Abstract

Multimodal sentiment analysis is an increasingly popular research area, which extends the conventional language-based definition of sentiment analysis to a multimodal setup where other relevant modalities accompany language. In this paper, we pose the problem of multimodal sentiment analysis as modeling intra-modality and inter-modality dynamics. We introduce a novel model, termed Tensor Fusion Network, which learns both such dynamics end-to-end. The proposed approach is tailored for the volatile nature of spoken language in online videos as well as accompanying gestures and voice. In the experiments, our model outperforms state-of-the-art approaches for both multimodal and unimodal sentiment analysis.

1 Introduction

Multimodal sentiment analysis (Morency et al., 2011; Zadeh et al., 2016b; Poria et al., 2015) is an increasingly popular area of affective computing research (Poria et al., 2017) that focuses on generalizing text-based sentiment analysis to opinionated videos, where three communicative modalities are present: language (spoken words), visual (gestures), and acoustic (voice).

This generalization is particularly vital to part of the NLP community dealing with opinion mining and sentiment analysis (Cambria et al., 2017) since there is a growing trend of sharing opinions in videos instead of text, specially in social media (Facebook, YouTube, etc.). The central challenge in multimodal sentiment analysis is to model the inter-modality dynamics: the interactions between language, visual and acoustic behaviors that change the perception of the expressed sentiment.

Figure 1 illustrates these complex inter-modality dynamics. The utterance “This movie is sick” can be ambiguous (either positive or negative) by itself, but if the speaker is also smiling at the same time, then it will be perceived as positive. On the other hand, the same utterance with a frown would be perceived negatively. A person speaking loudly “This movie is sick” would still be ambiguous. These examples are illustrating bimodal interactions. Examples of trimodal interactions are shown in Figure 1 when loud voice increases the sentiment to strongly positive. The complexity of inter-modality dynamics is shown in the second trimodal example where the utterance “This movie is fair” is still weakly positive, given the strong influence of the word “fair”.

A second challenge in multimodal sentiment analysis is efficiently exploring intra-modality dynamics of a specific modality (unimodal interaction). Intra-modality dynamics are particularly
challenging for the language analysis since multimodal sentiment analysis is performed on spoken language. A spoken opinion such as “I think it was alright . . . Hmmmm . . . let me think . . . yeah . . . no . . . ok yeah” almost never happens in written text. This volatile nature of spoken opinions, where proper language structure is often ignored, complicates sentiment analysis. Visual and acoustic modalities also contain their own intra-modality dynamics which are expressed through both space and time.

Previous works in multimodal sentiment analysis does not account for both intra-modality and inter-modality dynamics directly, instead they either perform early fusion (a.k.a., feature-level fusion) or late fusion (a.k.a., decision-level fusion). Early fusion consists in simply concatenating multimodal features mostly at input level (Morency et al., 2011; Pérez-Rosas et al., 2013; Poria et al., 2016). This fusion approach does not allow the intra-modality dynamics to be efficiently modeled. This is due to the fact that inter-modality dynamics can be more complex at input level and can dominate the learning process or result in overfitting. Late fusion, instead, consists in training unimodal classifiers independently and performing decision voting (Wang et al., 2016; Zadeh et al., 2016), and MOUD (Pérez-Rosas et al., 2013), however CMU-MOSI is the only English dataset with utterance-level sentiment labels. The newest multimodal sentiment analysis approaches have used deep neural networks, including convolutional neural networks (CNNs) with multiple-kernel learning (Poria et al., 2015), SAL-CNN (Wang et al., 2016) which learns generalizable features across speakers, and support vector machines (SVMs) with a multimodal dictionary (Zadeh, 2015).

In this paper, we introduce a new model, termed Tensor Fusion Network (TFN), which learns both the intra-modality and inter-modality dynamics end-to-end. Inter-modality dynamics are modeled with a new multimodal fusion approach, named Tensor Fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. Intra-modality dynamics are modeled through three Modality Embedding Subnetworks, for language, visual and acoustic modalities, respectively.

In our extensive set of experiments, we show (a) that TFN outperforms previous state-of-the-art approaches for multimodal sentiment analysis, (b) the characteristics and capabilities of our Tensor Fusion approach for multimodal sentiment analysis, and (c) that each of our three Modality Embedding Subnetworks (language, visual and acoustic) are also outperforming unimodal state-of-the-art unimodal sentiment analysis approaches.

