Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis

Soujanya Poria[a], Erik Cambria[b], Devamanyu Hazarika[c],
Navonil Majumder[d], Amir Zadeh[e], Louis-Philippe Morency[e]

[a] Temasek Laboratories, Nanyang Technological University, Singapore
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[c] School of Computing, National University of Singapore, Singapore
[d] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[e] Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—Multimodal sentiment analysis involves identifying sentiment in videos and is a developing field of research. Unlike current works, which model utterances individually, we propose a recurrent model that is able to capture contextual information among utterances. In this paper, we also introduce attention-based networks for improving both context learning and dynamic feature fusion. Our model shows 6-8% improvement over the state of the art on a benchmark dataset.**

I. Introduction

Emotion recognition and sentiment analysis have become a new trend in social media, avidly helping users to understand opinions expressed in social networks and user-generated content [1]. With the advancement of communication technology, abundance of smartphones and the rapid rise of social media, large amounts of data are uploaded by web users in the form of videos, rather than text [2]. For example, consumers tend to record their reviews on products using a web camera and upload them on social media platforms such as YouTube or Facebook to inform subscribers of their opinions. These videos often contain comparisons of products from competing brands, the pros and cons of product specifications, etc., which can aid prospective buyers in making an informed decision.

An utterance is a segment of speech bounded by breaths or pauses. A review video often contains multiple utterances. The goal of utterance-level sentiment analysis is to label each utterance by its sentiment label. Utterance-level sentiment analysis facilitates the understanding of the reviewer' sentiment dynamics on multiple aspects of the review. Recently, a number of approaches have been proposed in the field of multimodal sentiment analysis [3], [4], [5]. All such approaches consider each utterance as an independent entity and, hence, ignore the relationship and dependencies between them. In a video, however, utterances maintain a sequence and can be highly correlated due to the development of the speaker's idea, co-reference and discourse structure. In particular, the classification of an utterance can benefit from the contextual information of other utterances. Modeling such contextual relationship, however, may not be enough. Identifying relevant and important information from the pool of utterances is necessary in order to make a model more robust and accurate.

To this end, we propose an attention-based long short-term memory (LSTM) network which not only models the contextual relationship among utterances, but also prioritizes more relevant utterances for classifying the target utterance. Experimental results show that the proposed framework outperforms the state of the art on benchmark datasets by 6-8%. Below, we describe the major contributions of the paper:

- We propose a contextual attention-based LSTM (CAT-LSTM) network to model the contextual relationship among utterances and prioritize the important contextual information for classification.
- We introduce an attention-based fusion mechanism, termed AT-Fusion, which amplifies the higher quality and informative modalities during fusion in multimodal classification.

The remainder of this paper is organized as follows: Section II describes the proposed method in detail; experimental results and discussion are shown in Section III; finally, Section IV concludes the paper.

II. Method

In the following subsections, we discuss the problem definition and explain the proposed approach in detail.

A. Problem Definition

Let us assume a video to be considered as $V_j = [u_{j,1}\ u_{j,2}\ u_{j,3}, ..., u_{j,i}...u_{j,L_j}]$ where $u_{j,i}$ is the i^{th} utterance in video v_j and L_j is the number utterances in the video. The goal of this approach is to label each utterance $u_{j,i}$ with the sentiment expressed by the speaker. We claim that, in order to classify utterance $u_{j,i}$, the other utterances in the video, i.e., $[u_{j,k} \mid \forall k \leq L_j, k \neq i]$, serve as its context and provide key information for the classification.

B. Overview of the Approach

The overview of the proposed approach is as follows:

1. **Unimodal feature extraction** We first extract utterance-level unimodal features from the respective unimodal classifiers. This phase does not consider the contextual relationship among the utterances.

