Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling

N. Majumder[a], D. Hazarika[b], S. Poria[c], A. Gelbukh[a], E. Cambria[c], R. Zimmermann[b]

[a]*Instituto Politécnico Nacional, Mexico*
[b]*National University of Singapore, Singapore*
[c]*Nanyang Technological University, Singapore*

Abstract

Multimodal sentiment analysis is a very actively growing field of research. A promising area of opportunity in this field is to improve the multimodal fusion mechanism. We present a novel feature fusion strategy that proceeds in a hierarchical fashion, first fusing the modalities two in two and only then fusing all three modalities. On multimodal sentiment analysis of individual utterances, our strategy outperforms conventional concatenation of features by 1%, which amounts to 5% reduction in error rate. On utterance-level multimodal sentiment analysis of multi-utterance video clips, for which current state-of-the-art techniques incorporate contextual information from other utterances of the same clip, our hierarchical fusion gives up to 2.4% (almost 10% error rate reduction) over currently used concatenation. The implementation of our method is publicly available in the form of open-source code.

1. Introduction

On numerous social media platforms, such as YouTube, Facebook, or Instagram, people share their opinions on all kinds of topics in the form of posts, images, and video clips. With the proliferation of smartphones and tablets, which has greatly boosted content sharing, people increasingly share their opinions on newly released products or on other topics in form of video reviews or comments. This is an excellent opportunity for large companies to capitalize on, by extracting user sentiment, suggestions, and complaints on their products from these video reviews. This information also opens new horizons to improving our quality of life by making informed decisions on the choice of products we buy, services we use, places we visit, or movies we watch basing on the experience and opinions of other users.

Videos convey information through three channels: audio, video, and text (in the form of speech). Mining opinions from this plethora of multimodal data calls for a solid multimodal sentiment analysis technology. One of the major problems faced in multimodal sentiment analysis is the fusion of features pertaining to different modalities. For this, the majority of the recent works in

multimodal sentiment analysis have simply concatenated the feature vectors of different modalities. However, this does not take into account that different modalities may carry conflicting information. We hypothesize that the fusion method we present in this paper deals with this issue better, and present experimental evidence showing improvement over simple concatenation of feature vectors. Also, following the state of the art (Poria et al., 2017a), we employ recurrent neural network (RNN) to propagate contextual information between utterances in a video clip, which significantly improves the classification results and outperforms the state of the art by a significant margin of 1–2% for all the modality combinations.

In our method, we first obtain unimodal features for each utterance for all three modalities. Then, using RNN we extract context-aware utterance features. Thus, we transform the context-aware utterance vectors to the vectors of the same dimensionality. We assume that these transformed vectors contain abstract features representing the attributes relevant to sentiment classification. Next, we compare and combine each bimodal combination of these abstract features using fully-connected layers. This yields fused bimodal feature vectors. Similarly to the unimodal case, we use RNN to generate context-aware features. Finally, we combine these bimodal vectors into a trimodal vector using, again, fully-connected layers and use a RNN to pass contextual information between them. We empirically show that the feature vectors obtained in this manner are more useful for the sentiment classification task.

The implementation of our method is publicly available in the form of open-source code.[1]

This paper is structured as follows: Section 2 briefly discusses important previous work in multimodal feature fusion; Section 3 describes our method in details; Section 4 reports the results of our experiments and discuss their implications; finally, Section 5 concludes the paper and discusses future work.

2. Related Work

In recent years, sentiment analysis (Cambria et al., 2017a) has become increasingly popular for processing social media data on online communities, blogs, Wikis, microblogging platforms, and other online collaborative media. Sentiment analysis is a branch of affective computing research that aims to classify text, audio and video into either positive or negative, but sometimes also neutral (Chaturvedi et al., 2017). Text-based sentiment analysis systems can be broadly categorized into knowledge-based (Cambria et al., 2016), statistics-based (Oneto et al., 2016), and hybrid (Cambria and Hussain, 2015). While most works approach it as a simple categorization problem, sentiment analysis is actually a suitcase research problem (Cambria et al., 2017b) that requires tackling many NLP tasks, including aspect extraction (Poria et al., 2016a), named entity recognition (Ma et al., 2016), word polarity disambiguation (Xia et al., 2015),

[1] `https://www.github.com/xxx/xxx/` (will be revealed upon acceptance)

temporal tagging (Zhong et al., 2017), personality recognition (Majumder et al., 2017), and sarcasm detection (Poria et al., 2016b). Sentiment analysis has raised growing interest both within the scientific community, leading to many exciting open challenges, as well as in the business world, due to the remarkable benefits to be had from financial forecasting (Xing et al., 2017) and political forecasting (Ebrahimi et al., 2017), e-health (Cambria et al., 2010) and e-tourism (Valdivia et al., 2017), community detection (Cavallari et al., 2017) and user profiling (Mihalcea and Garimella, 2016), and more.

In the field of emotion recognition, early works by De Silva et al. (1997) and Chen et al. (1998) showed that fusion of audio and visual systems, creating a bimodal signal, yielded a higher accuracy than any unimodal system. Such fusion has been analyzed at both feature level (Kessous et al., 2010) and decision level (Schuller, 2011).

Although there is much work done on audio-visual fusion for emotion recognition, exploring contribution of text along with audio and visual modalities in multimodal emotion detection has been little explored. Wollmer et al. (2013) and Rozgic et al. (2012) fused information from audio, visual and textual modalities to extract emotion and sentiment. Metallinou et al. (2008) and Eyben et al. (2010a) fused audio and textual modalities for emotion recognition. Both approaches relied on a feature-level fusion. Wu and Liang (2011) fused audio and textual clues at decision level. Poria et al. (2015) uses convolutional neural network (CNN) to extract features from the modalities and then employs multiple-kernel learning (MKL) for sentiment analysis. The current state of the art, set forth by Poria et al. (2017a), extracts contextual information from the surrounding utterances using long short-term memory (LSTM). Poria et al. (2017b) fuses different modalities with deep learning-based tools.