2 Related Work

Sentiment Analysis is a well-studied research area in NLP (Pang et al., 2008). Various approaches have been proposed to model sentiment from language, including methods that focus on opinionated words (Hu and Liu, 2004; Taboada et al., 2011; Poria et al., 2014b; Cambria et al., 2016), n-grams and language models (Yang and Cardie, 2012), sentiment compositionality and dependency-based analysis (Socher et al., 2013; Poria et al., 2014a; Agarwal et al., 2015; Tai et al., 2015), and distributional representations for sentiment (Iyyer et al., 2015).

Multimodal Sentiment Analysis is an emerging research area that integrates verbal and nonverbal behaviors into the detection of user sentiment. There exist several multimodal datasets that include sentiment annotations, including the newly-introduced CMU-MOSI dataset (Zadeh et al., 2016b), as well as other datasets including ICT-MMMO (Wöllmer et al., 2013), YouTube (Morency et al., 2011), and MOUD (Pérez-Rosas et al., 2013), however CMU-MOSI is the only English dataset with utterance-level sentiment labels. The newest multimodal sentiment analysis approaches have used deep neural networks, including convolutional neural networks (CNNs) with multiple-kernel learning (Poria et al., 2015), SAL-CNN (Wang et al., 2016) which learns generalizable features across speakers, and support vector machines (SVMs) with a multimodal dictionary (Zadeh, 2015).

Audio-Visual Emotion Recognition is closely tied to multimodal sentiment analysis (Poria et al., 2017). Both audio and visual features have been shown to be useful in the recognition of emotions (Ghosh et al., 2016a). Using facial expressions and audio cues jointly has been the focus of many recent studies (Glok et al., 2011; Valstar et al., 2016; Nojavanasghari et al., 2016).

Multimodal Machine Learning has been a growing trend in machine learning research that is closely tied to the studies in this paper. Creative and novel applications of using multiple modalities have been among successful recent research directions in machine learning (You et al., 2016; Donahue et al., 2015; Antol et al., 2015; Specia et al., 2016; Tong et al., 2017).

3 CMU-MOSI Dataset

Multimodal Opinion Sentiment Intensity (CMU-MOSI) dataset is an annotated dataset of video
opinions from YouTube movie reviews (Zadeh et al., 2016a). Annotation of sentiment has closely followed the annotation scheme of the Stanford Sentiment Treebank (Socher et al., 2013), where sentiment is annotated on a seven-step Likert scale from very negative to very positive. However, whereas the Stanford Sentiment Treebank is segmented by sentence, the CMU-MOSI dataset is segmented by opinion utterances to accommodate spoken language where sentence boundaries are not as clear as text. There are 2199 opinion utterances for 93 distinct speakers in CMU-MOSI. There are an average 23.2 opinion segments in each video. Each video has an average length of 4.2 seconds. There are a total of 26,295 words in the opinion utterances. These utterance are annotated by five Mechanical Turk annotators for sentiment. The final agreement between the annotators is high in terms of Krippendorf’s alpha $\alpha = 0.77$. Figure 2 shows the distribution of sentiment across different opinions and different opinion sizes. CMU-MOSI dataset facilitates three prediction tasks, each of which we address in our experiments: 1) Binary Sentiment Classification 2) Five-Class Sentiment Classification (similar to Stanford Sentiment Treebank fine-grained classification with seven scale being mapped to five) and 3) Sentiment Regression in range $[-3, 3]$. For sentiment regression, we report Mean-Absolute Error (lower is better) and correlation (higher is better) between the model predictions and regression ground truth.

4 Tensor Fusion Network

Our proposed TFN consists of three major components: 1) Modality Embedding Subnetworks take as input unimodal features, and output a rich modality embedding. 2) Tensor Fusion Layer explicitly models the unimodal, bimodal and trimodal interactions using a 3-fold Cartesian product from modality embeddings. 3) Sentiment Inference Subnetwork is a network conditioned on the output of the Tensor Fusion Layer and performs sentiment inference. Depending on the task from Section 3 the network output changes to accommodate binary classification, 5-class classification or regression. Input to the TFN is an opinion utterance which includes three modalities of language, visual and acoustic. The following three subsections describe the TFN subnetworks and their inputs in detail.

4.1 Modality Embedding Subnetworks

Spoken Language Embedding Subnetwork:
Spoken text is different than written text (reviews, tweets) in compositionality and grammar. We revisit the spoken opinion: “I think it was alright … Hmmmm … let me think … yeah … no … ok yeah”. This form of opinion rarely happens in written language but variants of it are very common in spoken language. The first part conveys the actual message and the rest is speaker thinking out loud eventually agreeing with the first part. The key factor in dealing with this volatile nature of spoken language is to build models that are capable of operating in presence of unreliable and idiosyncratic speech traits by focusing on important parts of speech.

