Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph

Amir Zadeh[1], **Paul Pu Liang**[2], **Jonathan Vanbriesen**[1], **Soujanya Poria**[3], **Edmund Tong**[1], **Erik Cambria**[4], **Minghai Chen**[1], **Louis-Philippe Morency**[1]
{1- Language Technologies Institute, 2- Machine Learning Department}, CMU, USA
{3- A*STAR, 4- Nanyang Technological University}, Singapore
{abagherz,pliang,jvanbrie}@cs.cmu.edu, soujanya@sentic.net
edtong@cmu.edu, cambria@ntu.edu.sg, morency@cs.cmu.edu

Abstract

Analyzing human multimodal language is an emerging area of research in NLP. Intrinsically human communication is multimodal (heterogeneous), temporal and asynchronous; it consists of the language (words), visual (expressions), and acoustic (paralinguistic) modalities all in the form of asynchronous coordinated sequences. From a resource perspective, there is a genuine need for large scale datasets that allow for in-depth studies of multimodal language. In this paper we introduce CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), the largest dataset of sentiment analysis and emotion recognition to date. Using data from CMU-MOSEI and a novel multimodal fusion technique called the Dynamic Fusion Graph (DFG), we conduct experimentation to investigate how modalities interact with each other in human multimodal language. Unlike previously proposed fusion techniques, DFG is highly interpretable and achieves competitive performance compared to the current state of the art.

1 Introduction

Theories of language origin identify the combination of language and nonverbal behaviors (vision and acoustic modality) as the prime form of communication utilized by humans throughout evolution (Müller, 1866). In natural language processing, this form of language is regarded as human multimodal language. Modeling multimodal language has recently become a centric research direction in both NLP and multimodal machine learning (Hazarika et al., 2018; Zadeh et al., 2018a; Poria et al., 2017a; Baltrušaitis et al., 2017; Chen et al., 2017). Studies strive to model the dual dynamics of multimodal language: intra-modal dynamics (dynamics within each modality) and cross-modal dynamics (dynamics across different modalities). However, from a resource perspective, previous multimodal language datasets have severe shortcomings in the following aspects:

Diversity in the training samples: The diversity in training samples is crucial for comprehensive multimodal language studies due to the complexity of the underlying distribution. This complexity is rooted in variability of intra-modal and cross-modal dynamics for language, vision and acoustic modalities (Rajagopalan et al., 2016). Previously proposed datasets for multimodal language are generally small in size due to difficulties associated with data acquisition and costs of annotations.

Variety in the topics: Variety in topics opens the door to generalizable studies across different domains. Models trained on only few topics generalize poorly as language and nonverbal behaviors tend to change based on the impression of the topic on speakers' internal mental state.

Diversity of speakers: Much like writing styles, speaking styles are highly idiosyncratic. Training models on only few speakers can lead to degenerate solutions where models learn the identity of speakers as opposed to a generalizable model of multimodal language (Wang et al., 2016).

Variety in annotations Having multiple labels to predict allows for studying the relations between labels. Another positive aspect of having variety of labels is allowing for multi-task learning which has shown excellent performance in past research.

Our first contribution in this paper is to introduce the largest dataset of multimodal sentiment and emotion recognition called CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). CMU-MOSEI contains 23,453 annotated video segments from 1,000 distinct speakers and

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2236–2246
Melbourne, Australia, July 15 - 20, 2018.

250 topics. Each video segment contains manual transcription aligned with audio to phoneme level. All the videos are gathered from online video sharing websites [1]. The dataset is currently a part of the CMU Multimodal Data SDK and is freely available to the scientific community through Github [2].

Our second contribution is an interpretable fusion model called Dynamic Fusion Graph (DFG) to study the nature of cross-modal dynamics in multimodal language. DFG contains built-in efficacies that are directly related to how modalities interact. These efficacies are visualized and studied in detail in our experiments. Aside interpretability, DFG achieves superior performance compared to previously proposed models for multimodal sentiment and emotion recognition on CMU-MOSEI.

2 Background

In this section we compare the CMU-MOSEI dataset to previously proposed datasets for modeling multimodal language. We then describe the baselines and recent models for sentiment analysis and emotion recognition.

2.1 Comparison to other Datasets

We compare CMU-MOSEI to an extensive pool of datasets for sentiment analysis and emotion recognition. The following datasets include a combination of language, visual and acoustic modalities as their input data.

2.1.1 Multimodal Datasets

CMU-MOSI (Zadeh et al., 2016b) is a collection of 2199 opinion video clips each annotated with sentiment in the range [-3,3]. CMU-MOSEI is the next generation of CMU-MOSI. The **ICT-MMMO** (Wöllmer et al., 2013) consists of online social review videos annotated at the video level for sentiment. **YouTube** (Morency et al., 2011) contains videos from the social media web site YouTube that span a wide range of product reviews and opinion videos. **MOUD** (Perez-Rosas et al., 2013) consists of product review videos in Spanish. Each video consists of multiple segments labeled to display positive, negative or neutral sentiment. **IEMOCAP** (Busso et al., 2008) consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral) as well as valence, arousal and dominance.