2. **AT-Fusion – Multimodal fusion using the attention mechanism** In this step, the utterance-level unimodal features extracted at Step 1 are fused using an attention network (AT-Fusion) and the resulting output is used in the next step for sentiment classification.
3. **CAT-LSTM – Attention-based LSTM model for sentiment classification** CAT-LSTM is an attention-based LSTM network which accepts the features (output of Step 2) of a sequence of utterances per video and generates a new representation of those utterances based on the surrounding utterances.

C. Unimodal Feature Extraction

In this step, we extract unimodal features using dedicated unimodal feature extractors. The utterances are treated independently in this process.

1) **Textual Features Extraction:** We use a CNN for textual feature extraction, which takes utterances represented as a matrix of Google word2vec [6] vectors. Such vectors cover 87% of the vocabulary of CMU-MOSI dataset; the missing ones are initialized randomly. The convolution filters are then applied to this matrix of word vectors.

The CNN has two convolutional layers: the first layer has two kernels of size 3 and 4, with 50 feature maps each and the second layer has a kernel of size 2 with 100 feature maps. The convolution layers are interleaved with max-pooling layers of window 2×2. This is followed by a fully connected layer of size 500 and softmax output. We use ReLU as the activation function. The activation values of the fully-connected layer are taken as the features of utterances for text modality.

2) **Audio Feature Extraction:** Audio features are extracted with 30 Hz frame-rate and a sliding window of 100 ms using openSMILE toolkit. In order to identify samples with and without voice, voice normalization is performed using Z-standardization technique. The features extracted by openSMILE consist of several low-level descriptors, e.g., voice intensity, pitch, and their statistics, e.g., mean, root quadratic mean.

3) **Visual Feature Extraction:** There are various choices of deep networks specialized for image/video classification, e.g., cascaded CNN layers and recurrent neural networks (RNNs) such as LSTM and GRU. We chose 3D-CNN due to its proven ability to learn image representations (like 2D-CNN), along with the changes among the sequence of images (frames) in a video [7]. Let $V \in \mathbb{R}^{c \times f \times h \times w}$ represents an utterance video, where c = number of channels in an image (in our experiments $c = 3$, since the constituent images are RGB), f = number of frames, h = height of each frame, and w = width of each frame. We apply 3D convolutional filter F to video V, where $F \in \mathbb{R}^{f_m \times c \times f_d \times f_h \times f_w}$, f_m = number of feature maps, c = number of channels, f_d = number of frames, f_h = height of the filter, and f_w = width of the filter (we chose $F \in \mathbb{R}^{32 \times 3 \times 5 \times 5 \times 5}$). Following the philosophy of 2D-CNN, 3D-CNN slides filter F across video V and produces output $cvout \in \mathbb{R}^{fm \times c \times (f - f_d + 1) \times (h - f_h + 1) \times (w - f_w + 1)}$.

To discard irrelevant features, we apply max-pooling of window $3 \times 3 \times 3$ on $cvout$. Output of pooling layer is fed to a fully-connected layer of size 300, followed by a softmax layer for classification. The activations of the fully-connected layer is used as the features of video V.

D. AT-Fusion – Attention-Based Network for Multimodal Fusion

Attention mechanism has the ability to focus on the most important parts of an object relevant to the classification, improving the performance of the baseline deep neural networks. The attention mechanism has been successfully employed in NLP tasks such as sentiment analysis [8]. Not all modalities are equally relevant in the classification of sentiment. In order to prioritize only important modalities, we introduce an attention network, termed as AT-Fusion, which takes as an input audio, visual, and textual modalities and outputs an attention score for each modality.

We equalize the dimensions of the feature vectors of all three modalities prior to feeding them into the attention network. This is done using a fully-connected layer of size d. Let $B = [B_a, B_v, B_t]$ be the feature set after dimensionality equalization to size d, where B_a = acoustic features, B_v = visual features, and B_t = textual features; following $B \in \mathbb{R}^{d \times 3}$. The value of d when set to 300 gives best performance.