Our proposed approach to deal with challenges of spoken language is to learn a rich representation of spoken words at each word interval and use it as input to a fully connected deep network (Figure 3). This rich representation for $i$th word contains information from beginning of utterance through time, as well as $i$th word. This way as the model is discovering the meaning of the utterance through time, if it encounters unusable information in word $i + 1$ and arbitrary number of words after, the representation up until $i$ is not diluted or lost. Also, if the model encounters usable information again, it can recover by embedding those in the long short-term memory (LSTM). The time-dependent
If you like ..., see GloVe h1 h2 h3 hTl LSTM LSTM LSTM LSTM ...

... ...

... ...
zl 128 ReLU 128 ... of human voice and have been shown to be related to emotions (Ghosh et al., 2016b).

A LSTM network (Hochreiter and Schmidhuber, 1997) with a forget gate (Gers et al., 2000) is used to learn time-dependent language representations \( h_t \) for words according to the following LSTM formulation.

\[
\begin{align*}
    i_t &= \sigma(f_t \odot c_{t-1} + i_t \odot m) \\
    f_t &= \sigma(\odot \sigma m) \\
    o_t &= \sigma(\odot \sigma m) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot m \\
    h_t &= o_t \odot \sigma c_t \\
    h_t &= [h_1; h_2; h_3; \ldots; h_T] \\
    z_t &= U_t(1; W_l) \in \mathbb{R}^{128}
\end{align*}
\]

where \( W_l \) is the set of all weights in the \( U_l \) network (including \( W_{l_f}, W_{l_v}, W_{l_f} \), and \( b_{l_f} \)). \( \sigma \) is the sigmoid function.

Visual Embedding Subnetwork: Since opinion videos consist mostly of speakers talking to the audience through close-up camera, face is the most important source of visual information. The speaker’s face is detected for each frame (sampled at 30Hz) and indicators of the seven basic emotions (anger, contempt, disgust, fear, joy, sadness, and surprise) and two advanced emotions (frustration and confusion) (Ekman, 1992) are extracted using FACET facial expression analysis framework\(^1\). A set of 20 Facial Action Units (Ekman et al., 1980), indicating detailed muscle movements on the face, are also extracted using FACET. Estimates of head position, head rotation, and 68 facial landmark locations also extracted per frame using OpenFace (Baltrušaitis et al., 2016; Zadeh et al., 2017).

Let the visual features \( \hat{v}_j = [v_1^j, v_2^j, v_3^j, \ldots, v_p^j] \) for frame \( j \) of utterance video contain the set of \( p \) visual features, with \( T_v \) the number of total video frames in utterance. We perform mean pooling over the frames to obtain the expected visual features \( \mathbf{v} = [E[v_1], E[v_2], E[v_3], \ldots, E[v_p]] \). \( \mathbf{v} \) is then used as input to the Visual Embedding Subnetwork \( U_v \). Since information extracted using FACET from videos is rich, using a deep neural network would be sufficient to produce meaningful embeddings of visual modality. We use a deep neural network with three hidden layers of 32 ReLU units and weights \( W_v \). Empirically observed that making the model deeper or increasing the number of neurons in each layer does not lead to better visual performance. The subnetwork output provides the visual embedding \( z^v \):

\[
z^v = U_v(\mathbf{v}; W_v) \in \mathbb{R}^{32}
\]

Acoustic Embedding Subnetwork: For each opinion utterance audio, a set of acoustic features are extracted using COVAREP acoustic analysis framework (Degottex et al., 2014), including 12 MFCCs, pitch tracking and Voiced/UnVoiced segmenting features (using the additive noise robust Summation of Residual Harmonics (SRH) method (Drugman and Alwan, 2011)), glottal source parameters (estimated by glottal inverse filtering based on GCI synchronous IAIF (Drugman et al., 2012; Alku, 1992; Alku et al., 2002, 1997; Titze and Sundberg, 1992; Childers and Lee, 1991)), peak slope parameters (Degottex et al., 2014), maxima dispersion quotients (MDQ) (Kane and Gobl, 2013), and estimations of the \( R_d \) shape parameter of the Liljencrants-Fant (LF) glottal model (Fujisaki and Ljungqvist, 1986). These extracted features capture different characteristics of human voice and have been shown to be related to emotions (Ghosh et al., 2016b).

http://goo.gl/1rh1JN

\(^1\)
Unimodal Early Fusion Unimodal Tensor Fusion

Figure 4: Left: Commonly used early fusion (multimodal concatenation). Right: Our proposed tensor fusion with three types of subtensors: unimodal, bimodal and trimodal.