Dataset	# S	# Sp	Mod	Sent	Emo	TL (hh:mm:ss)
CMU-MOSEI	**23,453**	**1,000**	$\{l, v, a\}$	✓	✓	**65:53:36**
CMU-MOSI	2,199	98	$\{l, v, a\}$	✓	✗	02:36:17
ICT-MMMO	340	200	$\{l, v, a\}$	✓	✗	13:58:29
YouTube	300	50	$\{l, v, a\}$	✓	✗	00:29:41
MOUD	400	101	$\{l, v, a\}$	✓	✗	00:59:00
SST	11,855	–	$\{l\}$	✓	✗	–
Cornell	2,000	–	$\{l\}$	✓	✗	–
Large Movie	25,000	–	$\{l\}$	✓	✗	–
STS	5,513	–	$\{l\}$	✓	✗	–
IEMOCAP	10,000	10	$\{l, v, a\}$	✗	✓	11:28:12
SAL	23	4	$\{v, a\}$	✗	✓	11:00:00
VAM	499	20	$\{v, a\}$	✗	✓	12:00:00
VAM-faces	1,867	20	$\{v\}$	✗	✓	–
HUMAINE	50	4	$\{v, a\}$	✗	✓	04:11:00
RECOLA	46	46	$\{v, a\}$	✗	✓	03:50:00
SEWA	538	408	$\{v, a\}$	✗	✓	04:39:00
SEMAINE	80	20	$\{v, a\}$	✗	✓	06:30:00
AFEW	1,645	330	$\{v, a\}$	✗	✓	02:28:03
AM-FED	242	242	$\{v\}$	✗	✓	03:20:25
Mimicry	48	48	$\{v, a\}$	✗	✓	11:00:00
AFEW-VA	600	240	$\{v, a\}$	✗	✓	00:40:00

Table 1: Comparison of the CMU-MOSEI dataset with previous sentiment analysis and emotion recognition datasets. #S denotes the number of annotated data points. #Sp is the number of distinct speakers. Mod indicates the subset of modalities present from $\{(l)anguage, (v)ision, (a)udio\}$. Sent and Emo columns indicate presence of sentiment and emotion labels. TL denotes the total number of video hours.

2.1.2 Language Datasets

Stanford Sentiment Treebank (SST) (Socher et al., 2013) includes fine grained sentiment labels for phrases in the parse trees of sentences collected from movie review data. While SST has larger pool of annotations, we only consider the root level annotations for comparison. **Cornell Movie Review** (Pang et al., 2002) is a collection of 2000 movie-review documents and sentences labeled with respect to their overall sentiment polarity or subjective rating. **Large Movie Review** dataset (Maas et al., 2011) contains text from highly polar movie reviews. **Sanders Tweets Sentiment (STS)** consists of 5513 hand-classified tweets each classified with respect to one of four topics of Microsoft, Apple, Twitter, and Google.

2.1.3 Visual and Acoustic Datasets

The **Vera am Mittag (VAM)** corpus consists of 12 hours of recordings of the German TV talk-

[1] following creative commons license allows for personal unrestricted use and redistribution of the videos

[2] https://github.com/A2Zadeh/CMU-MultimodalDataSDK

show "Vera am Mittag" (Grimm et al., 2008). This audio-visual data is labeled for continuous-valued scale for three emotion primitives: valence, activation and dominance. VAM-Audio and VAM-Faces are subsets that contain on acoustic and visual inputs respectively. **RECOLA** (Ringeval et al., 2013) consists of 9.5 hours of audio, visual, and physiological (electrocardiogram, and electrodermal activity) recordings of online dyadic interactions. **Mimicry** (Bilakhia et al., 2015) consists of audiovisual recordings of human interactions in two situations: while discussing a political topic and while playing a role-playing game. **AFEW** (Dhall et al., 2012, 2015) is a dynamic temporal facial expressions data corpus consisting of close to real world environment extracted from movies.

Detailed comparison of CMU-MOSEI to the datasets in this section is presented in Table 1. CMU-MOSEI has longer total duration as well as larger number of data point in total. Furthermore, CMU-MOSEI has a larger variety in number of speakers and topics. It has all three modalities provided, as well as annotations for both sentiment and emotions.

2.2 Baseline Models

Modeling multimodal language has been the subject of studies in NLP and multimodal machine learning. Notable approaches are listed as follows and indicated with a symbol for reference in the Experiments and Discussion section (Section 5).

MFN: (Memory Fusion Network) (Zadeh et al., 2018a) synchronizes multimodal sequences using a multi-view gated memory that stores intra-view and cross-view interactions through time. ■ **MARN:** (Multi-attention Recurrent Network) (Zadeh et al., 2018b) models intra-modal and multiple cross-modal interactions by assigning multiple attention coefficients. Intra-modal and cross-modal interactions are stored in a hybrid LSTM memory component. ⋆ **TFN** (Tensor Fusion Network) (Zadeh et al., 2017) models inter and intra modal interactions by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions. ⋄ **MV-LSTM** (Multi-View LSTM) (Rajagopalan et al., 2016) is a recurrent model that designates regions inside a LSTM to different views of the data. § **EF-LSTM** (Early Fusion LSTM) concatenates the inputs from different modalities at each time-step and uses that as the input to a single LSTM (Hochreiter and Schmidhuber, 1997; Graves et al., 2013; Schuster and Paliwal, 1997). In case of unimodal models EF-LSTM refers to a single LSTM.