The attention weight vector α_{fuse} and the fused multimodal feature vector F are computed as follows:

$$P_F = tanh(W_F.B) \tag{1}$$

$$\alpha_{fuse} = softmax(w_F^T.P_F) \tag{2}$$

$$F = B.\alpha_{fuse}^T \tag{3}$$

Here, $W_F \in \mathbb{R}^{d \times d}$, $w_F \in \mathbb{R}^d$, $\alpha_{fuse}^T \in \mathbb{R}^3$, and $F \in \mathbb{R}^d$. We then feed the output F to the CAT-LSTM (Section II-E1, Figure 1) for final multimodal sentiment classification of the utterance.

E. Classifier: Context-Dependent Sentiment Classification

A speaker usually tries to gradually develop his/her idea and opinion about a product in the review, which makes the utterances in a video sequential, temporally and contextually dependent. This phenomenon motivates us to model inter-utterance relationship. To this end, we use a LSTM layer, in combination with the attention mechanism to amplify the important contextual evidences for sentiment classification of target utterance.

1) **Proposed CAT-LSTM for Sentiment Classification:** LSTM is a specialized RNN, which models long-range dependencies in a sequence. Specifically, LSTM solves the vanishing gradient problem of conventional RNNs, while modeling long-range dependencies. Current research in NLP indicates the benefit of using such networks to incorporate contextual information in the classification process [9], [10].

Let, $x \in \mathbb{R}^{d \times M}$ be input to the LSTM network, where M is the number of utterances in a video. The matrix x can be represented as $x = [x_1, x_2, ..., x_t, ...x_M]$, where $x_t \in \mathbb{R}^d$ for $t = 0$ to M.

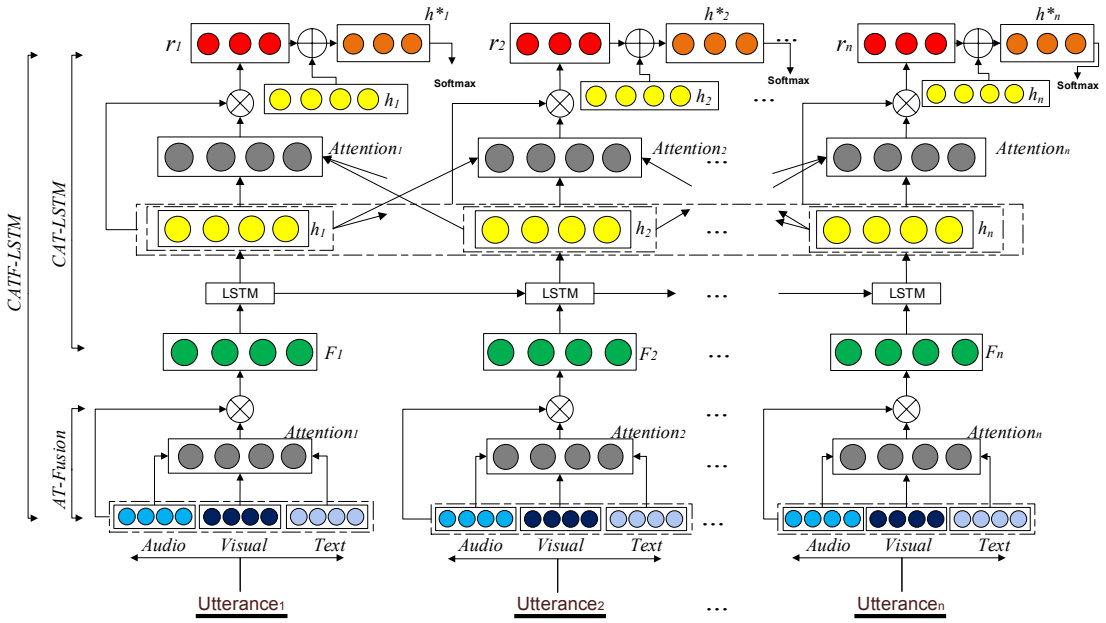


Fig. 1: CATF-LSTM takes input from multiple modalities, fuses them using AT-Fusion, and sends the output to CAT-LSTM for classification.