3. Our Method

In this section, we discuss our novel methodology behind solving the sentiment classification problem. First we discuss the overview of our method and then we discuss the whole method in details, step by step.

3.1. Overview

3.1.1. Unimodal Feature Extraction

We extract utterance-level features for three modalities. This step is discussed in Section 3.2.

3.1.2. Multimodal Fusion

Problems of early fusion. The majority of the work on multimodal data use concatenation, or early fusion (Fig. 1), as their fusion strategy. The problem with this simplistic approach is that it cannot filter out and conflicting or redundant information obtained from different modalities. To address this major issue, we devise an hierarchical approach which proceeds from unimodal to bimodal vectors and then bimodal to trimodal vectors.

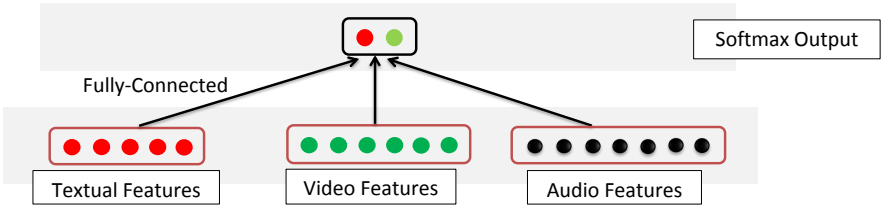

Figure 1: Utterance-level early fusion, or simple concatenation

Bimodal fusion. We fuse the utterance feature vectors for each bimodal combination, i.e., T+V, T+A, and A+V. This step is depicted in Fig. 2 and discussed in details in Section 3.4. We use the penultimate layer for Fig. 2 as bimodal features.

Trimodal fusion. We fuse the three bimodal features to obtain trimodal feature as depicted in Fig. 3. This step is discussed in details in Section 3.4.

Addition of context. We also improve the quality of feature vectors (both unimodal and multimodal) by incorporating information from surrounding utterances using RNN. We model the context using gated recurrent unit (GRU) as depicted in Fig. 4. The details of context modeling is discussed in Section 3.3 and the following subsections.

Classification. We classify the feature vectors using a softmax layer.

3.2. Unimodal Feature Extraction

In this section, we discuss the method of feature extraction for three different modalities: audio, video, and text.

3.2.1. Textual Feature Extraction

The source of textual modality is the transcription of the spoken words. To extract features from the textual modality, we use a deep convolutional neural network (CNN) (Karpathy et al., 2014). First, we represent each utterance as a concatenation of vectors of the constituent words, which in our experiments were the publicly available 300-dimensional `word2vec` vectors trained on 100 billion words from Google News (Mikolov et al., 2013).

The convolution kernels are thus applied to these concatenated word vectors instead of individual words. Each utterance is wrapped in a window of 50 words, which serves as the input to the CNN. 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 rectified linear unit (ReLU) (Teh and Hinton, 2001) as the activation function. The activation values of the fully-connected layer are taken as the

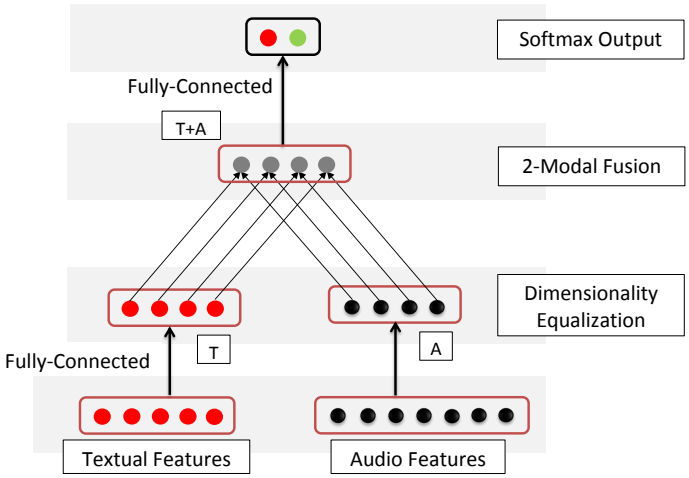

Figure 2: Utterance-level bimodal fusion

features of utterances for text modality. The convolution of the CNN over the utterance learns abstract representations of the phrases equipped with implicit semantic information, which with each successive layer spans over increasing number of words and ultimately the entire utterance.

3.2.2. Audio Feature Extraction with openSMILE

Audio features are extracted at 30 Hz frame rate with a sliding window of 100 ms. To compute the features, we use openSMILE (Eyben et al., 2010b), an open-source software that automatically extracts audio features such as pitch and voice intensity. Voice normalization is performed and voice intensity is thresholded to identify samples with and without voice. Z-standardization is used to perform voice normalization. Both of these tasks were performed using openSMILE.

The features extracted by openSMILE consist of several Low Level Descriptors (LLD) and statistical functionals of them. Some of the functionals are amplitude mean, arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, and linear regression slope. Specifically, we use "IS13-ComParE" configuration file in openSMILE. Taking into account all functionals of each LLD, we obtained 6372 features.

3.2.3. Visual Feature Extraction

We use 3D-CNN to obtain visual features from the video. We hypothesize that 3D-CNN will not only be able to learn relevant features from each frame, but will also be able to learn the changes among given number of consecutive frames.