We also compare to the following baseline models: † **BC-LSTM** (Poria et al., 2017b), ♣ **C-MKL** (Poria et al., 2016), ♭ **DF** (Nojavanasghari et al., 2016), ♡ **SVM** (Cortes and Vapnik, 1995; Zadeh et al., 2016b; Perez-Rosas et al., 2013; Park et al., 2014), • **RF** (Breiman, 2001), **THMM** (Morency et al., 2011), **SAL-CNN** (Wang et al., 2016), **3D-CNN** (Ji et al., 2013). For language only baseline models: ∪ **CNN-LSTM** (Zhou et al., 2015), **RNTN** (Socher et al., 2013), ×: **DynamicCNN** (Kalchbrenner et al., 2014), ▷ **DAN** (Iyyer et al., 2015), ≀ **DHN** (Srivastava et al., 2015), ◁ **RHN** (Zilly et al., 2016). For acoustic only baseline models: **AdieuNet** (Trigeorgis et al., 2016), **SER-LSTM** (Lim et al., 2016).

3 CMU-MOSEI Dataset

Understanding expressed sentiment and emotions are two crucial factors in human multimodal language. We introduce a novel dataset for multimodal sentiment and emotion recognition called CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). In the following subsections, we first explain the details of the CMU-MOSEI data acquisition, followed by details of annotation and feature extraction.

3.1 Data Acquisition

Social multimedia presents a unique opportunity for acquiring large quantities of data from various speakers and topics. Users of these social multimedia websites often post their opinions in the forms of monologue videos; videos with only one person in front of camera discussing a certain topic of interest. Each video inherently contains three modalities: language in the form of spoken text, visual via perceived gestures and facial expressions, and acoustic through intonations and prosody.

During our automatic data acquisition process, videos from YouTube are analyzed for the presence of one speaker in the frame using face detection to ensure the video is a monologue. We limit the videos to setups where the speaker's attention is exclusively towards the camera by rejecting videos that have moving cameras (such as camera on bikes or selfies recording while walking). We use a diverse set of 250 frequently used topics in online videos as the seed for acquisition. We restrict the

Figure 1: The diversity of topics of videos in CMU-MOSEI, displayed as a word cloud. Larger words indicate more videos from that topic. The most frequent 3 topics are reviews (16.2%), debate (2.9%) and consulting (1.8%) while the remaining topics are almost uniformly distributed.

number of videos acquired from each channel to a maximum of 10. This resulted in discovering 1,000 identities from YouTube. The definition of a identity is proxy to the number of channels since accurate identification requires quadratic manual annotations, which is infeasible for high number of speakers. Furthermore, we limited the videos to have manual and properly punctuated transcriptions provided by the uploader. The final pool of acquired videos included 5,000 videos which were then manually checked for quality of video, audio and transcript by 14 expert judges over three months. The judges also annotated each video for gender and confirmed that each video is an acceptable monologue. A set of 3228 videos remained after manual quality inspection. We also performed automatic checks on the quality of video and transcript which are discussed in Section 3.3 using facial feature extraction confidence and forced alignment confidence. Furthermore, we balance the gender in the dataset using the data provided by the judges (57% male to 43% female). This constitutes the final set of raw videos in CMU-MOSEI. The topics covered in the final set of videos are shown in Figure 1 as a Venn-style word cloud (Coppersmith and Kelly, 2014) with the size proportional to the number of videos gathered for that topic. The most frequent 3 topics are reviews (16.2%), debate (2.9%) and consulting (1.8%). The remaining topics are almost uniformly distributed [3].

The final set of videos are then tokenized into sentences using punctuation markers manually provided by transcripts. Due to the high quality of the transcripts, using punctuation markers showed better sentence quality than using the Stanford CoreNLP tokenizer (Manning et al., 2014). This was verified on a set of 20 random videos by two experts. After tokenization, a set of 23,453 sentences were chosen as the final sentences in the dataset. This was achieved by restricting each identity to contribute at least 10 and at most 50 sentences to the dataset. Table 2 shows high-level summary statistics of the CMU-MOSEI dataset.

Total number of sentences	23453
Total number of videos	3228
Total number of distinct speakers	1000
Total number of distinct topics	250
Average number of sentences in a video	7.3
Average length of sentences in seconds	7.28
Total number of words in sentences	447143
Total of unique words in sentences	23026
Total number of words appearing at least 10 times in the dataset	3413
Total number of words appearing at least 20 times in the dataset	1971
Total number of words appearing at least 50 times in the dataset	888

Table 2: Summary of CMU-MOSEI dataset statistics.

