Tensor Fusion Network for Multimodal Sentiment Analysis

Amir Zadeh[†], **Minghai Chen**[†]
Language Technologies Institute
Carnegie Mellon University
{abagherz,minghail}@cs.cmu.edu

Soujanya Poria
Temasek Laboratories,
NTU, Singapore
sporia@ntu.edu.sg

Erik Cambria
School of Computer Science and
Engineering, NTU, Singapore
cambria@ntu.edu.sg

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
morency@cs.cmu.edu

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.

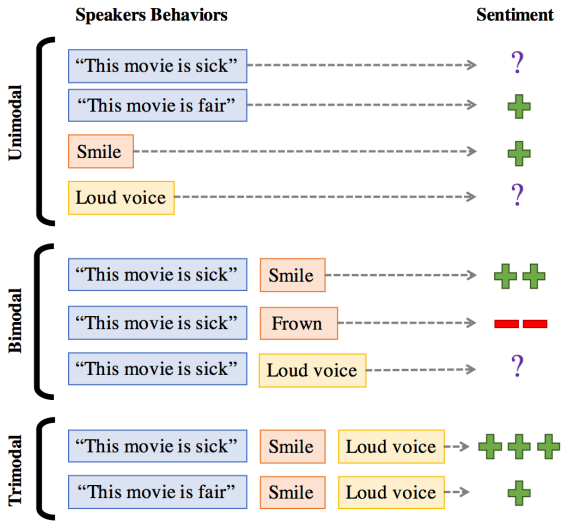

Figure 1: Unimodal, bimodal and trimodal interaction in multimodal sentiment analysis.

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

[†] means equal contribution

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1103–1114
Copenhagen, Denmark, September 7–11, 2017.

challenging for the language analysis since multimodal sentiment analysis is performed on spoken language. A spoken opinion such as "I think it was alright ... Hmmm ... 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., 2016a). This prevents the model from learning inter-modality dynamics in an efficient way by assuming that simple weighted averaging is a proper fusion approach.

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 (Glodek 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

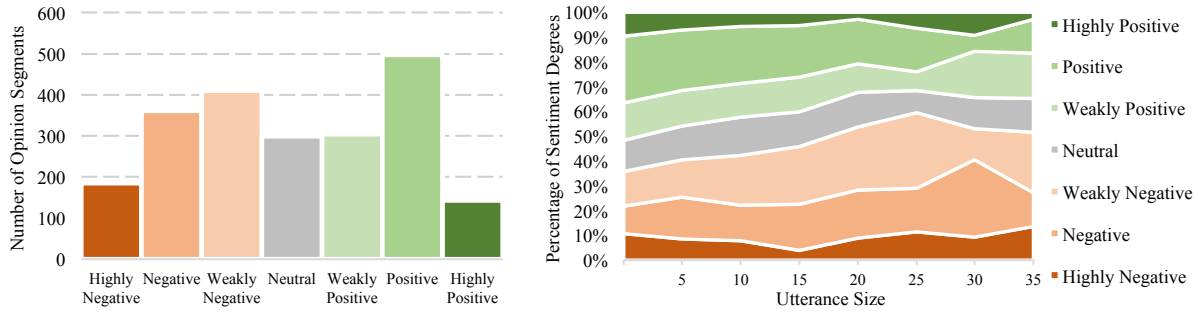

Figure 2: Distribution of sentiment across different opinions (left) and opinion sizes (right) in CMU-MOSI.

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 ... Hmmm ... 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 ith word contains information from beginning of utterance through time, as well as ith 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