In the past, 3D-CNN has been successfully applied to object classification on 3D data (Ji et al., 2013). Its ability to achieve state-of-the-art results motivated

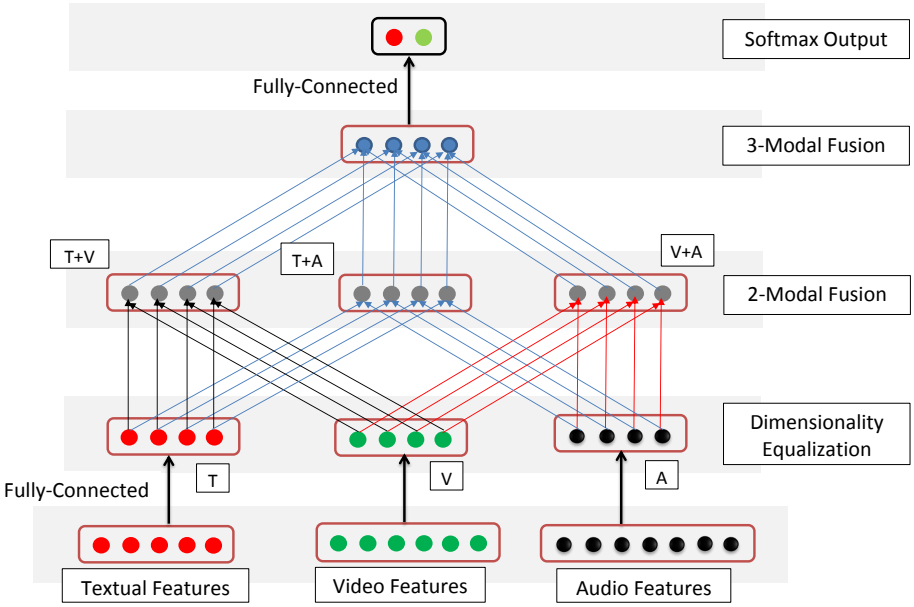

Figure 3: Utterance-level trimodal hierarchical fusion.[2]

us to use it.

Let $vid \in \mathbb{R}^{c \times f \times h \times w}$ be a video, where c = number of channels in an image (in our case $c = 3$, since we consider only RGB images), f = number of frames, h = height of the frames, and w = width of the frames. Again, we consider the 3D convolutional filter $filt \in \mathbb{R}^{fm \times c \times fd \times fh \times fw}$, where fm = number of feature maps, c = number of channels, fd = number of frames (in other words depth of the filter), fh = height of the filter, and fw = width of the filter. Similarly to 2D-CNN, $filt$ slides across video vid and generates output $convout \in \mathbb{R}^{fm \times c \times (f-fd+1) \times (h-fh+1) \times (w-fw+1)}$. Next, we apply max pooling to $convout$ to select only relevant features. The pooling will be applied only to the last three dimensions of the array $convout$.

In our experiments, we obtained best results with 32 feature maps (fm) with the filter size of $5 \times 5 \times 5$ (or $fd \times fh \times fw$). In other words, the dimension of the filter is $32 \times 3 \times 5 \times 5 \times 5$ (or $fm \times c \times fd \times fh \times fw$). Subsequently, we apply max pooling on the output of convolution operation, with window size being $3 \times 3 \times 3$. This is followed by a dense layer of size 300 and softmax. The activations of this dense layer are finally used as the video features for each utterance.

[2]Figure adapted from (Majumder, 2017) with permission.

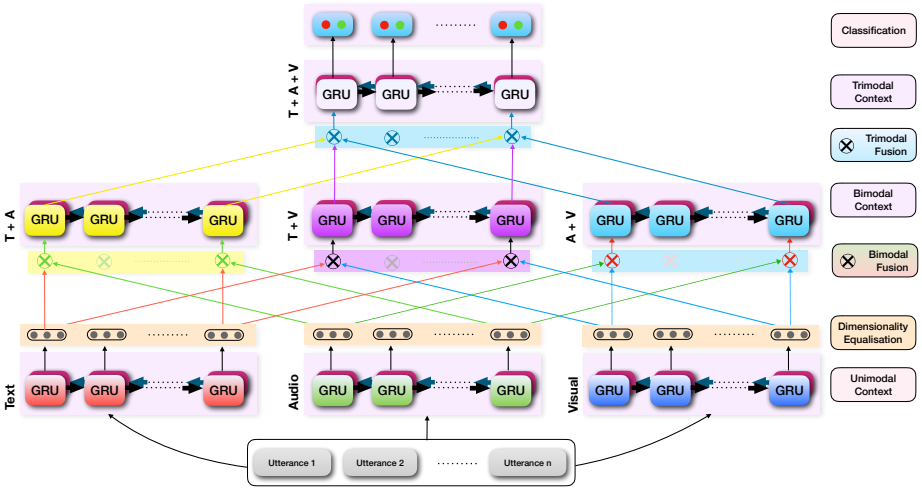

Figure 4: Context-aware hierarchical fusion

3.3. Context Modeling

Utterances in the videos are semantically dependent on each other. In other words, complete meaning of an utterance may be determined by taking preceding utterances into consideration. We call this the context of an utterance. Following Poria et al. (2017a), we use RNN, specifically GRU[3] to model semantic dependency among the utterances in a video.