Each cell in LSTM can be computed as follows:

$$X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{4}$$

$$f_t = \sigma(W_f.X + b_f) \tag{5}$$

$$i_t = \sigma(W_i.X + b_i) \tag{6}$$

$$o_t = \sigma(W_o.X + b_o) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c.X + b_c) \tag{8}$$

$$h_t = o_t \odot tanh(c_t) \tag{9}$$

where $W_i, W_f, W_o \in \mathbb{R}^{d \times 2d}$, $b_i, b_f, b_o \in \mathbb{R}^d$ are parameters to be learnt during the training, σ is the sigmoid function and $\odot$ is element-wise multiplication.

The output of this LSTM layer is represented as $H \in \mathbb{R}^{d \times M}$, where $H = [h_1, h_2, ..., h_t, ..., h_M]$ and $h_i \in \mathbb{R}^d$. We feed the sequence of M utterance-level features (fused features F, obtained in equation (3) or unimodal features) to LSTM and obtain contextually-aware utterance representations H.

a) **Attention Network:** All surrounding utterances are not equally relevant in the sentiment classification of the target utterance. In order to amplify the contribution of context-rich utterances, we use an attention network.

Let A_t denote the t^{th} Attention network for utterance represented by h_t. The attention mechanism of A_t produces an attention weight vector α_t and a weighted hidden representation r_t as follows:

$$P_t = tanh(W_h[t].H) \tag{10}$$

$$\alpha_t = softmax(w[t]^T.P_t) \tag{11}$$

$$r_t = H.\alpha_t^T \tag{12}$$

where, $P_t \in \mathbb{R}^{d \times M}, \alpha_t \in \mathbb{R}^M, r_t \in \mathbb{R}^d$. And, $W_h \in \mathbb{R}^{M \times d \times d}, w \in \mathbb{R}^{M \times d}$ are projection parameters with $W_h[t]$ and $w[t]$ being used by the t^{th} attention model.

Finally, the LSTM representation for t^{th} utterance is modified as:

$$h_t^* = tanh(W_p[t].r_t + W_x[t].h_t) \tag{13}$$

Similar to the results obtained by Rocktäschel et al. [11], addition of the term $W_x[t].h_t$ to $W_p[t].r_t$ gives better result in the experiments carried out. Here, $h_t^* \in \mathbb{R}^d$ and $W_p, W_x \in \mathbb{R}^{M \times d \times d}$ are weights to be learnt while training. In some experiments (e.g., Section II-F2b), we use the output h_t^* as contextual features for further processing.

b) **Classification:** Finally, each modified LSTM cell output h_t^* is sent into a softmax layer for sentiment classification.

$$Z_t = softmax((h_t^*)^T.W_{soft}[t] + b_{soft}[t]) \tag{14}$$

$$\hat{y}_t = \arg\max_j(Z_t[j]), \quad \forall j \in class \tag{15}$$

where, $Z_t \in \mathbb{R}^{ydim}, W_{soft} \in \mathbb{R}^{M \times d \times ydim}, b_{soft} \in \mathbb{R}^{M \times ydim}, ydim$ = number of classes, and $\hat{y}_t$ is the predicted class.

F. Training

1) **Unimodal Classification:** In our work, we perform classification on two types of data – unimodal and multimodal. To classify the unimodal input, the extracted unimodal features (Section II-C) are sent to the CAT-LSTM network as inputs.