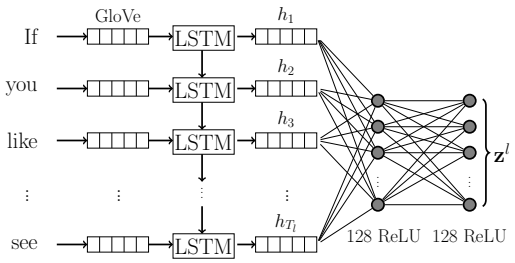

Figure 3: Spoken Language Embedding Subnetwork ($\mathcal{U}_l$)

encodings are usable by the rest of the pipeline by simply focusing on relevant parts using the non-linear affine transformation of time-dependent embeddings which can act as a dimension reducing attention mechanism. To formally define our proposed Spoken Language Embedding Subnetwork ($\mathcal{U}_l$), let $\mathbf{l} = \{l_1, l_2, l_3, \ldots, l_{T_l}; l_t \in \mathbb{R}^{300}\}$, where T_l is the number of words in an utterance, be the set of spoken words represented as a sequence of 300-dimensional GloVe word vectors (Pennington et al., 2014).

A LSTM network (Hochreiter and Schmidhuber, 1997) with a forget gate (Gers et al., 2000) is used to learn time-dependent language representations $\mathbf{h_l} = \{h_1, h_2, h_3, \ldots, h_{T_l}; h_t \in \mathbb{R}^{128}\}$ for words according to the following LSTM formulation.

$$\begin{pmatrix} i \\ f \\ o \\ m \end{pmatrix} = \begin{pmatrix} sigmoid \\ sigmoid \\ sigmoid \\ tanh \end{pmatrix} W_{l_d} \begin{pmatrix} X_t W_{l_e} \\ h_{t-1} \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot m$$

$$h_t = o \otimes tanh(c_t)$$

$$\mathbf{h_l} = [h_1; h_2; h_3; \ldots; h_{T_l}]$$

$\mathbf{h_l}$ is a matrix of language representations formed from concatenation of $h_1, h_2, h_3, \ldots h_{T_l}$. $\mathbf{h_l}$ is then used as input to a fully-connected network that generates language embedding $\mathbf{z}^l$:

$$\mathbf{z}^l = \mathcal{U}_l(\mathbf{l};\ W_l) \in \mathbb{R}^{128}$$

where W_l is the set of all weights in the $\mathcal{U}_l$ network (including W_{l_d}, W_{l_e},$W_{l_{fc}}$, and $b_{l_{fc}}$), σ 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 $\hat{\mathbf{v}}_j = [v_j^1, v_j^2, v_j^3, \ldots, v_j^p]$ 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 $\mathbf{v} = [\mathbb{E}[v^1], \mathbb{E}[v^2], \mathbb{E}[v^3], \ldots, \mathbb{E}[v^l]]$. $\mathbf{v}$ is then used as input to the Visual Embedding Subnetwork $\mathcal{U}_v$. 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 we 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 $\mathbf{z}^v$:

$$\mathbf{z}^v = \mathcal{U}_v(\mathbf{v};\ W_v) \in \mathbb{R}^{32}$$

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).