For each opinion segment with $T_a$ audio frames (sampled at 100Hz; i.e., 10ms), we extract the set of $q$ acoustic features $a_j = [a_j^1, a_j^2, a_j^3, \ldots, a_j^q]$ for audio frame $j$ in utterance. We perform mean pooling per utterance on these extracted acoustic features to obtain the expected acoustic features $a = [E[a_1], E[a_2], E[a_3], \ldots, E[q]]$. Here, $a$ is the input to the Audio Embedding Subnetwork $U_a$. Since COVAREP also extracts rich features from audio, using a deep neural network is sufficient to model the acoustic modality. Similar to $U_v$, $U_a$ is a network with 3 layers of 32 ReLU units with weights $W_a$.

Here, we also empirically observed that making the model deeper or increasing the number of neurons in each layer does not lead to better performance. The subnetwork produces the audio embedding $z^a$:

$$z^a = U_a(a; W_a) \in \mathbb{R}^{32}$$

4.2 Tensor Fusion Layer

While previous works in multimodal research have used feature concatenation as an approach for multimodal fusion, we aim to build a fusion layer in TFN that disentangles unimodal, bimodal and trimodal dynamics by modeling each of them explicitly. We call this layer Tensor Fusion, which is defined as the following vector field using three-fold Cartesian product:

$$\left\{ (z^l, z^v, z^a) \mid z^l \in \left[ \begin{array}{c} z^l \\ 1 \end{array} \right], z^v \in \left[ \begin{array}{c} z^v \\ 1 \end{array} \right], z^a \in \left[ \begin{array}{c} z^a \\ 1 \end{array} \right] \right\}$$

The extra constant dimension with value 1 generates the unimodal and bimodal dynamics. Each neural coordinate $(z^l, z^v, z^a)$ can be seen as a 3-D point in the 3-fold Cartesian space defined by the language, visual, and acoustic embeddings dimensions $[z^l]^T, [z^v]^T$, and $[z^a]^T$.

This definition is mathematically equivalent to a differentiable outer product between $z^l$, the visual representation $z^v$, and the acoustic representation $z^a$.

$$z^n = \left[ \begin{array}{c} z^l \\ 1 \end{array} \right] \otimes \left[ \begin{array}{c} z^v \\ 1 \end{array} \right] \otimes \left[ \begin{array}{c} z^a \\ 1 \end{array} \right]$$

Here $\otimes$ indicates the outer product between vectors and $z^n \in \mathbb{R}^{129 \times 33 \times 33}$ is the 3D cube of all possible combination of unimodal embeddings with seven semantically distinct subregions in Figure 4. The first three subregions $z^l$, $z^v$, and $z^a$ are unimodal embeddings from Modality Embedding Subnetworks forming unimodal interactions in Tensor Fusion. Three subregions $z^l \otimes z^v$, $z^l \otimes z^a$, and $z^v \otimes z^a$ capture bimodal interactions in Tensor Fusion. Finally, $z^l \otimes z^v \otimes z^a$ captures trimodal interactions.

Early fusion commonly used in multimodal research dealing with language, vision and audio, can be seen as a special case of Tensor Fusion with only unimodal interactions. Since Tensor Fusion is mathematically formed by an outer product, it has no learnable parameters and we empirically observed that although the output tensor is high dimensional, chances of overfitting are low.
We argue that this is due to the fact that the output neurons of Tensor Fusion are easy to interpret and semantically very meaningful (i.e., the manifold that they lie on is not complex but just high dimensional). Thus, it is easy for the subsequent layers of the network to decode the meaningful information.

4.3 Sentiment Inference Subnetwork

After Tensor Fusion layer, each opinion utterance can be represented as a multimodal tensor \( z^m \). We use a fully connected deep neural network called Sentiment Inference Subnetwork \( U_s \) with weights \( W_s \) conditioned on \( z^m \). The architecture of the network consists of two layers of 128 ReLU activation units connected to decision layer. The likelihood function of the Sentiment Inference Subnetwork is defined as follows, where \( \phi \) is the sentiment prediction:

\[
\arg \max_{\phi} p(\phi \mid z^m; W_s) = \arg \max_{\phi} U_s(z^m; W_s)
\]

In our experiments, we use three variations of the \( U_s \) network. The first network is trained for binary sentiment classification, with a single sigmoid output neuron using binary cross-entropy loss. The second network is designed for five-class sentiment classification, and uses a softmax probability function using categorical cross-entropy loss. The third network uses a single sigmoid output, using mean-squared error loss to perform sentiment regression.