2) **Multimodal Classification:** For multimodal classification, the extracted unimodal features are first fed to the AT-Fusion to produce fused multimodal features. Then, such features are fed to the CAT-LSTM network for sentiment classification. We call this multimodal sentiment classification model as Contextual Attentive Fusion LSTM, i.e., CATF-LSTM. The CATF-LSTM is shown in Figure 1. Multimodal classification be accomplished using two different frameworks:

a) **Single-Level Framework:** In this framework, we fuse context-independent unimodal features as explained in Section II-D and feed those to CATF-LSTM for multimodal fusion and classification.

b) **Multi-Level Framework:** Contextual unimodal features can further improve the performance of the multimodal fusion framework explained in Section II-F2a. In this fusion scheme, we first send context-independent unimodal features extracted from every modality to CAT-LSTM. The contextual features yielded from CAT-LSTM are then fed to CATF-LSTM for fusion and final classification. Both the unimodal and multimodal classifiers are trained in an end-to-end manner using back propagation, with objective function being log-loss:

$$loss = -\sum_{i}\sum_{j} log(Z_t[y_i^j]) + \lambda||\theta||^2 \quad (16)$$

where, y = target class, Z_t = predicted distribution of j^{th} utterance from video V_i s.t. $i \in [0, N]$ and $j \in [0, L_i]$. λ is the L_2 regularization term and θ is the parameter set $\theta = \{W_i, b_i, W_f, b_f, W_o, b_o, W_F, w_F, W_h, w, W_p, W_x, W_{soft}, b_{soft}\}$.

In our experiments, we pad videos with dummy utterances to enable batch processing. Hence, we also use bit-masks to mitigate proliferation of noise in the network. The network is typically trained for 500-700 epochs with an early-stopping patience of 20 epochs. As optimizer, we use AdaGrad which is known to have improved robustness over SGD, given its ability to adapt the learning rate based on the parameters.

III. Experimental Results

In this section, we present the experimental results on different network variants in contrast with various baselines.

A. Dataset details

We perform person-independent experiments to emulate unseen conditions. Our train/test splits of the dataset are completely disjoint with respect to speakers.

a) **CMU-MOSI Dataset:** Zadeh et al. [12] constructed a multimodal sentiment analysis dataset termed Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI), consisting 2199 opinionated utterances, 93 videos by 89 speakers. The videos address a large array of topics, such as movies, books, and products. Videos were crawled from YouTube and segmented into utterances. Each of the 2199 utterances were labeled with its sentiment label, i.e., positive and negative. The train set comprises of the first 62 individuals in the dataset. So, the test set comprises of 31 videos by 27 speakers. In particular, we use 1447 utterances in the training and 752 utterances to test the models out of which 467 are negative and 285 are negative.

B. Different Models and Network Architectures

We have carried out experiments with both unidirectional and bi-directional LSTM with the later giving 0.3-0.7% better performance in all kinds of experiments. As this is an expected and non-critical outcome, we present all the results below using bi-directional LSTM variant. Additionally, we consider the following models in our experiments:

a) **Poria et al. (2015):** We have implemented and compared our method with the current state of the art approach, proposed by Poria et al. [4], who extracted visual features using CLM-Z, audio features using openSMILE, and textual features using CNN. Multiple kernel learning was then applied on the features obtained from the concatenation of the unimodal features. However, authors did not conduct any speaker-independent experiments.

b) **Poria et al. (2016):** This is an extended approach with respect to [4], which introduces a CNN-RNN feature extractor to extract visual features. We reimplemented this approach in our experiments.

c) **Unimodal-SVM:** We extract unimodal features (Section II-C) and concatenate them to produce multimodal features. A support vector machine (SVM) is applied on the resulting feature vector for the final sentiment classification.

d) **Simple-LSTM:** In this configuration, the extracted unimodal and multimodal features of the utterances are fed to a LSTM without attention mechanism.

e) **CAT-LSTM:** This is the simple contextual attention-based LSTM framework as described in Section II-E1. For multimodal setting, it accepts input generated by appending unimodal features.

f) **CATF-LSTM:** This model is used for multimodal classification. As explained in Section II-F, it consists of AT-Fusion and CAT-LSTM, where the output of AT-Fusion is fed to CAT-LSTM.

g) **ATS-Fusion:** In this variant, instead of feeding the output of AT-Fusion to the cells of CAT-LSTM, we feed to softmax classifiers. The utterances are treated independently in this case.

h) **Poria et al. (2015) + Our best model:** In order to perform a fair comparison with Poria et al. (2015), we feed the features extracted by their method to our best performing model.

i) **Poria et al. (2016) + Our best model:** This model is similar to the model *Poria et al. (2015) + Our best model*, except it uses the features extraction process from Poria et al. (2016).