3.2 Annotation

Annotation of CMU-MOSEI follows closely the annotation of CMU-MOSI (Zadeh et al., 2016a) and Stanford Sentiment Treebank (Socher et al., 2013). Each sentence is annotated for sentiment on a [-3,3] Likert scale of: [−3: highly negative, −2 negative, −1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 highly positive]. Ekman emotions (Ekman et al., 1980) of {happiness, sadness, anger, fear, disgust, surprise} are annotated on a [0,3] Likert scale for presence of emotion x: [0: no evidence of x, 1: weakly x, 2: x, 3: highly x]. The annotation was carried out by 3 crowdsourced judges from Amazon Mechanical Turk platform. To avert implicitly biasing the judges and to capture the raw perception of the crowd, we avoided extreme annotation training and instead provided the judges with a 5 minutes training video on how to use the annotation system. All the annotations have been carried out by only master workers with higher than 98% approval rate to assure high quality annotations [4].

Figure 2 shows the distribution of sentiment and emotions in CMU-MOSEI dataset. The distribution

[3] more detailed analysis such as exact percentages and number of videos per topic are available in the supplementary material

[4] Extensive statistics of the dataset including the crawling mechanism, the annotation UI, training procedure for the workers, agreement scores are available in submitted supplementary material available on arXiv.

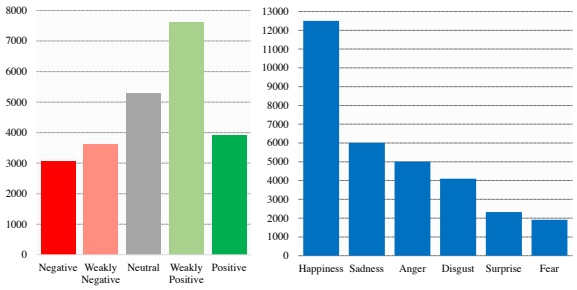

Figure 2: Distribution of sentiment and emotions in the CMU-MOSEI dataset. The distribution shows a natural skew towards more frequently used emotions. However, the least frequent emotion, fear, still has 1,900 data points which is an acceptable number for machine learning studies.

shows a slight shift in favor of positive sentiment which is similar to distribution of CMU-MOSI and SST. We believe that this is an implicit bias in online opinions being slightly shifted towards positive, since this is also present in CMU-MOSI. The emotion histogram shows different prevalence for different emotions. The most common category is happiness with more than 12,000 positive sample points. The least prevalent emotion is fear with almost 1900 positive sample points which is an acceptable number for machine learning studies.

3.3 Extracted Features

Data points in CMU-MOSEI come in video format with one speaker in front of the camera. The extracted features for each modality are as follows (for other benchmarks we extract the same features):

Language: All videos have manual transcription. Glove word embeddings (Pennington et al., 2014) were used to extract word vectors from transcripts. Words and audio are aligned at phoneme level using P2FA forced alignment model (Yuan and Liberman, 2008). Following this, the visual and acoustic modalities are aligned to the words by interpolation. Since the utterance duration of words in English is usually short, this interpolation does not lead to substantial information loss.

Visual: Frames are extracted from the full videos at 30Hz. The bounding box of the face is extracted using the MTCNN face detection algorithm (Zhang et al., 2016). We extract facial action units through Facial Action Coding System (FACS) (Ekman et al., 1980). Extracting these action units allows for accurate tracking and understanding of the facial expressions (Baltrušaitis et al., 2016). We also extract a set of six basic emotions purely from static faces using Emotient FACET (iMotions, 2017). MultiComp OpenFace (Baltrušaitis et al., 2016) is used to extract the set of 68 facial landmarks, 20 facial shape parameters, facial HoG features, head pose, head orientation and eye gaze (Baltrušaitis et al., 2016). Finally, we extract face embeddings from commonly used facial recognition models such as DeepFace (Taigman et al., 2014), FaceNet (Schroff et al., 2015) and SphereFace (Liu et al., 2017).

Acoustic: We use the COVAREP software (Degottex et al., 2014) to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features (Drugman and Alwan, 2011), glottal source parameters (Drugman et al., 2012; Alku et al., 1997, 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013). All extracted features are related to emotions and tone of speech.

4 Multimodal Fusion Study

From the linguistics perspective, understanding the interactions between language, visual and audio modalities in multimodal language is a fundamental research problem. While previous works have been successful with respect to accuracy metrics, they have not created new insights on how the fusion is performed in terms of what modalities are related and how modalities engage in an interaction during fusion. Specifically, to understand the fusion process one must first understand the n-modal dynamics (Zadeh et al., 2017). n-modal dynamics state that there exists different combination of modalities and that all of these combinations must be captured to better understand the multimodal language. In this paper, we define building the n-modal dynamics as a hierarchical process and propose a new fusion model called the Dynamic Fusion Graph (DFG). DFG is easily interpretable through what is called efficacies in graph connections. To utilize this new fusion model in a multimodal language framework, we build upon Memory Fusion Network (MFN) by replacing the original fusion component in the MFN with our DFG. We call this resulting model the Graph Memory Fusion Network (Graph-MFN). Once the model is trained end to end, we analyze the efficacies in the DFG to study the fusion mechanism learned for modalities in multimodal language. In addition to being an interpretable fusion mechanism,