5 Experiments

In this paper, we devise three sets of experiments each addressing a different research question:

**Experiment 1**: We compare our TFN with previously proposed multimodal sentiment analysis models. We compare to the following baselines:

- **C-MKL** (Poria et al., 2015) Convolutional MKL-based model is a multimodal sentiment classification model which uses a CNN to extract textual features and uses multiple kernel learning for sentiment analysis. It is current SOTA (state of the art) on CMU-MOSI.

- **SAL-CNN** (Wang et al., 2016) Select-Additive Learning is a multimodal sentiment analysis model that attempts to prevent identity-dependent information from being learned in a deep neural network. We retrain the model for 5-fold cross-validation using the code provided by the authors on github.

- **SVM-MD** (Zadeh et al., 2016b) is a SVM model trained on multimodal features using early fusion. The model used in (Morency et al., 2011) and (Pérez-Rosas et al., 2013) also similarly use SVM on multimodal concatenated features. We also present the results of Random Forest **RF-MD** to compare to another non-neural approach.

The results first experiment are reported in Table 1. TFN outperforms previously proposed neural and non-neural approaches. This difference is specifically visible in the case of 5-class classification.

**5.1 E1: Multimodal Sentiment Analysis**

In this section, we compare the performance of TFN model with previously proposed multimodal sentiment analysis models. We compare to the following baselines:

- **C-MKL** (Poria et al., 2015) Convolutional MKL-based model is a multimodal sentiment classification model which uses a CNN to extract textual features and uses multiple kernel learning for sentiment analysis. It is current SOTA (state of the art) on CMU-MOSI.

- **SAL-CNN** (Wang et al., 2016) Select-Additive Learning is a multimodal sentiment analysis model that attempts to prevent identity-dependent information from being learned in a deep neural network. We retrain the model for 5-fold cross-validation using the code provided by the authors on github.

- **SVM-MD** (Zadeh et al., 2016b) is a SVM model trained on multimodal features using early fusion. The model used in (Morency et al., 2011) and (Pérez-Rosas et al., 2013) also similarly use SVM on multimodal concatenated features. We also present the results of Random Forest **RF-MD** to compare to another non-neural approach.

The results first experiment are reported in Table 1. TFN outperforms previously proposed neural and non-neural approaches. This difference is specifically visible in the case of 5-class classification.

**5.2 E2: Tensor Fusion Evaluation**

Table 4 shows the results of our ablation study. The first three rows are showing the performance of each modality, when no intermodality dynamics are modeled. From this first experiment, we observe that the language modality is the most predictive.
<table>
<thead>
<tr>
<th>Baseline</th>
<th>Binary Acc(%)</th>
<th>5-class Acc(%)</th>
<th>Regression F1</th>
<th>Regression MAE</th>
<th>Regression r</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFNlanguage</td>
<td>74.8</td>
<td>75.6</td>
<td>38.5</td>
<td>0.99</td>
<td>0.61</td>
</tr>
<tr>
<td>TFNvisual</td>
<td>66.8</td>
<td>70.4</td>
<td>30.4</td>
<td>1.13</td>
<td>0.48</td>
</tr>
<tr>
<td>TFNacoustic</td>
<td>65.1</td>
<td>67.3</td>
<td>27.5</td>
<td>1.23</td>
<td>0.36</td>
</tr>
<tr>
<td>TFNunimodal</td>
<td>75.2</td>
<td>76.0</td>
<td>39.6</td>
<td>0.92</td>
<td>0.65</td>
</tr>
<tr>
<td>TFNbimodal</td>
<td>74.5</td>
<td>75.0</td>
<td>38.9</td>
<td>0.93</td>
<td>0.65</td>
</tr>
<tr>
<td>TFNtrimodal</td>
<td>76.3</td>
<td>76.2</td>
<td>39.7</td>
<td>0.919</td>
<td>0.66</td>
</tr>
<tr>
<td>TFN</td>
<td>77.1</td>
<td>77.9</td>
<td>42.0</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>TFNearly</td>
<td>75.2</td>
<td>76.2</td>
<td>39.0</td>
<td>0.96</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 2: Comparison of TFN with its subtensor variants. All the unimodal, bimodal and trimodal subtensors are important. TFN also outperforms early fusion.