C. Single-Level vs Multi-level Framework

Multi-level framework outperforms single-level framework in our experiments given the presence of contextual unimodal features (see Table II). Hence, for brevity, apart from Table II, we present only the results of multi-level framework.

D. AT-Fusion Performance

AT-Fusion employs the attention mechanism to fuse multiple modalities. In order to assess the effectiveness of AT-Fusion, we compare it with a simple fusion technique where the feature vectors from different modalities are appended and fed to the sentiment classifier, i.e., CAT-LSTM.

Table II presents the performance of CATF-LSTM, which utilizes AT-Fusion for feature fusion followed by CAT-LSTM for sentiment classification. Given AT-Fusion's ability to amplify the contribution of the important modalities during fusion, it unsurprisingly outperforms the simple fusion method.

It should be noted that AT-Fusion can be integrated with the other network variants, i.e., Simple-LSTM (Table III). Table III also shows that AT-Fusion with softmax output, i.e., ATS-Fusion, which outperforms the unimodal-SVM thanks to the superiority of the AT-Fusion over simple feature-append fusion.

E. Comparison Among the Models

a) **Comparison with the state of the art:** As shown in Table I, the proposed approach has outperformed the state of the art [4], [13] in the range of 6.25%-7.5%. We use the same set of textual and audio features used in [4], [13]. Notably, apart from using a different fusion mechanism, our method also uses a different visual feature extraction method. On the CMU-MOSI dataset, the proposed visual feature extraction method has outperformed the CLM-Z (used in [4]) and CNN-RNN (proposed by [13]). When we employ our best classifier, i.e., CATF-LSTM, on the features extracted by [4] and [13], performance of those methods have improved. Using CATF-LSTM on the features extracted by those methods, we obtained better results than both of them for audio-visual, visual-textual bimodal experiments. According to [4], [13] trimodal classifier outperforms all unimodal and bimodal classifiers. Hence, we compare our proposed method with those works in the trimodal experiment. From these experimental results (Table I), it is evident that the proposed contextual attention-based LSTM network and fusion methodology are the key to outperform the state of the art.

Models	A+V+T
Poria et al. (2015)	73.55%
Poria et al. (2016)	75.13%
Features of Poria et al. (2015)+ CATF-LSTM	79.40%
Features of Poria et al. (2016) + CATF-LSTM	80.25%
Our features + CATF-LSTM	**81.30%**

TABLE I: Comparison of state-of-the-art on multimodal classification with our network: CATF-LSTM. Metric used: macro-fscore. A=Audio;V=Visual;T=Textual.

b) **unimodal-SVM:** Our unimodal-SVM model yields comparable performance with the state of the art. However, simple-LSTM outperforms unimodal-SVM in all the experiments (Table I) as the latter is incapable of grasping the context information while classifying an utterance.

Modality	Single-Level		Multi-Level	
	Feat Append	AT-Fusion	Feat Append	AT-Fusion
A+V	61.0	**61.6**	62.4	**62.9**
A+T	78.5	**79.2**	79.5	**80.1**
V+T	77.6	**78.3**	79.6	**79.9**
A+V+T	78.9	**79.3**	81.0	**81.3**

TABLE II: Comparison between *single-level* and *multi-level* fusion mentioned in Section II-F2 using CAT-LSTM network. Feat Append=Unimodal features are appended and sent to CAT-LSTM. AT-Fusion is used with CAT-LSTM network. The table reports the macro-fscore of classification. A=Audio;V=Visual;T=Textual.

