Computational Modeling of Human Multimodal Language: The MOSEI Dataset and Interpretable Dynamic Fusion

Paul Pu Liang, Ruslan Salakhutdinov
Machine Learning Department
Carnegie Mellon University
{pliang,rsalakhu}@cs.cmu.edu

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

Abstract

Computational modeling of human multimodal language is an emerging research area in natural language processing spanning the language, visual and acoustic modalities. Comprehending multimodal language requires not only the modeling of interactions within each modality (intra-modal interactions), but more importantly the interactions between modalities (cross-modal interactions). Modeling these interactions lie at the core of multimodal language analysis. From a resource perspective, there is a genuine need for large scale datasets that allow for in-depth studies of human multimodal language. In this paper we introduce CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), the largest dataset for multimodal sentiment analysis and emotion recognition. In addition, we propose a novel multimodal fusion technique called the Graph Memory Fusion Network (GMFN) that dynamically fuses modalities in a hierarchical manner. Using data from CMU-MOSEI and GMFN, we conduct experiments to investigate the hierarchical interactions between modalities in human multimodal language. Unlike previously proposed fusion techniques, GMFN is highly interpretable and achieves superior performance when compared to the previous state of the art, demonstrating that GMFN is highly suitable for multimodal language analysis.

1 Introduction

Theories in the origin of language have identified the combination of verbal and nonverbal behaviors (visual and acoustic modality) as the prime form of communication utilized by humans throughout evolution (Müller, 1866). From a computational perspective, the modeling of human language across both verbal and nonverbal behaviors is an upcoming research area that extends the boundaries of natural language processing. This research area focuses on modeling tasks such as multimodal sentiment analysis (Morency et al., 2011; Glorot et al., 2011), emotion recognition (Busso et al., 2008) and personality traits recognition (Park et al., 2014) from multimodal temporal signals including language (spoken words), visual (facial expressions, gestures), and acoustic (prosody, speech tone).

At its core, these multimodal signals are highly structured with two prime forms of interactions: intra-modal and cross-modal interactions (Rajagopalan et al., 2016b). Intra-modal interactions refer to information within a specific modality, independent of other modalities. For example, the arrangement of words in a sentence according to the generative grammar of the language (language modality) (Chomsky, 1957) or the sequence of facial muscles for the presentation of a frown (vision modality). Cross-modal interactions refer to interactions between modalities. For example, the simultaneous co-occurrence of a smile with a positive sentence or the delayed occurrence of a laughter after the end of sentence. Modeling these intra-modal and cross-modal interactions lie at the heart of human multimodal language analysis and has recently become a centric research direction in both NLP (Hazarika et al., 2018; Pham et al., 2018; Chen et al., 2017) and multimodal machine learning (Tsai et al., 2018; Srivastava and Salakhutdinov, 2012; Ngiam et al., 2011).

However, from a resource perspective, datasets made for modeling multimodal language have severe shortcomings in the following aspects:

Diversity in training samples: Having a rich diversity in training samples is crucial for com-

prehensive studies of human multimodal language. This is due to the high complexity of the underlying data distribution. This complexity is rooted in the variability of intra-modal and cross-modal dynamics for language, vision and acoustic modalities (Rajagopalan et al., 2016b). Previously proposed datasets for multimodal language are generally small in size due to the difficulties associated with data acquisition, costs of annotations, as well as significant the amount of feature extraction and post-processing required.

Diversity in topics: Having a diverse set of topics allows our models to generalizable across different domains. Models trained on only a few topics generalize poorly as the learnt language and nonverbal features might be highly correlated to the topics of videos they were trained on. For example, a lack of diversity in topics might cause a model to always associate movie reviews with negative sentiment. Increasing the diversity of topics with rich examples from each topic allows our models the generalize across multiple domains.

Diversity in speakers: Human multimodal language is highly idiosyncratic: individuals prefer their own writing and speaking styles. Training models on only few speakers can lead to a poor solutions where models memorize the identity of speakers. Having a diverse set of speakers opens the door towards generalizable models of multimodal language (Zadeh et al., 2016) and allows us to draw reasonable conclusions over the rich expressiveness of human multimodal language.

Diversity in annotations: Human multimodal language is broadly defined across expressions of sentiment, emotions and personality traits. The expression of each intent is unique. It is crucial that our methods learn to generalize across different expressions of intent such as sentiment, emotions and various speaker personality traits. Furthermore, having multiple labels to predict allows our computational methods to discover potential relationships between labels. For example, there are certainly strong correlations between positive sentiment and positive emotions. Such a variety of labels could allow for multi-task learning and a step towards deeper understanding of human language from multiple perspectives.