As a second set of ablation experiments, we test our TFN approach when only the bimodal subtensors are used (TFN_{bimodal}) or when only the trimodal subtensor is used (TFN_{trimodal}). We observe that bimodal subtensors are more informative when used without other subtensors. The most interesting comparison is between our full TFN model and a variant (TFN_{notrimodal}) where the trimodal subtensor is removed (but all the unimodal and bimodal subtensors are present). We observe a big improvement for the full TFN model, confirming the importance of the trimodal dynamics and the need for all components of the full tensor.

We also perform a comparison with the early fusion approach (TFN_{early}) by simply concatenating all three modality embeddings \(<z^l, z^a, z^n>\) and passing it directly as input to \(U_n\). This approach was depicted on the left side of Figure 4. When looking at Table 4 results, we see that our TFN approach outperforms the early fusion approach².

5.3 E3: Modality Embedding Subnetworks Evaluation

In this experiment, we compare the performance of our Modality Embedding Networks with state-of-the-art approaches for language-based, visual-based and acoustic-based sentiment analysis.

5.3.1 Language Sentiment Analysis

We selected the following state-of-the-art approaches to include variety in their techniques,

²We also performed other comparisons with variants of the early fusion model TFN_{early} where we increased the number of parameters and neurons to replicate the numbers from our TFN model. In all cases, the performances were similar to TFN_{early} (and lower than our TFN model). Because of space constraints, we could not include them in this paper.

Table 3: Language Sentiment Analysis. Comparison of with state-of-the-art approaches for language sentiment analysis. \(\Delta_{SOTA}^{\text{language}}\) shows improvement.

<table>
<thead>
<tr>
<th>Language Baseline</th>
<th>Binary Acc(%)</th>
<th>5-class Acc(%)</th>
<th>Regression F1</th>
<th>Regression MAE</th>
<th>Regression r</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNTN</td>
<td>(73.7)</td>
<td>(73.4)</td>
<td>(35.2)</td>
<td>(0.99)</td>
<td>(0.59)</td>
</tr>
<tr>
<td>DAN</td>
<td>73.4</td>
<td>73.8</td>
<td>39.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D-CNN</td>
<td>(68.8)</td>
<td>(68.4)</td>
<td>(36.7)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CMKL-L</td>
<td>71.2</td>
<td>72.4</td>
<td>34.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SAL-CNN-L</td>
<td>73.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM-MD-L</td>
<td>70.6</td>
<td>71.2</td>
<td>33.1</td>
<td>1.18</td>
<td>0.46</td>
</tr>
<tr>
<td>TFN_{language}</td>
<td>74.8</td>
<td>75.6</td>
<td>38.5</td>
<td>0.98</td>
<td>0.62</td>
</tr>
</tbody>
</table>

\(\Delta_{SOTA}^{\text{language}}\) ↑ 1.1 ↑ 1.8 ↓ 0.7 ↓ 0.01 ↑ 0.03

Based on dependency parsing (RNTN), distributional representation of text (DAN), and convolutional approaches (DynamicCNN). When possible, we retrain them on the CMU-MOSI dataset (performances of the original pre-trained models are shown in parenthesis in Table 3) and compare them to our language only TFN_{language}.

RNTN (Socher et al., 2013) The Recursive Neural Tensor Network is among the most well-known sentiment analysis methods proposed for both binary and multi-class sentiment analysis that uses dependency structure.

DAN (Iyyer et al., 2015) The Deep Average Network approach is a simple but efficient sentiment analysis model that uses information only from distributional representation of the words and not from the compositionality of the sentences.

DynamicCNN (Kalchbrenner et al., 2014) DynamicCNN is among the state-of-the-art models in text-based sentiment analysis which uses a convolutional architecture adopted for the semantic modeling of sentences.

CMKL-L, SAL-CNN-L and SVM-MD-L are multimodal models from section using only language modality 5.1.

Results in Table 3 show that our model using only language modality outperforms state-of-the-art approaches for the CMU-MOSI dataset. While previous models are well-studied and suitable models for sentiment analysis in written language, they underperform in modeling the sentiment in spoken language. We suspect that this underperformance is due to: RNTN and similar approaches rely heavily on dependency structure, which may not be present.
**Table 4: Visual Sentiment Analysis.** Comparison with state-of-the-art approaches for visual sentiment analysis and emotion recognition. $\Delta_{\text{visual}}$ shows the improvement.

