73SF8w" randomly selected

Modalities	Accuracy	F1
L	**82.66**	**83.07**
L+A+V	82.20	82.78

Table 12: Performance using other modalities in Siamese component (all numbers in %)

from the test set. This dialogue has four utterances, and we see a shift from positive to negative emotion between utterances two and three and a shift from negative to positive emotion between utterances three and four. As seen in the left graph in Figure 3, the emotion shift component learns to set a low reset gate value when there is an emotion shift (namely timestamps $t = 3$ and $t = 4$). This low reset gate value helps to weigh down the contribution of the previous emotion state for the predictions at the current timestamp. Comparing it to the case when we remove the emotion shift component (right graph in Figure 3), the reset gate activations learned by the GRU do not follow the same trend, indicating that the previous emotion state will still significantly contribute to predictions at the current timestamp. Overall, the emotion shift component plays a vital role in effectively controlling information from the past.

Figure 3: Reset Gate activations on the dialogue 'm7SJs73SF8w" from CMU-MOSEI test dataset

7 Discussion

The presence of emotion shifts in human-to-human conversation is prominent in the conversational datasets. The existing works based on sequential modeling often suffer from these shifts, leading to poor performance for utterances with emotion shifts. In this work, we try to control the effect of previous utterances using an independent emotion shift module. As highlighted in Tables 8 and 9, the proposed architecture performs significantly better on emotion shift cases when compared to Multilogue-Net (20% improvement in negative-positive and 4% improvement on positive-negative shifts). The novel design of the emotion shift-based gating mechanism in the GRU unit helps boost the prediction performance for utterances with emotion shifts. As noticed in Fig. 3, the reset and update gates provide a significant signal when there is an emotional shift in conversation.

Modularity: The modular design and idea of the proposed emotion shift component can further be used to improve any emotion prediction systems that have poor performance in emotion shift cases. Moreover, the designed emotion shift component
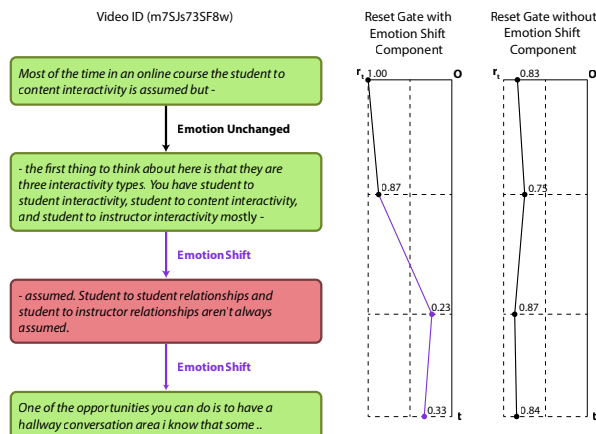
works considering only the textual modality, making it applicable to both multimodal as well as unimodal systems.

Application to Real-Time Systems: A notable limitation of all the existing Emotion Recognition state-of-the-art systems often comes from the incapability of their implementations for real-time use cases as they require the entire context to be given in the form of multiple utterances to the model. For future approaches where the models will target the real-time setting, the proposed emotion shift component can be handy as it only uses two consecutive utterances to predict the emotion shift.

8 Conclusion and Future Directions

In this paper, we proposed a deep learning based model for multimodal emotion recognition in conversations. We proposed a new emotion shift component (modeled using the Siamese net) that captures the emotional arc in a conversation and steers the main emotion recognition model. We performed a battery of experiments on two main emotion recognition datasets. Results and analysis show the importance of the emotion shift component. Currently, the emotion shift component uses only the text modality for predicting the shift and we plan to explore more sophisticated ways of using information from multiple modalities.

9 Acknowledgements

References

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

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*.

Jane Bromley, J.W. Bentz, Leon Bottou, I. Guyon, Yann Lecun, C. Moore, Eduard Sackinger, and R. Shah. 1993. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4).

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile – the munich versatile and fast open-source audio feature extractor. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 1459–1462.

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

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Navonil Majumder, Devamanyu Hazarika, Alexander F. Gelbukh, Erik Cambria, and Soujanya Poria. 2018a. Multimodal sentiment analysis using hierarchical fusion with context modeling. *CoRR*, abs/1806.06228.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2018b. Dialoguernn: An attentive RNN for emotion detection in conversations. *CoRR*, abs/1811.00405.

Marvin Minsky. 2007. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani,

Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Amir Hussain, and Alexander F. Gelbukh. 2018. Multimodal sentiment analysis: Addressing key issues and setting up baselines. *CoRR*, abs/1803.07427.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *CoRR*, abs/1905.02947.

Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. 2017. Improving speech recognition by revising gated recurrent units. *CoRR*, abs/1710.00641.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Garima Sharma and Abhinav Dhall. 2021. A survey on automatic multimodal emotion recognition in the wild. In *Advances in Data Science: Methodologies and Applications*, pages 35–64. Springer.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021a. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91, Online. Association for Computational Linguistics.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2021b. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea Samson. 2017. Facial expression analysis with affdex and facet: A validation study. *Behavior Research Methods*, 50.