[1] http://goo.gl/1rh1JN

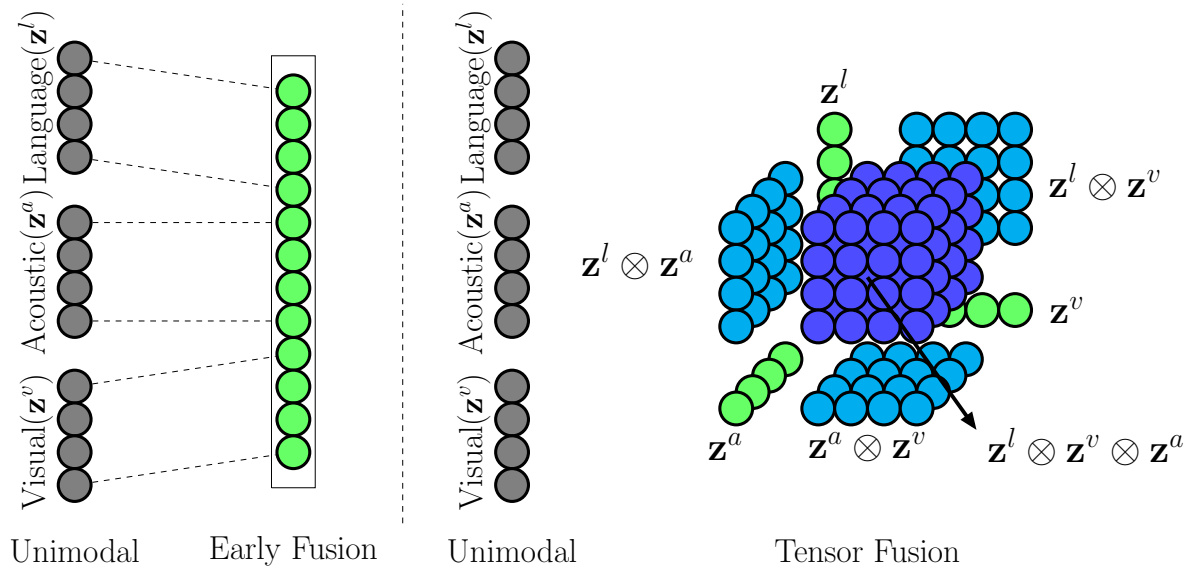

Figure 4: Left: Commonly used early fusion (multimodal concatenation). Right: Our proposed tensor fusion with three types of subtensors: unimodal, bimodal and trimodal.

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

$$\mathbf{z}^a = \mathcal{U}_a(\mathbf{a}; W_a) \in \mathbb{R}^{32}$$

4.2 Tensor Fusion Layer

While previous works in multimodal research has 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:

$$\left\{(z^l, z^v, z^a) \mid z^l \in \begin{bmatrix}\mathbf{z}^l \\ 1\end{bmatrix}, z^v \in \begin{bmatrix}\mathbf{z}^v \\ 1\end{bmatrix}, z^a \in \begin{bmatrix}\mathbf{z}^a \\ 1\end{bmatrix}\right\}$$

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 $[\mathbf{z}^l 1]^T$, $[\mathbf{z}^v 1]^T$, and $[\mathbf{z}^a 1]^T$.

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

$$\mathbf{z}^m = \begin{bmatrix}\mathbf{z}^l \\ 1\end{bmatrix} \otimes \begin{bmatrix}\mathbf{z}^v \\ 1\end{bmatrix} \otimes \begin{bmatrix}\mathbf{z}^a \\ 1\end{bmatrix}$$

Here $\otimes$ indicates the outer product between vectors and $\mathbf{z}^m \in \mathbb{R}^{129\times33\times33}$ is the 3D cube of all possible combination of unimodal embeddings with seven semantically distinct subregions in Figure 4. The first three subregions $\mathbf{z}^l$, $\mathbf{z}^v$, and $\mathbf{z}^a$ are unimodal embeddings from Modality Embedding Subnetworks forming unimodal interactions in Tensor Fusion. Three subregions $\mathbf{z}^l \otimes \mathbf{z}^v$, $\mathbf{z}^l \otimes \mathbf{z}^a$, and $\mathbf{z}^v \otimes \mathbf{z}^a$ capture bimodal interactions in Tensor Fusion. Finally, $\mathbf{z}^l \otimes \mathbf{z}^v \otimes \mathbf{z}^a$ 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 $\mathbf{z}^m$. We use a fully connected deep neural network called Sentiment Inference Subnetwork $\mathcal{U}_s$ with weights W_s conditioned on $\mathbf{z}^m$. 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 ϕ is the sentiment prediction:

$$\arg\max_{\phi} \ \mathrm{p}(\phi \mid \mathbf{z}^m; W_s) = \arg\max_{\phi} \ \mathcal{U}_s(\mathbf{z}^m; W_s)$$

In our experiments, we use three variations of the $\mathcal{U}_s$ 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-squarred 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 previous state-of-the-art approaches in multimodal sentiment analysis.

Experiment 2: We study the importance of the TFN subtensors and the impact of each individual modality (see Figure 4). We also compare with the commonly-used early fusion approach.

Experiment 3: We compare the performance of our three modality-specific networks (language, visual and acoustic) with state-of-the-art unimodal approaches.

Section 5.4 describes our experimental methodology which is kept constant across all experiments. Section 6 will discuss our results in more details with a qualitative analysis.