<table>
<thead>
<tr>
<th>Visual Baseline</th>
<th>Binary Acc(%)</th>
<th>5-class Acc(%)</th>
<th>Reg F1</th>
<th>MAE</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-CNN</td>
<td>56.1</td>
<td>24.9</td>
<td>1.31</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>60.7</td>
<td>25.1</td>
<td>1.27</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>LSTM-FA</td>
<td>62.1</td>
<td>26.2</td>
<td>1.23</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>CMKL-V</td>
<td>52.6</td>
<td>29.3</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SAL-CNN-V</td>
<td>63.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SVM-MD-V</td>
<td>59.2</td>
<td>25.6</td>
<td>1.24</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td><strong>TFN\text{visual}</strong></td>
<td><strong>69.4</strong></td>
<td><strong>31.0</strong></td>
<td><strong>1.12</strong></td>
<td><strong>0.50</strong></td>
<td></td>
</tr>
</tbody>
</table>

$\Delta_{\text{SOTA}}$ visual $\uparrow 5.6 \uparrow 7.7 \uparrow 1.7 \uparrow 0.11 \uparrow 0.14$

**Table 5: Acoustic Sentiment Analysis.** Comparison with state-of-the-art approaches for audio sentiment analysis and emotion recognition. $\Delta_{\text{acoustic}}$ shows improvement.

<table>
<thead>
<tr>
<th>Acoustic Baseline</th>
<th>Binary Acc(%)</th>
<th>5-class Acc(%)</th>
<th>Reg F1</th>
<th>MAE</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL-RNN</td>
<td>63.4</td>
<td>25.9</td>
<td><strong>1.21</strong></td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Adieu-Net</td>
<td>59.2</td>
<td>25.1</td>
<td>1.29</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>SER-LSTM</td>
<td>55.4</td>
<td>24.2</td>
<td>1.36</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>CMKL-A</td>
<td>52.6</td>
<td>29.1</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SAL-CNN-A</td>
<td>62.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SVM-MD-A</td>
<td>56.3</td>
<td>24.6</td>
<td>1.29</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td><strong>TFN\text{acoustic}</strong></td>
<td><strong>65.1</strong></td>
<td><strong>27.5</strong></td>
<td><strong>1.23</strong></td>
<td><strong>0.36</strong></td>
<td></td>
</tr>
</tbody>
</table>

$\Delta_{\text{SOTA}}$ acoustic $\uparrow 1.7 \uparrow 3.1 \downarrow 1.6 \uparrow 0.02 \uparrow 0.02$

5.3.2 Visual Sentiment Analysis

We compare the performance of our models using visual information (TFN\text{visual}) with the following well-known approaches in visual sentiment analysis and emotion recognition (retrained for sentiment analysis):

**3DCNN** (Byeon and Kwak, 2014) a network using 3D CNN is trained using the face of the speaker. Face of the speaker is extracted in every 6 frames and resized to $64 \times 64$ and used as the input to the proposed network.

**CNN-LSTM** (Ebrahimi Kahou et al., 2015) is a recurrent model that at each timestamp performs convolutions over facial region and uses output to an LSTM. Face processing is similar to 3DCNN.

**LSTM-FA** similar to both baselines above, information extracted by FACET is used every 6 frames as input to an LSTM with a memory dimension of 100 neurons.

**SAL-CNN-V, SVM-MD-V, CMKL-V, RF-V** use only visual modality in multimodal baselines from Section 5.1.

The results in Table 5 show that $U_v$ is able to outperform state-of-the-art approaches on visual sentiment analysis.

5.3.3 Acoustic Sentiment Analysis

We compare the performance of our models using visual information (TFN\text{acoustic}) with the following well-known approaches in audio sentiment analysis and emotion recognition (retrained for sentiment analysis):

**HL-RNN** (Lee and Tashev, 2015) uses an LSTM on high-level audio features. We use the same features extracted for $U_a$, averaged over time slices of every 200 intervals.

**Adieu-Net** (Trigeorgis et al., 2016) is an end-to-end approach for emotion recognition in audio using directly PCM features.

**SER-LSTM** (Lim et al., 2016) is a model that uses recurrent neural networks on top of convolution operations on spectrogram of audio.