Let the following items represent unimodal features:

$$\begin{aligned} f_A &\in \mathbb{R}^{N \times d_A} \quad \text{(acoustic features)}, \\ f_V &\in \mathbb{R}^{N \times d_V} \quad \text{(visual features)}, \\ f_T &\in \mathbb{R}^{N \times d_T} \quad \text{(textual features)}, \end{aligned}$$

where N = maximum number of utterances in a video. We pad the shorter videos with dummy utterances represented by null vectors of corresponding length. For each modality, we feed the unimodal utterance features f_m (where $m \in \{A, V, T\}$) (discussed in Section 3.2) of a video to GRU_m with output size D_m, which is defined as

$$\begin{aligned} z_m &= \sigma(f_{mt} U^{mz} + s_{m(t-1)} W^{mz}), \\ r_m &= \sigma(f_{mt} U^{mr} + s_{m(t-1)} W^{mr}), \\ h_{mt} &= \tanh(f_{mt} U^{mh} + (s_{m(t-1)} * r_m) W^{mh}), \\ F_{mt} &= \tanh(h_{mt} U^{mx} + u^{mx}), \\ s_{mt} &= (1 - z_m) * F_{mt} + z_m * s_{m(t-1)}, \end{aligned}$$

[3]LSTM does not perform well

where $U^{mz} \in \mathbb{R}^{d_m \times D_m}$, $W^{mz} \in \mathbb{R}^{D_m \times D_m}$, $U^{mr} \in \mathbb{R}^{d_m \times D_m}$, $W^{mr} \in \mathbb{R}^{D_m \times D_m}$, $U^{mh} \in \mathbb{R}^{d_m \times D_m}$, $W^{mh} \in \mathbb{R}^{D_m \times D_m}$, $U^{mx} \in \mathbb{R}^{d_m \times D_m}$, $u^{mx} \in \mathbb{R}^{D_m}$, $z_m \in \mathbb{R}^{D_m}$, $r_m \in \mathbb{R}^{D_m}$, $h_{mt} \in \mathbb{R}^{D_m}$, $F_{mt} \in \mathbb{R}^{D_m}$, and $s_{mt} \in \mathbb{R}^{D_m}$. This yields hidden outputs F_{mt} as context-aware unimodal features for each modality. Hence, we define $F_m = GRU_m(f_m)$, where $F_m \in \mathbb{R}^{N \times D_m}$. Thus, the context-aware multimodal features can be defined as

$$\begin{aligned} F_A &= GRU_A(f_A), \\ F_V &= GRU_V(f_V), \\ F_T &= GRU_T(f_T). \end{aligned}$$

3.4. Multimodal Fusion

In this section, we use context-aware unimodal features F_A, F_V, and F_T to a unified feature space.

The unimodal features may have different dimensions, i.e., $D_A \neq D_V \neq D_T$. Thus, we map them to the same dimension, say D (we obtained best results with $D = 400$), using fully-connected layer as follows:

$$\begin{aligned} g_A &= \tanh(F_A W_A + b_A), \\ g_V &= \tanh(F_V W_V + b_V), \\ g_T &= \tanh(F_T W_T + b_T), \end{aligned}$$

where $W_A \in \mathbb{R}^{D_A \times D}$, $b_A \in \mathbb{R}^D$, $W_V \in \mathbb{R}^{D_V \times D}$, $b_V \in \mathbb{R}^D$, $W_T \in \mathbb{R}^{D_T \times D}$, and $b_T \in \mathbb{R}^D$. We can represent the mapping for each dimension as

$$g_x = \begin{bmatrix} c^x_{11} & c^x_{21} & c^x_{31} & \cdots & c^x_{D1} \\ c^x_{12} & c^x_{22} & c^x_{32} & \cdots & c^x_{D2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ c^x_{1N} & c^x_{2N} & c^x_{3N} & \cdots & c^x_{DN} \end{bmatrix},$$

where $x \in \{V, A, T\}$ and c^x_{lt} are scalars for all $l = 1, 2, \ldots, D$ and $t = 1, 2, \ldots, N$. Also, in g_x the rows represent the utterances and the columns the feature values. We can see these values c^x_{lt} as more abstract feature values derived from fundamental feature values (which are the components of f_A, f_V, and f_T). For example, an abstract feature can be the angriness of a speaker in a video. We can infer the degree of angriness from visual features (f_V; facial muscle movements), acoustic features (f_A, such as pitch and raised voice), or textual features (f_T, such as the language and choice of words). Therefore, the degree of angriness can be represented by c^x_{lt}, where x is A, V, or T, l is some fixed integer between 1 and D, and t is some fixed integer between 1 and N.

Now, the evaluation of abstract feature values from all the modalities may not have the same merit or may even contradict each other. Hence, we need the network to make comparison among the feature values derived from different modalities to make a more refined evaluation of the degree of anger. To this end, we take each bimodal combination (which are audio–video, audio–text,

and video–text) at a time and compare and combine each of their respective abstract feature values (i.e. c_{lt}^V with c_{lt}^T, c_{lt}^V with c_{lt}^A, and c_{lt}^A with c_{lt}^T) using fully-connected layers as follows:

$$i_{lt}^{VA} = \tanh(w_l^{VA}.[c_{lt}^V, c_{lt}^A]^\intercal + b_l^{VA}), \tag{1}$$
$$i_{lt}^{AT} = \tanh(w_l^{AT}.[c_{lt}^A, c_{lt}^T]^\intercal + b_l^{AT}), \tag{2}$$
$$i_{lt}^{VT} = \tanh(w_l^{VT}.[c_{lt}^V, c_{lt}^T]^\intercal + b_l^{VT}), \tag{3}$$

where $w_l^{VA} \in \mathbb{R}^2$, b_l^{VA} is scalar, $w_l^{AT} \in \mathbb{R}^2$, b_l^{AT} is scalar, $w_l^{VT} \in \mathbb{R}^2$, and b_l^{VT} is scalar, for all $l = 1, 2, \ldots, D$ and $t = 1, 2, \ldots, N$. We hypothesize that it will enable the network to compare the decisions from each modality against the others and help achieve a better fusion of modalities.