This paper addresses the lack of multimodal resources with diversity in samples, topics, speakers and annotations. Our first contribution is to present the scientific community with the largest dataset of multimodal sentiment and emotion recognition called CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). CMU-MOSEI contains 23,500 annotated sentence utterance video segments from 1,000 distinct speakers and 250 topics. This diverse set of speakers, topics, annotations, and samples allows for generalizable studies of human multimodal language. All the videos are gathered from online video sharing websites and follow creative commons license that allows for personal unrestricted use and redistribution. The multimodal dataset and a multimodal data loading framework are provided to the scientific community to encourage valuable research in human multimodal language analysis.

Our second contribution is an interpretable fusion model called the Graph Memory Fusion Network (GMFN). As compared to previously proposed models, GMFN is specifically designed with interpretability in mind: fusion is performed in a hierarchical manner so that the importance of every combination of modalities can be analyzed. Specifically, GMFN contains a Dynamic Fusion Graph module with built-in efficacies that allow us to interpret the interactions between modalities during fusion. This allows us to study the nature of cross-modal dynamics in multimodal language which we do so in detail in our experiments. Furthermore, GMFN achieves superior performance compared to previously proposed models on CMU-MOSEI as well as excellent results on 6 additional datasets relating to multimodal sentiment analysis, multimodal emotion recognition and multimodal speaker trait recognition.

2 Related Works

In this section we compare the CMU-MOSEI to previously proposed dataset 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., 2016) is a collection of 2199 opinion video clips each annotated with

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
IEMOCAP	10,000	10	$\{l, v, a\}$	✗	✓	11:28:12
SST	11,855	–	$\{l\}$	✓	✗	–
Cornell	2,000	–	$\{l\}$	✓	✗	–
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

Table 1: Comparison between the CMU-MOSEI dataset with standard 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 $\{l, v, a\}$. Sent and Emo columns indicate presence of sentiment and emotion labels. TL denotes the total number of video hours.

sentiment in the range [-3,3]. 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.

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-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 a longer total duration as well as a larger number of total data points. Furthermore, CMU-MOSEI has a significantly greater variety in the number of speakers and topics. It has features from all three modalities of language, visual and acoustic. Finally, CMU-MOSEI is annotated 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.

The first category of models simplify the temporal aspect of videos by averaging each modality's information through time (Abburi et al., 2016). These approaches then use this as input to a non-temporal learning model such as Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; **?**; Perez-Rosas et al., 2013; Park et al., 2014), Random Forests (Breiman, 2001) or Neural Networks (Nojavanasghari et al., 2016; Poria et al., 2016; Wang et al., 2016). These models are successful in understanding co-occurrences of multimodal information. However, the lack of temporal modeling is a major flaw as these models do not have the resolution to work with multiple contradictory evidences, eg. if a smile and frown happen together in a segments. Furthermore, the representation over long periods of time become less informative and there-

fore the performances of these approaches decrease as the segments increase in length.

The second category of research models temporal information using probabilistic models. In particular, the application of graphical models in temporal sequence modeling has been an important research problem due to their interpretability and mathematical interpretations. Hidden Markov Models (HMMs) (Baum and Petrie, 1966), Conditional Random Fields (CRFs) (Lafferty et al., 2001) and Hidden Conditional Random Fields (HCRFs) (Quattoni et al., 2007) were shown to work well on unimodal sequences such as language (Misawa et al., 2017; Ma and Hovy, 2016; Huang et al., 2015) and audio (Yuan and Liberman, 2008). These temporal graphical models have been extended to work with multimodal data (Morency et al., 2011). Several methods have been proposed including multi-view HCRFs where the potentials of the HCRF are changed to facilitate multiple views. (Song et al., 2012) extends the HCRF where view-shared and view specific sub-structures are explicitly learned to capture the interaction between views. (Song et al., 2013) proposes using multi-layered CRFs with latent variables to learn hidden spatio-temporal dynamics. Feature representations are learned at every layer and this is repeated to obtain a hierarchical sequence summary (HSS) representation. (Song et al., 2013) extends this for multi-view data.

The third category has emerged recently with the advent of deep learning. Recurrent Neural Networks, specially Long-short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have been extensively used for language (Socher et al., 2013; Kalchbrenner et al., 2014; Zhou et al., 2015; Srivastava et al., 2015; Zilly et al., 2016) and speech (Trigeorgis et al., 2016; Lim et al., 2016) based sequence modeling due to their superior performance in capturing long term dependencies. Extensions of LSTMs have also been proposed in a multi-view setting. The 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). The Multi-view LSTM partitions the LSTM memory into different components for explicit modeling of different views (Rajagopalan et al., 2016a). The Recurrent Multistage Fusion Network performs multiple stages of fusion through highlighting mechanisms (Liang et al., 2018a). The Multimodal Local-Global Fusion model uses Bayesian ranking algorithms to model both speaker-dependent and speaker-independent dynamics (Liang et al., 2018b). The Memory Fusion Network (Zadeh et al., 2018a) studied the synchronization of multimodal sequences using a multi-view gated memory that stores intra-view and cross-view interactions through time. Other methods proposed learning binary gating mechanisms to remove noisy modalities that either provide contradictory evidences or are redundant to multimodal prediction (Chen et al., 2017), using generative-discriminative objective functions to learn joint (Pham et al., 2018) or factorized multimodal representations (Tsai et al., 2018), or modeling inter and intra modal interactions by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions (Liu et al., 2018; Zadeh et al., 2017).