**SAL-CNN-A, SVM-MD-A, CMKL-A, RF-A** use only acoustic modality in multimodal baselines from Section 5.1.

5.4 Methodology

All the models in this paper are tested using five-fold cross-validation proposed by CMU-MOSI (Zadeh et al., 2016a). All of our experiments are performed independent of speaker identity, as no speaker is shared between train and test sets for generalizability of the model to unseen speakers in real-world. The best hyperparameters are chosen using grid search based on model performance on a validation set (using last 4 videos in train fold). The TFN model is trained using the Adam optimizer (Kingma and Ba, 2014) with the learning rate $5e4$. $U_v$ and $U_s$, $U_a$ subnetworks are regularized using dropout on all hidden layers with $p = 0.15$ and L2 norm coefficient 0.01. The train, test and validation folds are exactly the same for all baselines.

6 Qualitative Analysis

We analyze the impact of our proposed TFN multimodal fusion approach by comparing it with the
Table 6: Examples from the CMU-MOSI dataset. The ground truth sentiment labels are between strongly negative (-3) and strongly positive (+3). For each example, we show the prediction output of the three unimodal models (TFNaoustic, TFNvisual and TFNlanguage), the early fusion model TFNearly and our proposed TFN approach. TFNearly seems to be mostly replicating language modality while our TFN approach successfully integrate intermodality dynamics to predict the sentiment level.

<table>
<thead>
<tr>
<th>#</th>
<th>Spoken words + acoustic and visual behaviors</th>
<th>TFN-Acoustic</th>
<th>TFN-Visual</th>
<th>TFN-Language</th>
<th>TFN-Early</th>
<th>TFN</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“You can’t even tell funny jokes” + frowning expression</td>
<td>-0.375</td>
<td>-1.760</td>
<td>-0.558</td>
<td>-0.839</td>
<td>-1.661</td>
<td>-1.800</td>
</tr>
<tr>
<td>2</td>
<td>“I gave it a B” + smile expression + excited voice</td>
<td>1.967</td>
<td>1.245</td>
<td>0.438</td>
<td>0.467</td>
<td>1.215</td>
<td>1.400</td>
</tr>
<tr>
<td>3</td>
<td>“But I must say those are some pretty big shoes to fill so I thought maybe it has a chance” + headshake</td>
<td>-0.378</td>
<td>-1.034</td>
<td>1.734</td>
<td>1.385</td>
<td>0.608</td>
<td>0.400</td>
</tr>
<tr>
<td>4</td>
<td>“The only actor who can really sell their lines is Erin Eckart” + frown + low-energy voice</td>
<td>-0.970</td>
<td>-0.716</td>
<td>0.175</td>
<td>-0.031</td>
<td>-0.825</td>
<td>-1.000</td>
</tr>
</tbody>
</table>

early fusion approach TFNearly and the three unimodal models. Table 6 shows examples taken from the CMU-MOSI dataset. Each example is described with the spoken words as well as the acoustic and visual behaviors. The sentiment predictions and the ground truth labels range between strongly negative (-3) and strongly positive (+3).

As a first general observation, we observe that the early fusion model TFNearly shows a strong preference for the language modality and seems to be neglecting the intermodality dynamics. We can see this trend by comparing it with the language unimodal model TFNlanguage. In comparison, our TFN approach seems to capture more complex interaction through bimodal and trimodal dynamics and thus performs better. Specifically, in the first example, the utterance is weakly negative where the speaker is referring to lack of funny jokes in the movie. This example contains a bimodal interaction where the visual modality shows a negative expression (frowning) which is correctly captured by our TFN approach.

In the second example, the spoken words are ambiguous since the model has no clue what a B is except a token, but the acoustic and visual modalities are bringing complementary evidences. Our TFN approach correctly identify this trimodal interaction and predicts a positive sentiment. The third example is interesting since it shows an interaction where language predicts a positive sentiment but the strong negative visual behaviors bring the final prediction of our TFN approach almost to a neutral sentiment. The fourth example shows how the acoustic modality is also influencing our TFN predictions.

7 Conclusion

We introduced a new end-to-end fusion method for sentiment analysis which explicitly represents unimodal, bimodal, and trimodal interactions between behaviors. Our experiments on the publicly-available CMU-MOSI dataset produced state-of-the-art performance when compared against both multimodal approaches. Furthermore, our approach brings state-of-the-art results for language-only, visual-only and acoustic-only multimodal sentiment analysis on CMU-MOSI.

Acknowledgments

This project was partially supported by Oculus research grant. We would like to thank the reviewers for their valuable feedback.

References


Michel Valstar, Jonathan Gratch, BjörnSchuller, Fabien Ringeval, DennisLalanne, MercedesTorres Torres, StefanScherer, GiotaStratou, Roddy Cowie, and MajaPantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop