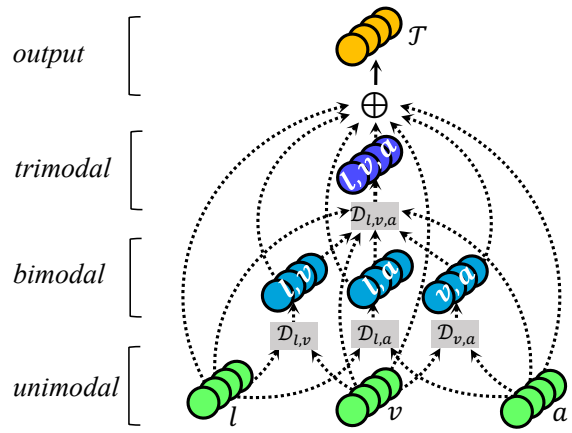

Figure 3: The structure of Dynamic Fusion Graph (DFG) for three modalities of $\{(l)anguage, (v)ision, (a)coustic\}$. Dashed lines in DFG show the dynamic connections between vertices controlled by the efficacies (α).

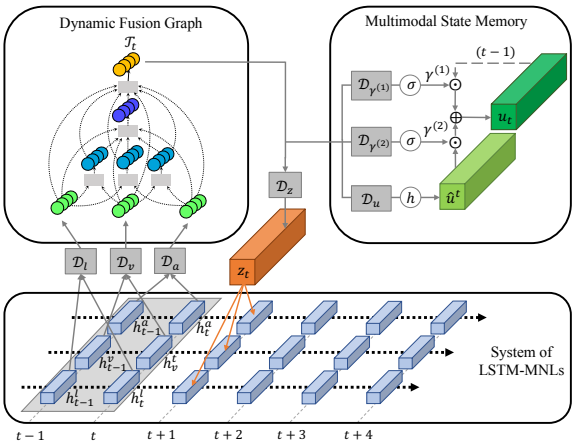

Figure 4: The overview of Graph Memory Fusion Network (Graph-MFN) pipeline. Graph-MFN replaces the fusion block in MFN with a Dynamic Fusion Graph (DFG). For description of variables and memory formulation please refer to the original Memory Fusion Network paper (Zadeh et al., 2018a).

Graph-MFN also outperforms previously proposed state-of-the-art models for sentiment analysis and emotion recognition on the CMU-MOSEI.

4.1 Dynamic Fusion Graph

In this section we discuss the internal structure of the proposed Dynamic Fusion Graph (DFG) neural model (Figure 3. DFG has the following properties: 1) it explicitly models the n-modal interactions, 2) does so with an efficient number of parameters (as opposed to previous approaches such as Tensor Fusion (Zadeh et al., 2017)) and 3) can dynamically alter its structure and choose the proper fusion graph based on the importance of each n-modal dynamics during inference. We assume the set of modalities to be $M = \{(l)anguage, (v)ision, (a)coustic\}$. The unimodal dynamics are denoted as $\{l\}, \{v\}, \{a\}$, the bimodal dynamics as $\{l, v\}, \{v, a\}, \{l, a\}$ and trimodal dynamics as $\{l, v, a\}$. These dynamics are in the form of latent representations and are each considered as vertices inside a graph $G = (V, E)$ with V the set of vertices and E the set of edges. A directional neural connection is established between two vertices v_i and v_j only if $v_i \subset v_j$. For example, $\{l\} \subset \{l, v\}$ which results in a connection between $< language >$ and $< language, vision >$. This connection is denoted as an edge e_{ij}. $\mathcal{D}_j$ takes as input all v_i that satisfy the neural connection formula above for v_j.

We define an efficacy for each edge e_{ij} denoted as α_{ij}. v_i is multiplied by α_{ij} before being used as input to D_j. Each α is a sigmoid activated probability neuron which indicates how strong or weak the connection is between v_i and v_j. αs are the main source of interpretability in DFG. The vector of all αs is inferred using a deep neural network $\mathcal{D}_\alpha$ which takes as input singleton vertices in V (l, v, and a). We leave it to the supervised training objective to learn parameters of $\mathcal{D}_\alpha$ and make good use of efficacies, thus dynamically controlling the structure of the graph. The singleton vertices are chosen for this purpose since they have no incoming edges thus no efficacy associated with those edges (no efficacy is needed to infer the singleton vertices). The same singleton vertices l, v, and a are the inputs to the DFG. In the next section we discuss how these inputs are given to DFG. All vertices are connected to the output vertex $\mathcal{T}_t$ of the network via edges scaled by their respective efficacy. The overall structure of the vertices, edges and respective efficacies is shown in Figure 3. There are a total of 8 vertices (counting the output vertex), 19 edges and subsequently 19 efficacies.

4.2 Graph-MFN

To test the performance of DFG, we use a similar recurrent architecture to Memory Fusion Network (MFN). MFN is a recurrent neural model with three main components 1) System of LSTMs: a set of parallel LSTMs with each LSTM modeling a single modality. 2) Delta-memory Attention Network is the component that performs multimodal fusion

Dataset	MOSEI Sentiment						MOSEI Emotions											
Task	Sentiment						Anger		Disgust		Fear		Happy		Sad		Surprise	
Metric	A^2	F1	A^5	A^7	MAE	r	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
LANGUAGE																		
SOTA2	74.1§	74.1$^{\triangleright}$	43.1$^{\wr}$	42.9$^{\wr}$	0.75§	0.46$^{\wr}$	56.0$^{\cup}$	71.0$^{\times}$	59.0§	67.1$^{\triangleright}$	56.2§	79.7§	53.0$^{\triangleright}$	44.1$^{\triangleright}$	53.8$^{\wr}$	49.9$^{\wr}$	53.2$^{\times}$	70.0$^{\triangleright}$
SOTA1	74.3$^{\triangleright}$	74.1§	43.2§	43.2§	0.74$^{\triangleright}$	0.47§	56.6$^{\wr}$	71.8$^{\bullet}$	64.0$^{\triangleright}$	72.6$^{\bullet}$	58.8$^{\times}$	89.8$^{\bullet}$	54.0§	47.0§	54.0§	61.2$^{\bullet}$	**54.3**$^{\triangleright}$	85.3$^{\bullet}$
VISUAL																		
SOTA2	73.8§	73.5§	42.5$^{\triangleright}$	42.5$^{\triangleright}$	0.78$^{\wr}$	0.41$^{\heartsuit}$	54.4$^{\wr}$	64.6§	54.4$^{\heartsuit}$	71.5$^{\triangleleft}$	51.3§	78.4§	53.4$^{\wr}$	40.8§	54.3$^{\triangleright}$	60.8$^{\bullet}$	51.3$^{\triangleright}$	84.2§
SOTA1	73.9$^{\triangleright}$	73.7$^{\triangleright}$	42.7$^{\wr}$	42.7$^{\wr}$	0.78§	0.43$^{\wr}$	60.0§	71.0$^{\bullet}$	60.3$^{\wr}$	72.4$^{\bullet}$	**64.2**$^{\heartsuit}$	89.8$^{\bullet}$	57.4$^{\bullet}$	49.3$^{\bullet}$	57.7§	61.5$^{\triangleleft}$	51.8§	85.4$^{\bullet}$
ACOUSTIC																		
SOTA2	74.2$^{\wr}$	73.8$^{\triangle}$	42.1$^{\triangle}$	42.1$^{\triangle}$	0.78$^{\triangleright}$	0.43§	55.5$^{\triangleleft}$	51.8$^{\triangle}$	58.9$^{\triangleright}$	72.4$^{\bullet}$	58.5$^{\triangleright}$	89.8$^{\bullet}$	57.2$^{\cap}$	55.5$^{\cap}$	58.9$^{\triangleleft}$	65.9$^{\triangleleft}$	52.2$^{\heartsuit}$	83.6$^{\cap}$
SOTA1	74.2$^{\triangle}$	73.9$^{\wr}$	42.4$^{\cap}$	42.4$^{\cap}$	0.74$^{\cap}$	0.43$^{\triangleright}$	56.4$^{\triangle}$	71.9$^{\bullet}$	60.9§	72.4$^{\bullet}$	62.7§	89.8$^{\triangleleft}$	61.5§	61.4§	**62.0**$^{\cap}$	**69.2**$^{\cap}$	**54.3**$^{\triangleleft}$	85.4$^{\bullet}$
MULTIMODAL																		
SOTA2	76.0$^{\#}$	76.0$^{\#}$	44.7†	44.6†	0.72*	0.52*	56.0$^{\diamond}$	71.4$^{\flat}$	65.2$^{\#}$	71.4$^{\#}$	56.7§	**89.9**$^{\#}$	57.8§	66.6*	58.9*	60.8$^{\#}$	52.2*	85.4$^{\bullet}$
SOTA1	76.4$^{\diamond}$	76.4$^{\diamond}$	44.8*	44.7*	0.72$^{\#}$	0.52$^{\#}$	60.5*	72.0$^{\bullet}$	67.0$^{\flat}$	73.2$^{\bullet}$	60.0$^{\heartsuit}$	**89.9**$^{\bullet}$	**66.5***	**71.0**$^{\blacksquare}$	59.2§	61.8$^{\bullet}$	53.3$^{\#}$	85.4$^{\#}$
Graph-MFN	**76.9**	**77.0**	**45.1**	**45.0**	**0.71**	**0.54**	**62.6**	**72.8**	**69.1**	**76.6**	62.0	**89.9**	66.3	66.3	60.4	66.9	53.7	**85.5**

Table 3: Results for sentiment analysis and emotion recognition on the MOSEI dataset (reported results are as of 5/11/2018. please check the CMU Multimodal Data SDK github for current state of the art and new features for CMU-MOSEI and other datasets). SOTA1 and SOTA2 refer to the previous best and second best state-of-the-art models (from Section 2) respectively. Compared to the baselines Graph-MFN achieves superior performance in sentiment analysis and competitive performance in emotion recognition. For all metrics, higher values indicate better performance except for MAE where lower values indicate better performance.

by assigning coefficients to highlight cross-modal dynamics. 3) Multiview Gated Memory is a component that stores the output of multimodal fusion. We replace the Delta-memory Attention Network with DFG and refer to the modified model as Graph Memory Fusion Network (Graph-MFN). Figure 4 shows the overall architecture of the Graph-MFN.

Similar to MFN, Graph-MFN employs a system of LSTMs for modeling individual modalities. c_l, c_v, and c_a represent the memory of LSTMs for language, vision and acoustic modalities respectively. D_m, $m \in \{l, v, a\}$ is a fully connected deep neural network that takes in $h^m_{[t-1,t]}$ the LSTM representation across two consecutive timestamps, which allows the network to track changes in memory dimensions across time. The outputs of D_l, D_v and D_a are the singleton vertices for the DFG. The DFG models cross-modal interactions and encodes the cross-modal representations in its output vertex $\mathcal{T}_t$ for storage in the Multi-view Gated Memory u_t. The Multi-view Gated Memory functions using a network D_u that transforms $\mathcal{T}_t$ into a proposed memory update $\hat{u}_t$. γ_1 and γ_2 are the Multi-view Gated Memory's retain and update gates respectively and are learned using networks D_{γ_1} and D_{γ_2}. Finally, a network D_z transforms $\mathcal{T}_t$ into a multimodal representation z_t to update the system of LSTMs. The output of Graph-MFN in all the experiments is the output of each LSTM h^m_T as well as contents of the Multi-view Gated Memory at time T (last recurrence timestep), u_T. This output is subsequently connected to a classification or regression layer for final prediction (for sentiment and emotion recognition).

5 Experiments and Discussion

In our experiments, we seek to evaluate how modalities interact during multimodal fusion by studying the efficacies of DFG through time.

Table 3 shows the results on CMU-MOSEI. Accuracy is reported as A^x where x is the number of sentiment classes as well as F1 measure. For regression we report MAE and correlation (r). For emotion recognition due to the natural imbalances across various emotions, we use weighted accuracy (Tong et al., 2017) and F1 measure. Graph-MFN shows superior performance in sentiment analysis and competitive performance in emotion recognition. Therefore, DFG is both an effective and interpretable model for multimodal fusion.

To better understand the internal fusion mechanism between modalities, we visualize the behavior of the learned DFG efficacies in Figure 5 for various cases (deep red denotes high efficacy and deep blue denotes low efficacy).

Multimodal Fusion has a Volatile Nature: The first observation is that the structure of the DFG is changing case by case and for each case over time. As a result, the model seems to be selectively prioritizing certain dynamics over the others. For example, in case (I) where all modalities are informative, all efficacies seem to be high, imply-

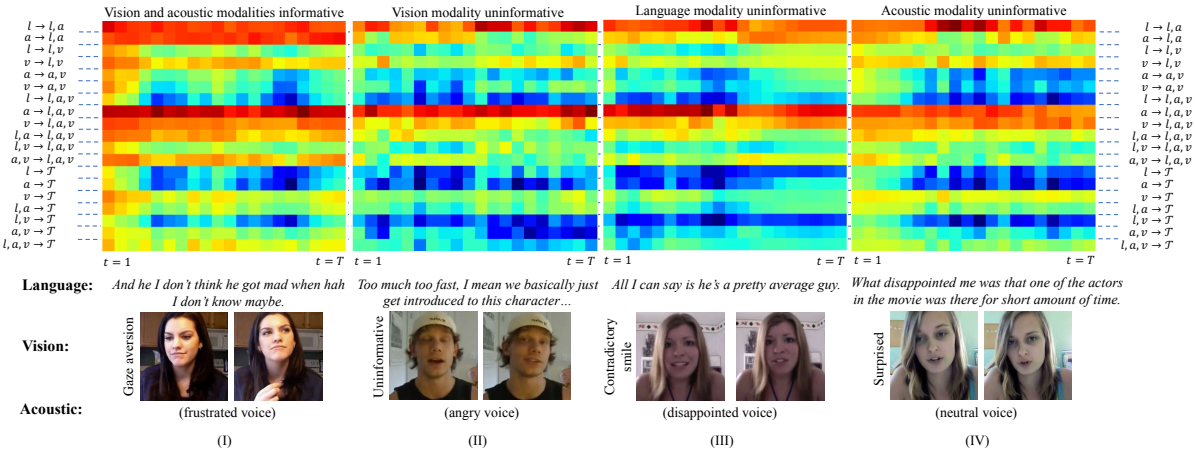

Figure 5: Visualization of DFG efficacies across time. The efficacies (thus the DFG structure) change over time as DFG is exposed to new information. DFG is able choose which n-modal dynamics to rely on. It also learns priors about human communication since certain efficacies (thus edges in DFG) remain unchanged across time and across data points.

ing that the DFG is able to find useful information in unimodal, bimodal and trimodal interactions. However, in cases (II) and (III) where the visual modality is either uninformative or contradictory, the efficacies of $v \rightarrow l, v$ and $v \rightarrow l, a, v$ and $l, a \rightarrow l, a, v$ are reduced since no meaningful interactions involve the visual modality.

Priors in Fusion: Certain efficacies remain unchanged across cases and across time. These are priors from Human Multimodal Language that DFG learns. For example the model always seems to prioritize fusion between language and audio in $(l \rightarrow l, a)$, and $(a \rightarrow l, a)$. Subsequently, DFG gives low values to efficacies that rely unilaterally on language or audio alone: the $(l \rightarrow \tau)$ and $(a \rightarrow \tau)$ efficacies seem to be consistently low. On the other hand, the visual modality appears to have a partially isolated behavior. In the presence of informative visual information, the model increases the efficacies of $(v \rightarrow \tau)$ although the values of other visual efficacies also increase.

Trace of Multimodal Fusion: We trace the dominant path that every modality undergoes during fusion: 1) *language* tends to first fuse with audio via $(l \rightarrow l, a)$ and the language and acoustic modalities together engage in higher level fusions such as $(l, a \rightarrow l, a, v)$. Intuitively, this is aligned with the close ties between language and audio through word intonations. 2) The *visual* modality seems to engage in fusion only if it contains meaningful information. In cases (I) and (IV), all the paths involving the visual modality are relatively active while in cases (II) and (III) the paths involving the visual modality have low efficacies. 3) The *acoustic* modality is mostly present in fusion with the language modality. However, unlike language, the acoustic modality also appears to fuse with the visual modality if both modalities are meaningful, such as in case (I).

An interesting observation is that in almost all cases the efficacies of unimodal connections to terminal $\mathcal{T}$ is low, implying that $\mathcal{T}$ prefers to not rely on just one modality. Also, DFG always prefers to perform fusion between language and audio as in most cases both $l \rightarrow l, a$ and $a \rightarrow l, a$ have high efficacies; intuitively in most natural scenarios language and acoustic modalities are highly aligned. Both of these cases show unchanging behaviors which we believe DFG has learned as natural priors of human communicative signal.

With these observations, we believe that DFG has successfully learned how to manage its internal structure to model human communication.

6 Conclusion

In this paper we presented the largest dataset of multimodal sentiment analysis and emotion recognition called CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). CMU-MOSEI consists of 23,453 annotated sentences from more than 1000 online speakers and 250 different topics. The dataset expands the horizons of Human Multimodal Language studies in NLP. One such study was presented in this paper where we analyzed the structure of multimodal fusion in sentiment analysis and emotion recognition. This was

done using a novel interpretable fusion mechanism called Dynamic Fusion Graph (DFG). In our studies we investigated the behavior of modalities in interacting with each other using built-in efficacies of DFG. Aside analysis of fusion, DFG was trained in the Memory Fusion Network pipeline and showed superior performance in sentiment analysis and competitive performance in emotion recognition.

Acknowledgments

This material is based upon work partially supported by the National Science Foundation (Award #1833355) and Oculus VR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or Oculus VR, and no official endorsement should be inferred.

References

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.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406* .

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.

Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern Recognition Letters* 66(Supplement C):52 – 61. Pattern Recognition in Human Computer Interaction. https://doi.org/https://doi.org/10.1016/j.patrec.2015.03.005

Leo Breiman. 2001. Random forests. *Mach. Learn.* 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42(4):335–359. https://doi.org/10.1007/s10579-008-9076-6.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI 2017, pages 163–171. https://doi.org/10.1145/3136755.3136801.

Glen Coppersmith and Erin Kelly. 2014. Dynamic wordclouds and vennclouds for exploratory data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 22–29.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20(3):273–297. https://doi.org/10.1023/A:1022627411411.

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.

A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia* 19(3):34–41. https://doi.org/10.1109/MMUL.2012.26.

Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI '15, pages 423–426. https://doi.org/10.1145/2818346.2829994.

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.

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

A. Graves, A. r. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. pages 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947.

Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2008. The vera am mittag german audio-visual emotional speech database. In *ICME*. IEEE, pages 865–868.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmerman. 2018. Memn: Multimodal emotional memory network for emotion recognition in dyadic conversational videos. In *NAACL*.

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

iMotions. 2017. Facial expression analysis. goo.gl/1rh1JN.

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.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* .

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.

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.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 142–150. http://www.aclweb.org/anthology/P11-1015.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

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 Interactions*. ACM, pages 169–176.

Friedrich Max Müller. 1866. *Lectures on the science of language: Delivered at the Royal Institution of Great Britain in April, May, & June 1861*, volume 1. Longmans, Green.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI 2016, pages 284–288. https://doi.org/10.1145/2993148.2993176.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*. pages 79–86.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI '14, pages 50–57. https://doi.org/10.1145/2663204.2663260.

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

Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*. Sofia, Bulgaria.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context dependent sentiment analysis in user generated videos. In *Association for Computational Linguistics*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Association for Computational Linguistics*.

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.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *FG*. IEEE Computer Society, pages 1–8.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*. IEEE Computer Society, pages 815–823.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.* 45(11):2673–2681. https://doi.org/10.1109/78.650093.

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, volume 1631, page 1642.

Rupesh K Srivastava, Klaus Greff, and Juergen Schmidhuber. 2015. Training very deep networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 2377–2385. http://papers.nips.cc/paper/5850-training-very-deep-networks.pdf.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, CVPR '14, pages 1701–1708. https://doi.org/10.1109/CVPR.2014.220.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1547–1556.

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.

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.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927* .

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923* .

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.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23(10):1499–1503.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *CoRR* abs/1511.08630.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2016. Recurrent Highway Networks. *arXiv preprint arXiv:1607.03474* .