Bimodal fusion. Eqs. (1) to (3) are used for bimodal fusion. The bimodal fused features for video–audio, audio–text, video–text are defined as

$$f_{VA} = (f_{VA1}, f_{VA2}, \ldots, f_{VA(N)}), \text{ where } f_{VAt} = (i_{1t}^{VA}, i_{2t}^{VA}, \ldots, i_{Dt}^{VA}),$$
$$f_{AT} = (f_{AT1}, f_{AT2}, \ldots, f_{AT(N)}), \text{ where } f_{ATt} = (i_{1t}^{AT}, i_{2t}^{AT}, \ldots, i_{Dt}^{AT}),$$
$$f_{VT} = (f_{VT1}, f_{VT2}, \ldots, f_{VT(N)}), \text{ where } f_{VTt} = (i_{1t}^{VT}, i_{2t}^{VT}, \ldots, i_{Dt}^{VT}).$$

We further employ GRU_m(Section 3.3) ($m \in \{VA, VT, TA\}$), to incorporate contextual information among the utterances in a video with

$$F_{VA} = (F_{VA1}, F_{VA2}, \ldots, F_{VA(N)}) = GRU_{VA}(f_{VA}),$$
$$F_{VT} = (F_{VT1}, F_{VT2}, \ldots, F_{VT(N)}) = GRU_{VT}(f_{VT}),$$
$$F_{TA} = (F_{TA1}, F_{TA2}, \ldots, F_{TA(N)}) = GRU_{TA}(f_{TA}),$$

where

$$F_{VAt} = (I_{1t}^{VA}, I_{2t}^{VA}, \ldots, I_{D_2t}^{VA}),$$
$$F_{VTt} = (I_{1t}^{AT}, I_{2t}^{AT}, \ldots, I_{D_2t}^{AT}),$$
$$F_{TAt} = (I_{1t}^{VT}, I_{2t}^{VT}, \ldots, I_{D_2t}^{VT}),$$

F_{VA}, F_{VT}, and F_{TA} are context-aware bimodal features represented as vectors and I_{nt}^m is scalar for $n = 1, 2, \ldots, D_2$, $D_2 = 500$, $t = 1, 2, \ldots, N$, and $m =$ VA,VT,TA.

Trimodal fusion. We combine all three modalities using fully-connected layers as follows:

$$z_{lt} = \tanh(w_l^{AVT}.[I_{lt}^{VA}, I_{lt}^{AT}, I_{lt}^{VT}]^\intercal + b_l^{AVT}),$$

where $w_l^{AVT} \in \mathbb{R}^3$ and b_l^{AVT} is a scalar for all $l = 1, 2, \ldots, D_2$ and $t = 1, 2, \ldots, N$. So, we define the fused features as

$$f_{AVT} = (f_{AVT1}, f_{AVT2}, \ldots, f_{AVT(N)}),$$

where $f_{AVTt} = (z_{1t}, z_{2t}, \ldots, z_{D_2 t})$, z_{nt} is scalar for $n = 1, 2, \ldots, D_2$ and $t = 1, 2, \ldots, N$.

Similarly to bimodal fusion (Section 3.4), after trimodal fusion we pass the fused features through GRU_{AVT} to incorporate contextual information in them, which yields

$$F_{AVT} = (F_{AVT1}, F_{AVT2}, \ldots, F_{AVT(N)}) = GRU_{AVT}(f_{AVT}),$$

where $F_{AVTt} = (Z_{1t}, Z_{2t}, \ldots, Z_{D_3 t})$, Z_{nt} is scalar for $n = 1, 2, \ldots, D_3$, $D_3 = 550$, $t = 1, 2, \ldots, N$, and F_{AVT} is the context-aware trimodal feature vector.

3.5. Classification

In order to perform classification, we feed the fused features F_{mt} (where $m = AV, VT, TA,$ or AVT and $t = 1, 2, \ldots, N$) to a softmax layer with $C = 2$ outputs. The classifier can be described as follows:

$$\begin{aligned} \mathcal{P} &= \text{softmax}(W_{softmax} F_{mt} + b_{softmax}), \\ \hat{y} &= \underset{j}{\text{argmax}}(\mathcal{P}[j]), \end{aligned}$$

where $W_{softmax} \in \mathbb{R}^{C \times D}$, $b_{softmax} \in \mathbb{R}^C$, $\mathcal{P} \in \mathbb{R}^C$, j = class value (0 or 1), and $\hat{y}$ = estimated class value.

3.6. Training

We employ categorical cross-entropy as loss function (J) for training,

$$J = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{C-1} y_{ij} \log \mathcal{P}_i[j],$$

where N = number of samples, i = index of a sample, j = class value, and

$$y_{ij} = \begin{cases} 1, & \text{if expected class value of sample } i \text{ is } j \\ 0, & \text{otherwise.} \end{cases}$$

Adam (Kingma and Ba, 2014) is used as optimizer due to its ability to adapt learning rate for each parameter individually. We train the network for 200 epochs with early stopping, where we optimize the parameter set

$$\begin{aligned} \theta = & \bigcup_{m \in M} \left(\bigcup_{j \in \{z,r,h\}} \{U^{mj}, W^{mj}\} \cup \{U^{mx}, u^{mx}\} \right) \\ & \cup \bigcup_{m \in M_2} \bigcup_{i=1}^{D_2} \{w_i^m\} \cup \bigcup_{i=1}^{D_3} \{w_i^{AVT}\} \cup \bigcup_{m \in M_1} \{W_m, b_m\} \\ & \cup \{W_{softmax}, b_{softmax}\}, \end{aligned}$$

where $M = \{A, V, T, VA, VT, TA, AVT\}$, $M_1 = \{A, V, T\}$, and $M_2 = \{VA, VT, TA\}$. Algorithm 1 summarizes our method.[4]

4. Experiments

4.1. Dataset Details

Most research works in multimodal sentiment analysis are performed on datasets where train and test splits may share certain speakers. Since, each individual has an unique way of expressing emotions and sentiments, finding generic and person-independent features for sentiment analysis is crucial.