Modalities	Sentiment, on CMU-MOSI					
	Uni-SVM	*Simple-LSTM*		*CAT-LSTM*		
	feat-app	feat-app	AT-Fusion	feat-app	AT-Fusion	ATS-Fusion
A	58.1	59.5	-	60.1	-	-
V	53.4	54.9	-	55.5	-	-
T	75.5	77.2	-	79.1	-	-
A + V	58.6	61.4	61.8	62.4	62.9	59.1
A + T	75.8	78.5	79.1	79.5	80.1	76.3
V + T	76.7	78.7	79.1	79.6	79.9	77.5
A + V + T	**77.9**	**80.1**	**80.6**	**81.0**	**81.3**	78.3

TABLE III: Comparison of models mentioned in Section III-B. The table reports the macro-fscore of classification. Note: feat-appen=fusion by appending unimodal features. Multi-level framework is employed (See Section II-F2). A=Audio;V=Visual;T=Textual.

c) **CAT-LSTM vs Simple-LSTM:** From Table III, we can see that CAT-LSTM outperforms Simple-LSTM by 0.6-1.1% in unimodal experiments; 0.2-1% in bimodal experiments, and 0.9% in trimodal experiment. This again confirms that, even though both networks have access to contextual information, CAT-LSTM outperforms Simple-LSTM because of its attention capability to capture important contexts. As expected, CATF-LSTM further improves (0.3-0.6%) the performance of CAT-LSTM as it employs attention mechanism in fusion.

F. Importance of the Modalities

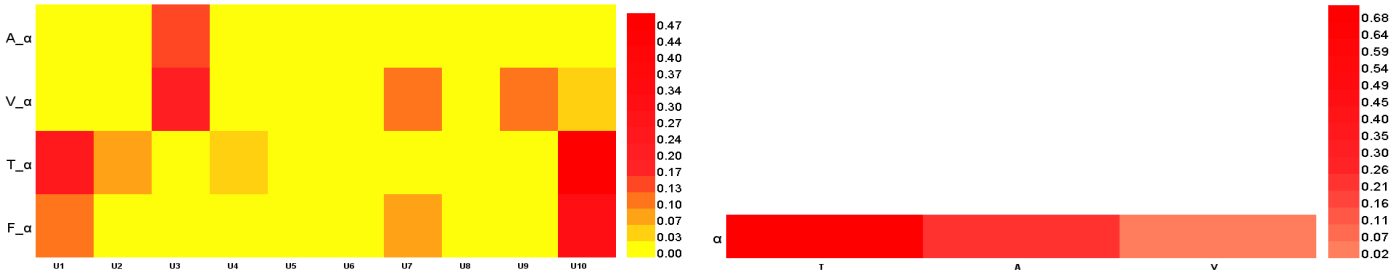
As expected, bimodal classifiers dominate unimodal classifiers and trimodal classifiers perform the best among all. Across modalities, textual modality performs better than the other two, thus indicating the need for better feature extraction for audio and video modalities.

G. Qualitative Analysis and Case Studies

In this section, we provide analysis and interesting observations on the learnt attention parameters for both contextual attention (CAT-LSTM) and attention fusion (AT-Fusion). Following, we list some of these observations. The need for considering context dependency (see Section I) is primal for utterance classification. For example, the utterance "*Whoever wrote this isn't the writer definitely*" has the sentiment expressed implicitly and, hence, baseline unimodal-SVM and state of the art fail to classify it correctly[1]. Information from neighboring utterances, e.g., *"And the dialogue threw me off"* and *"The whole movie had a really dry dialogue"*, indicate the negative context for the utterance. Such contextual relationships are prevalent throughout the dataset.