Kurt Vonnegut. 1995. Shapes of stories. *Vonnegut's Shapes of Stories*.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at iest 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning.

Appendix

A Evaluation Metrics

Consider $\{y_n\}_{n=1}^N$ as the true labels and $\{\hat{y}\}_{n=1}^N$ as the predicted labels for the N datapoints. Note that $y_n, \hat{y}_n \in \{1, 2, \ldots, K\}$ where K is the number of classes.

The **Accuracy score** for predictions if given by:

$$\text{Accuracy} = \frac{\sum_{n=1}^N \mathbb{I}[y_n = \hat{y}_n]}{N}$$

We use the `accuracy_score` method of Python based Scikit-learn library (Buitinck et al., 2013) for its evaluation. The **F1 score** for a class k is given by:

$$\text{F1}_k = \frac{2 \times \text{precision}_k \times \text{recall}_k}{\text{precision}_k + \text{recall}_k}$$

where, precision_k is the precision for class k and recall_k is the recall for class k. These are calculated using:

$$\text{precision}_k = \frac{\sum_{\hat{y}_n = k} \mathbb{I}[y_n = \hat{y}_n]}{\sum_{\hat{y}_n = k} 1}$$

$$\text{recall}_k = \frac{\sum_{y_n = k} \mathbb{I}[y_n = \hat{y}_n]}{\sum_{y_n = k} 1}$$

Finally, the **weighted F1 score** is defined as

$$\text{weighted F1} = \sum_{k=1}^K f_k \times \text{F1}_k$$

where f_k is the relative frequency of class k

We use the `F1_score` method of Scikit-learn library for its evaluation.

B Experiment Reproducibility

B.1 Input and hidden states dimensions

The Input modality dimensions for the different datasets we experimented are as follows:

CMU-MOSEI

- Text (BERT): 768
- Audio (OpenSmile): 384
- Video (OpenFace2.0): 711

IEMOCAP:

- Text (BERT): 768
- Audio (OpenSmile): 100
- Video (3D CNN): 512

The dimension of the hidden states and GRU states are as follows:

- Siamese hidden state ($\mathcal{H}_t$): 300
- Party state for each modality $s_t^{q_t, m}$: 150
- Context State for each modality c_t^m: 150
- Emotion State for each modality e_t^m: 100

All other weights and parameters are such that the equations given in §3 hold.

There are a total of 5578803 parameters in the model.

B.2 Training the main model

All experiments are implemented using the Py-Torch library (Paszke et al., 2019). All weights are initialized randomly using PyTorch's default methods, and we use the Adam optimizer (Kingma and Ba, 2014) for training these weights.

The following hyper-parameters are used for the optimizer:

- Learning rate (lr) : 0.0001
- Weight decay (weight_decay): 0.0001
- β_1, β_2 (betas): (0.9, 0.999)

Here, the names in parenthesis denote the arguments corresponding to the hyper-parameters in the Adam Optimizer object of the PyTorch library.

We use a batch size of 128 for training across all experiments. The number of epochs for which the model was trained varies across datasets. These are listed as follows:

- CMU-MOSEI - 50 epochs
- IEMOCAP - 500 epochs

Training time per epoch was approximately 2.5 minutes for CMU-MOSEI and 15 seconds for IEMOCAP.

The model is evaluated at every epoch on the validation set (constructed using an 80:20 random split of the training data). The model giving the best weighted average F1 score across all classes is checkpointed. All the randomizations in the training procedure are reproducible using a seed value of 42 for libraries NumPy and PyTorch.

B.3 Training of the emotion shift component

This section provides the hyper-parameters for the pre-training procedure of the emotion shift component described in 4. We use a batch size of 8, and the model is pre-trained for five epochs. The model is checkpointed against the best F1 score.

B.4 Hyperparameter Tuning

Hyperparameters like the size of siamese hidden state ($\mathcal{H}_t$), size of context/party/hidden states ($s_t^{q_t,m}, c_t^m, e_t^m$) are tuned manually. The best weighted average F1 score over the validation set across all epochs was used as the criterion to select the best hyperparameter configuration.

To tune the hyperparameters used in the optimizer (learning rate, weight decay, β_1, β_2), we started with the default values used in the PyTorch library. These values are:

- learning rate: 0.001

- weight decay: 0

- β_1: 0.9

- β_2: 0.999

On manual tuning, we found that decreasing the learning rate to 0.0001 and increasing the weight decay to 0.0001 helped in better convergence and superior validation performance. Changing the values of β_1 and β_2 did not lead to any improvement. So these were kept the same as the default values.

B.5 Machine Specification

All experiments were performed on a server using Intel i7-5820K CPU @ 3.30GHz, Nvidia GeForce GTX TITAN X GPU, and CUDA 11.