4.1.1. CMU-MOSI

CMU-MOSI dataset (Zadeh et al., 2016) is rich in sentimental expressions, where 89 people review various topics in English. The videos are segmented into utterances where each utterance is annotated with scores between -3 (strongly negative) and $+3$ (strongly positive) by five annotators. We took the average of these five annotations as the sentiment polarity and considered only two classes (positive and negative). Given every individual's unique way of expressing sentiments, real world applications should be able to model generic person independent features and be robust to person variance. To this end, we perform person-independent experiments to emulate unseen conditions. Our train/test splits of the dataset are completely disjoint with respect to speakers. The train/validation set consists of the first 62 individuals in the dataset. The test set contains opinionated videos by rest of the 31 speakers. In particular, 1447 and 752 utterances are used for training and test respectively.

4.1.2. IEMOCAP

IEMOCAP (Busso et al., 2008) contains two way conversations among ten speakers, segmented into utterances. The utterances are tagged with the labels anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. We consider the first four ones to compare with the state of the art (Poria et al., 2017a) and other works. It contains 1083 angry, 1630 happy, 1083 sad, and 1683 neutral videos. Only the videos by the first eight speakers are considered for training.

4.2. Baselines

We compare our method with the following strong baselines.

Early fusion. We extract unimodal features (Section 3.2) and simply concatenate them to produce multimodal features. Followed by support vector machine (SVM) being applied on this feature vector for the final sentiment classification.

[4]Implementation of this algorithm is available at `https://www.github.com/xxx/xxx/` (will be revealed upon acceptance)

Algorithm 1 Context-Aware Hierarchical Fusion Algorithm

procedure TRAINANDTESTMODEL(U, V) ▷ U = train set, V = test set

  Unimodal feature extraction:

  for `i:[1,N]` **do** ▷ extract baseline features

    $f_A^i \leftarrow AudioFeatures(u_i)$

    $f_V^i \leftarrow VideoFeatures(u_i)$

    $f_T^i \leftarrow TextFeatures(u_i)$

  for `m` $\in \{A, V, T\}$ **do**

    $F_m = GRU_m(f_m)$

  Fusion:

  $g_A \leftarrow MapToSpace(F_A)$ ▷ dimensionality equalization

  $g_V \leftarrow MapToSpace(F_V)$

  $g_T \leftarrow MapToSpace(F_T)$

  $f_{VA} \leftarrow BimodalFusion(g_V, g_A)$ ▷ bimodal fusion

  $f_{AT} \leftarrow BimodalFusion(g_A, g_T)$

  $f_{VT} \leftarrow BimodalFusion(g_V, g_T)$

  for `m` $\in \{VA, AT, VT\}$ **do**

    $F_m = GRU_m(f_m)$

  $f_{AVT} \leftarrow TrimodalFusion(F_{VA}, F_{AT}, F_{VT})$ ▷ trimodal fusion

  $F_{AVT} = GRU_{AVT}(f_{AVT})$

  for `i:[1,N]` **do** ▷ softmax classification

    $\hat{y}^i = \underset{j}{\text{argmax}}(softmax(F_{AVT}^i)[j])$

  $TestModel(V)$

procedure MAPTOSPACE(x_z) ▷ for modality z

  $g_z \leftarrow \tanh(W_z x_z + b_z)$

  return g_z

procedure BIMODALFUSION(g_{z_1}, g_{z_2}) ▷ for modality z_1 and z_2, where $z_1 \neq z_2$

  for `i:[1,D]` **do**

    $f_{z_1 z_2}^i \leftarrow \tanh(w_i^{z_1 z_2}.[g_{z_1}^i, g_{z_2}^i]^\intercal + b_i^{z_1 z_2})$

  $f_{z_1 z_2} \leftarrow (f_{z_1 z_2}^1, f_{z_1 z_2}^2, \ldots, f_{z_1 z_2}^D)$

  return $f_{z_1 z_2}$

procedure TRIMODALFUSION(f_{z_1}, f_{z_2}, f_{z_3}) ▷ for modality combination z_1, z_2, and z_3, where $z_1 \neq z_2 \neq z_3$

  for `i:[1,D]` **do**

    $f_{z_1 z_2 z_3}^i \leftarrow \tanh(w_i.[f_{z_1}^i, f_{z_2}^i, f_{z_3}^i]^\intercal + b_i)$

  $f_{z_1 z_2 z_3} \leftarrow (f_{z_1 z_2 z_3}^1, f_{z_1 z_2 z_3}^2, \ldots, f_{z_1 z_2 z_3}^D)$

  return $f_{z_1 z_2 z_3}$

procedure TESTMODEL(V)

  Similarly to training phase, V is passed through the learnt models to get the features and classification outputs. Section 3.6 mentions the trainable parameters (θ).

Method from (Poria et al., 2015). We have implemented and compared our method with the approach proposed by Poria et al. (2015). In their approach, they extracted visual features using CLM-Z, audio features using openSMILE, and textual features using CNN. MKL was then applied to the features obtained

Table 1: Comparison in terms of accuracy of Hierarchical Fusion (HFusion) with other fusion methods for CMU-MOSI dataset; bold font signifies best accuracy for the corresponding feature set and modality or modalities, where T stands for text, V for video, and A for audio.