There are also cases where utterances are very ambiguous if considered separately because of the lack of context, e.g., "*You never know whats gonna happen*". In such cases, our context attention network attends to relevant utterances throughout the video to find contextual dependencies.

[1]RNTN classifies it as neutral. It can be seen here http://nlp.stanford.edu:8080/sentiment/rntnDemo.html



(a) The visualization of the attention scores of unimodal CAT-LSTM and trimodal CATF-LSTM. A_α, T_α, V_α, F_α represent attention scores of audio, textual, and visual fusion, respectively.

(b) Visualization of the attention weights of the AT-Fusion network in trimodal CATF-LSTM.

Fig. 2: Target utterance for classification - U4: You never know whats gonna happen. The input utterances are - U1: This is the most engaging endeavor yet. U2: And he puts them in very strange and unusual characters. U3: Umm what really need about this movie is the chemical brothers did the soundtracks whats pulse pounding the entire way through. U4: You never know whats gonna happen. U5: Theres all these colorful characters. U6: Now it isn't a great fantastic tell everybody about that kind of movie. U7: But I think its one of those movies thats so unique. U8: Its colourful. U9: Its in you face. U10: And something that I can't find anything else to compare it to.

Figure 2a shows the attention weights across the video for the above-mentioned utterance. While audio and visual provide decent attention vectors, text modality provides improved attention. It can be clearly seen that utterances like U_{10}, U_1 (Figure 2a) are the most relevant ones, which multimodal attention has been able to capture, thus proving its effectiveness. Interestingly, in this case the most important utterance U_{10} is located far from the target utterance U_4, proving the effectiveness of LSTM in modeling long distance sequence.

Figure 2b shows the contribution of each modality for the multimodal classification. Rightly, text has been given the highest weight by the attention fusion network, followed by audio and visual. Although the context dependency among utterances can be modeled in a simple LSTM network, there are often cases where utterances with complementary contexts are sparsely distributed across the video. In such situations, neighboring irrelevant utterances may provide negative bias for the utterance to be classified.

Our model, instead, provides an attention framework which focuses only on the relevant utterances throughout the video. For example, in one of the videos, the first utterance *"I am gonna give the reasons why I like him"* has its answers from the 7^{th} utterance onwards, with the intermediate utterances being irrelevant. In such situations, CAT-LSTM performs better than simple-LSTM model.

The effectiveness of AT-Fusion can also be seen in multiple cases. In one such instance, the audio quality of the utterance *"Sigh it probably would have helped if I went with someone"* was affected by loud background noise. In simple feature-append fusion models (e.g., unimodal-SVM), this utterance is misclassified because of the high noise present in the fusion. However, the multimodal attention-based fusion network (CATF-LSTM) correctly attends the video and text modality giving negligible attention on audio modality. This trend is also observed in many other cases. We finally observe that, in some cases, textual CAT-LSTM classifier performs better than trimodal CATF-LSTM for the presence of noise in the audio modality or when the speaker does not look directly at the camera while speaking.

IV. Conclusion

Multimodal sentiment analysis is a very hot research topic whose challenges are often underestimated. In this paper, we discarded the oversimplified assumption that utterances in a video are independent from each other. Hence, we developed a new framework that models contextual information obtained from other relevant utterances while classifying one target utterance. As demonstrated in the experiments, our framework outperforms the state of the art and paves the path for a more context-aware and meaning-preserving multimodal analysis.

References

[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[2] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[3] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, vol. 2, 2017, pp. 873–883.

[4] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015, pp. 2539–2544.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[8] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," *arXiv preprint arXiv:1605.08900*, 2016.

[9] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *ACL*, 2016, pp. 207–213.

[10] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," in *arXiv preprint arXiv:1708.02709*, 2017.

[11] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiskỳ, and P. Blunsom, "Reasoning about entailment with neural attention," *arXiv preprint arXiv:1509.06664*, 2015.

[12] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[13] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *ICDM*, 2016, pp. 439–448.