Multimodal Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
Random	50.2	48.7	23.9	1.88	-
C-MKL	73.1	75.2	35.3	-	-
SAL-CNN	73.0	-	-	-	-
SVM-MD	71.6	72.3	32.0	1.10	0.53
RF	71.4	72.1	31.9	1.11	0.51
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
Human	85.7	87.5	53.9	0.71	0.82
Δ^{SOTA}	↑ 4.0	↑ 2.7	↑ 6.7	↓ 0.23	↑ 0.17

Table 1: Comparison with state-of-the-art approaches for multimodal sentiment analysis. TFN outperforms both neural and non-neural approaches as shown by Δ^{SOTA}.

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.

Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\text{TFN}_{language}$	74.8	75.6	38.5	0.99	0.61
TFN_{visual}	66.8	70.4	30.4	1.13	0.48
$\text{TFN}_{acoustic}$	65.1	67.3	27.5	1.23	0.36
$\text{TFN}_{bimodal}$	75.2	76.0	39.6	0.92	0.65
$\text{TFN}_{trimodal}$	74.5	75.0	38.9	0.93	0.65
$\text{TFN}_{notrimodal}$	75.3	76.2	39.7	0.919	0.66
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
TFN_{early}	75.2	76.2	39.0	0.96	0.63

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 ($\text{TFN}_{bimodal}$) or when only the trimodal subtensor is used ($\text{TFN}_{bimodal}$). 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 ($\text{TFN}_{notrimodal}$) 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 $< z^l, z^a, z^v >$ and passing it directly as input to $\mathcal{U}_s$. 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 approach[2].

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, 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 $\text{TFN}_{language}$.

Language Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
RNTN	- (73.7)	- (73.4)	- (35.2)	- (0.99)	- (0.59)
DAN	73.4 (68.8)	73.8 (68.4)	**39.2** (36.7)	- -	- -
D-CNN	65.5 (62.1)	66.9 (56.4)	32.0 (32.4)	- -	- -
CMKL-L	71.2	72.4	34.5	-	-
SAL-CNN-L	73.5	-	-	-	-
SVM-MD-L	70.6	71.2	33.1	1.18	0.46
$\text{TFN}_{language}$	**74.8**	**75.6**	38.5	**0.98**	**0.62**
$\Delta^{SOTA}_{language}$	↑ 1.1	↑ 1.8	↓ 0.7	↓ 0.01	↑ 0.03

Table 3: Language Sentiment Analysis. Comparison of with state-of-the-art approaches for language sentiment analysis. $\Delta^{SOTA}_{language}$ shows improvement.

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.

CMK-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

[2]We 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.

Visual Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
3D-CNN	56.1	58.4	24.9	1.31	0.26
CNN-LSTM	60.7	61.2	25.1	1.27	0.30
LSTM-FA	62.1	63.7	26.2	1.23	0.33
CMKL-V	52.6	58.5	29.3	-	-
SAL-CNN-V	63.8	-	-	-	-
SVM-MD-V	59.2	60.1	25.6	1.24	0.36
TFN$_{visual}$	**69.4**	**71.4**	**31.0**	**1.12**	**0.50**
Δ_{visual}^{SOTA}	↑ 5.6	↑ 7.7	↑ 1.7	↓ 0.11	↑ 0.14

Table 4: Visual Sentiment Analysis. Comparison with state-of-the-art approaches for visual sentiment analysis and emotion recognition. Δ_{visual}^{SOTA} shows the improvement.

Acoustic Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
HL-RNN	63.4	64.2	25.9	**1.21**	0.34
Adieu-Net	59.2	60.6	25.1	1.29	0.31
SER-LSTM	55.4	56.1	24.2	1.36	0.23
CMKL-A	52.6	58.5	**29.1**	-	-
SAL-CNN-A	62.1	-	-	-	-
SVM-MD-A	56.3	58.0	24.6	1.29	0.28
TFN$_{acoustic}$	**65.1**	**67.3**	27.5	1.23	**0.36**
$\Delta_{acoustic}^{SOTA}$	↑ 1.7	↑ 3.1	↓ 1.6	↑ 0.02	↑ 0.02

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

in spoken language; DAN and similar sentence embeddings approaches can easily be diluted by words that may not relate directly to sentiment or meaning; D-CNN and similar convolutional approaches rely on spatial proximity of related words, which may not always be present in spoken language.