Modality	(Poria et al., 2015) feature set		Our feature set	
Combination	(Poria et al., 2015)	HFusion	Early fusion	HFusion
T	N/A		75.0%	
V	N/A		55.3%	
A	N/A		56.9%	
T+V	73.2%	**74.4%**	77.1%	**77.8%**
T+A	73.2%	**74.2%**	77.1%	**77.3%**
A+V	55.7%	**57.5%**	56.5%	**56.8%**
A+V+T	73.5%	**74.6%**	77.0%	**77.9%**

from concatenation of the unimodal features. However, they did not conduct speaker independent experiments.

In order to perform a fair comparison with (Poria et al., 2015), we employ our fusion method on the features extracted by Poria et al. (2015).

Method from (Poria et al., 2017a). We have compared our method with (Poria et al., 2017a), which takes advantage of contextual information obtained from the surrounding utterances. This context modeling is achieved using LSTM. We reran the experiments of Poria et al. (2017a) without using SVM for classification since using SVM with neural networks is usually discouraged. This provides a fair comparison with our model which does not use SVM.

4.3. Experimental Setting

We considered two variants of experimental setup while evaluating our model.

HFusion. In this setup, we evaluated hierarchical fusion without context-aware features with CMU-MOSI dataset. We removed all the GRUs from the model described in Sections 3.3 and 3.4 forwarded utterance specific features directly to the next layer. This setup is described in depicted in Fig. 3.

CHFusion. This setup is exactly as the model is described in Section 3.

4.4. Results and Discussion

We discuss the results for the different experimental settings discussed in Section 4.3.

4.4.1. Hierarchical Fusion (HFusion)

The results of our experiments are presented in Table 1. We evaluated this setup with CMU-MOSI dataset (Section 4.1.1) and two feature sets: the feature set used in (Poria et al., 2015) and the set of unimodal features discussed in Section 3.2.

Table 2: Comparison of Context-Aware Hierarchical Fusion (CHFusion) in terms of accuracy with the state of the art for CMU-MOSI and IEMOCAP dataset; bold font signifies best accuracy for the corresponding dataset and modality or modalities, where T stands text, V for video, A for audio, and SOTA for state of the art (Poria et al., 2017a).

Modality	CMU-MOSI		IEMOCAP	
Combination	SOTA	CHFusion	SOTA	CHFusion
T	76.5%		73.6%	
V	54.9%		53.3%	
A	55.3%		57.1%	
T+V	77.8%	**79.3%**	74.1%	**75.9%**
T+A	77.3%	**79.1%**	73.7%	**76.1%**
A+V	57.9%	**58.8%**	68.4%	**69.5%**
A+V+T	78.7%	**80.0%**	74.1%	**76.5%**

Our model outperformed (Poria et al., 2015), which employed MKL, for all bimodal and trimodal scenarios by a margin of 1–1.8%. This leads us to present two observations. Firstly, the features used in (Poria et al., 2015) is inferior to the features extracted in our approach. Second, our hierarchical fusion method is better than their fusion method.

It is already established in the literature (Poria et al., 2015; Pérez-Rosas et al., 2013) that multimodal analysis outperforms unimodal analysis. We also observe the same trend in our experiments where trimodal and bimodal classifiers outperform unimodal classifiers. The textual modality performed best among others with a higher unimodal classification accuracy of 75%. Although other modalities contribute to improve the performance of multimodal classifiers, that contribution is little in compare to the textual modality.

On the other hand, we compared our model with early fusion (Section 4.2) for aforementioned feature sets (Section 3.2). Our fusion mechanism consistently outperforms early fusion for all combination of modalities. This supports our hypothesis that our hierarchical fusion method captures the inter-relation among the modalities and produce better performance vector than early fusion. Text is the strongest individual modality, and we observe that the text modality paired with remaining two modalities results in consistent performance improvement.

Overall, the results give a strong indication that the comparison among the abstract feature values dampens the effect of less important modalities, which was our hypothesis. For example, we can notice that for early fusion T+V and T+A both yield the same performance. However, with our method text with video performs better than text with audio, which is more aligned with our expectations, since facial muscle movements usually carry more emotional nuances than voice.

4.4.2. Context-Aware Hierarchical Fusion (CHFusion)

The results of this experiment are shown in Table 2. This setting fully utilizes the model described in Section 3. We applied this experimental setting for two

datasets, namely CMU-MOSI (Section 4.1.1) and IEMOCAP (Section 4.1.2). We used the feature set discussed in Section 3.2, which was also used by Poria et al. (2017a).

CMU-MOSI. We achieve 1–2% performance improvement over the state of the art (Poria et al., 2017a) for all the modality combinations having textual component. For A+V modality combination we achieve better but similar performance to the state of the art. We suspect that it is due to both audio and video modality being severely less informative than textual modality. It is evident from the unimodal performance where we observe that textual modality on its own performs around 21% better than both audio and video modality. Also, audio and video modality performs close to majority baseline. On the other hand, it is important to notice that with all modalities combined we achieve about 3.5% higher accuracy than text alone.

For example, consider the following utterance: *so overall new moon even with the bigger better budgets huh it was still too long.* The speaker discusses her opinion on the movie Twilight New Moon. Textually the utterance is abundant with positive words however audio and video comprises of a frown which is observed by the hierarchical fusion based model.

IEMOCAP. Here as well, we achieve performance improvement consistent with CMU-MOSI. This method performs 1–2.4% better than the state of the art for all the modality combinations. Also, trimodal accuracy is 3% higher than the same for textual modality. One key observation for IEMOCAP dataset is that its A+V modality combination performs significantly better than the same of CMU-MOSI dataset. We think that this is due to audio and video modality of IEMOCAP being richer than the same of CMU-MOSI.

4.4.3. HFusion vs. CHFusion

We compare HFusion and CHFusion models over CMU-MOSI dataset. We observe that CHFusion performs 1–2% better than HFusion model for all the modality combinations. This performance boost is achieved by the inclusion of utterance-level contextual information in HFusion model by adding GRUs in different levels of fusion hierarchy.

5. Conclusion

Multimodal fusion strategy is an important issue in multimodal sentiment analysis. However, little work has been done so far in this direction. In this paper, we have presented a novel and comprehensive fusion strategy. Our method outperforms the widely used early fusion on both datasets typically used to test multimodal sentiment analysis methods. Moreover, with the addition of context modeling with GRU, our method outperforms the state of the art in multimodal sentiment analysis and emotion detection by significant margin.

In our future work, we plan to improve the quality of unimodal features, especially textual features, which will further improve the accuracy of classification. We will also experiment with more sophisticated network architectures.

Acknowledgement

The work was partially supported by the Instituto Politécnico Nacional via grant SIP 20172008 to A. Gelbukh.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, 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), Association for Computational Linguistics, Vancouver, Canada, 873–883, URL `http://aclweb.org/anthology/P17-1081`, 2017a.

E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, A Practical Guide to Sentiment Analysis, Springer, Cham, Switzerland, 2017a.

I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Bayesian network based extreme learning machine for subjectivity detection, Journal of The Franklin Institute .

E. Cambria, S. Poria, R. Bajpai, B. Schuller, SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives, in: COLING, 2666–2677, 2016.

L. Oneto, F. Bisio, E. Cambria, D. Anguita, Statistical learning theory and ELM for big social data analysis, IEEE Computational Intelligence Magazine 11 (3) (2016) 45–55.

E. Cambria, A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, Cham, Switzerland, 2015.

E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment Analysis is a Big Suitcase, IEEE Intelligent Systems 32 (6).

S. Poria, I. Chaturvedi, E. Cambria, F. Bisio, Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis, in: IJCNN, 4465–4473, 2016a.

Y. Ma, E. Cambria, S. Gao, Label embedding for zero-shot fine-grained named entity typing, in: COLING, Osaka, 171–180, 2016.

Y. Xia, E. Cambria, A. Hussain, H. Zhao, Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features, Cognitive Computation 7 (3) (2015) 369–380.

X. Zhong, A. Sun, E. Cambria, Time expression analysis and recognition using syntactic token types and general heuristic rules, in: ACL, 420–429, 2017.

N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning based document modeling for personality detection from text, IEEE Intelligent Systems 32 (2) (2017) 74–79.

S. Poria, E. Cambria, D. Hazarika, P. Vij, A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks, in: COLING, 1601–1612, 2016b.

F. Xing, E. Cambria, R. Welsch, Natural Language Based Financial Forecasting: A Survey, Artificial Intelligence Review .

M. Ebrahimi, A. Hossein, A. Sheth, Challenges of sentiment analysis for dynamic events, IEEE Intelligent Systems 32 (5).

E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, J. Munro, Sentic Computing for Patient Centered Application, in: IEEE ICSP, Beijing, 1279–1282, 2010.

A. Valdivia, V. Luzon, F. Herrera, Sentiment analysis in TripAdvisor, IEEE Intelligent Systems 32 (4) (2017) 2–7.

S. Cavallari, V. Zheng, H. Cai, K. Chang, E. Cambria, Joint Node and Community Embedding on Graphs, in: CIKM, 2017.

R. Mihalcea, A. Garimella, What men say, what women hear: Finding gender-specific meaning shades, IEEE Intelligent Systems 31(4), pp. 62-67 (2016) 31 (4) (2016) 62–67.

L. C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multi-modal information, in: Proceedings of ICICS, vol. 1, IEEE, 397–401, 1997.

L. S. Chen, T. S. Huang, T. Miyasato, R. Nakatsu, Multimodal human emotion/expression recognition, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 366–371, 1998.

L. Kessous, G. Castellano, G. Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, Journal on Multimodal User Interfaces 3 (1-2) (2010) 33–48.

B. Schuller, Recognizing affect from linguistic information in 3D continuous space, IEEE Transactions on Affective Computing 2 (4) (2011) 192–205.

M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intelligent Systems 28 (3) (2013) 46–53.

V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Ensemble of SVM trees for multimodal emotion recognition, in: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, IEEE, 1–4, 2012.

A. Metallinou, S. Lee, S. Narayanan, Audio-visual emotion recognition using gaussian mixture models for face and voice, in: Tenth IEEE International Symposium on ISM 2008, IEEE, 250–257, 2008.

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues, Journal on Multimodal User Interfaces 3 (1-2) (2010a) 7–19.

C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Transactions on Affective Computing 2 (1) (2011) 10–21.

S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of EMNLP, 2539–2544, 2015.

S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017b) 98–125.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 1725–1732, 2014.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 .

V. Teh, G. E. Hinton, Rate-coded restricted Boltzmann machines for face recognition, in: T. Leen, T. Dietterich, V. Tresp (Eds.), Advances in neural information processing system, vol. 13, 908–914, 2001.

F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the Munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 1459–1462, 2010b.

S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2013) 221–231.

N. Majumder, Multimodal Sentiment Analysis in Social Media using Deep Learning with Convolutional Neural Networks, Master's thesis, CIC, Instituto Politécnico Nacional, 2017.

D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, CoRR abs/1412.6980, URL `http://arxiv.org/abs/1412.6980`.

A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intelligent Systems 31 (6) (2016) 82–88.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335–359.

V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-Level Multimodal Sentiment Analysis, in: ACL (1), 973–982, 2013.