5.3.2 Visual Sentiment Analysis

We compare the performance of our models using visual information (TFN$_{visual}$) 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 64×64 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 $\mathcal{U}_v$ 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$_{acoustic}$) 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 $\mathcal{U}_a$ 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. $\mathcal{U}_v$ and $\mathcal{U}_a, \mathcal{U}_s$ 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

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

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 ($\text{TFN}_{acoustic}$, TFN_{visual} and $\text{TFN}_{language}$), the early fusion model TFN_{early} and our proposed TFN approach. TFN_{early} seems to be mostly replicating language modality while our TFN approach successfully integrate intermodality dynamics to predict the sentiment level.

early fusion approach TFN_{early} 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 TFN_{early} 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 $\text{TFN}_{language}$. 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

Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. 2015. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation* 7(4):487–499.

Paavo Alku. 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication* 11(2-3):109–118.

Paavo Alku, Tom Bäckström, and Erkki Vilkman. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America* 112(2):701–710.

Paavo Alku, Helmer Strik, and Erkki Vilkman. 1997. Parabolic spectral parameter—a new method for quantification of the glottal flow. *Speech Communication* 22(1):67–79.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2425–2433.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, pages 1–10.

Young-Hyen Byeon and Keun-Chang Kwak. 2014. Facial expression recognition using 3d convolutional neural network. *International Journal of Advanced Computer Science and Applications* 5(12).

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. *A Practical Guide to Sentiment Analysis*. Springer, Cham, Switzerland.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*. pages 2666–2677.

Donald G Childers and CK Lee. 1991. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America* 90(5):2394–2410.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 960–964.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 2625–2634.

Thomas Drugman and Abeer Alwan. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*. pages 1973–1976.

Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing* 20(3):994–1006.

Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 467–474.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6(3-4):169–200.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology* 39(6):1125–1134.

Hiroya Fujisaki and Mats Ljungqvist. 1986. Proposal and evaluation of models for the glottal source waveform. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*. IEEE, volume 11, pages 1605–1608.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation* 12(10):2451–2471.

Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016a. Representation learning for speech emotion recognition. In *Interspeech 2016*. pages 3603–3607. https://doi.org/10.21437/Interspeech.2016-692.

Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016b. Representation learning for speech emotion recognition. *Interspeech 2016* pages 3603–3607.

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, Springer, pages 359–368.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 168–177.

Mohit Iyyer, Varun Manjunatha, Jordan L Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *ACL (1)*. pages 1681–1691.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences pages 656–666.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing* 21(6):1170–1179.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Jinkyu Lee and Ivan Tashev. 2015. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH*. pages 1537–1540.

Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and recurrent neural networks. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, pages 1–4.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, pages 169–176.

Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, pages 137–144.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *ACL (1)*. pages 973–982.

Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. 2014a. Dependency-based semantic parsing for concept-level text analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 113–127.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37:98–125.

Soujanya Poria, Erik Cambria, and Alexander F Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*. pages 2539–2544.

Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. 2014b. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69:45–63.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, pages 439–448.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Citeseer, pages 1631–1642.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks pages 1556–1566.

Ingo R Titze and Johan Sundberg. 1992. Vocal intensity in speakers and singers. *the Journal of the Acoustical Society of America* 91(5):2936–2946.

Edmund Tong, Amir Zadeh, and Louis-Philippe Morency. 2017. Combating human trafficking with deep multimodal models. In *Association for Computational Linguistics*.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pages 5200–5204.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop

and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 3–10.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244* .

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28(3):46–53.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1335–1345.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4651–4659.

Amir Zadeh. 2015. Micro-opinion sentiment intensity analysis and summarization in online videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 587–591.

Amir Zadeh, Tadas Baltrušaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for facial landmark detection. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*. IEEE.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* .

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.