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). CMU-MOSEI is the largest to date with both attributes annotated. In the following subsections, we first explain the details of the CMU-MOSEI data acquisition, crawling mechanism, followed by details of annotation, inter-annotator agreement and feature extraction.

3.1 Data Acquisition

Social multimedia presents a unique opportunity for acquiring large quantities of data from various speakers and topics. We only use websites that support creative commons license which allows for open usage of data. 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 number of videos acquired from each channel to a maximum of 10 to prevent excessive duplicates in identity, providing a diverse set of speakers. 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 accurately punctuated transcriptions provided by the uploader in order to facilitate computational language analysis on the collected dataset.

3.1.1 Crawl System

We developed a crawler that can search YouTube and filter videos with only one person in front of the camera. This filter is implemented by extracting a number of frames from each video and using OpenCV's (Itseez, 2015) Haar cascades to estimate how many faces are in each video. The crawler is supplied a search term which it then forwards to the YouTube Data API. The search terms provide a rough estimate of topics since they are directly connected to meta-data provided by the uploader.

Figure 1 shows the distribution of the video topics used in CMU-MOSEI. The diversity of the video topics brings the following generalizability advantages: (1) models trained on CMU-MOSEI will be generalizable across different topics and the notion of dataset domain is marginalized, (2) the diversity of topics bring variety of speakers, which allows the trained models to be generalizable across different speakers, and (3) the diversity in topics brings diversity in recording setups which allows the trained models to be generalizable across microphones and cameras with different intrinsic parameters. This diversity makes CMU-MOSEI a one-of-a-kind dataset for sentiment analysis and emotion recognition.

3.1.2 Transcripts

The crawled videos are limited to only videos with user-provided transcripts (which we rely on the YouTube Data API for). However to ensure that the user-provided transcript is reliable, we further post-process with the following filters: 1) punctuation – we use various heuristics about punctuation to ensure that the transcript is high quality, 2) alignment – we ensure that the forced alignment using P2FA (Yuan and Liberman, 2008) passes with high confidence. These filters allow us to filter out videos with bad transcripts.

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 5 most frequent topics are: reviews (16.2%), debate (2.9%), consulting (1.8%), financial (1.8%) and speech (1.6%). The remaining topics are almost uniformly distributed at around 0.5%-1.5% each.

3.2 Dataset Splits

The CMU-MOSEI Mega Corpus facilitates both machine learning and behavioral studies. The dataset in its complete form can be used for machine learning research as it contains a rather balanced distribution across various sentiment scores. However, this does not represent the true distribution of monologue videos on YouTube. To compensate for this bias, CMU-MOSEI Natural Split is a subset of the dataset that was crawled without any form of sentiment and emotion guidance, and reflects a more representative sample of the distribution of sentiment and emotion polarity in YouTube monologues. This subset contains fewer polarized videos with a majority of displaying neutral sentiment. In the following subsections, we first discuss the statistics of CMU-MOSEI Natural Split, and then discuss how we acquired videos with more polarized sentiment and emotion in order to balance the complete CMU-MOSEI dataset.

3.2.1 CMU-MOSEI Natural Split

CMU-MOSEI Natural Split contains sentences randomly sampled from YouTube monologues. The

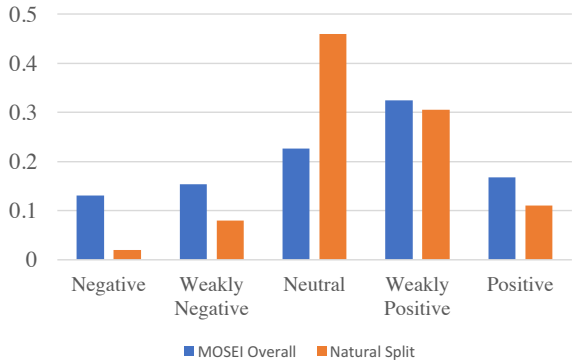

Figure 2: Distribution of sentiment labels for CMU-MOSEI Overall and Natural split. This figure does not represent magnitude (Overall 23,500 sentences but Natural has 7,500 sentences), only ratio.

distribution of annotated sentiment for this split is skewed towards neutral sentences. Due to a combination of factors such as video topic and gender in the selection of these videos, we believe this is the distribution of uttered sentences in YouTube monologues. From a machine learning perspective, this distribution is not ideal since there are many neutral sentences. This dataset contains a total of 7,500 sentences from more than 245 topics and 721 different speakers with almost an equal number of male/female speakers (57.2% male and 42.8% female). Figure 2 shows the distribution of sentiment in Natural Split compared to the overall dataset.

3.2.2 CMU-MOSEI Guided Crawl

To compensate for the lack of videos with polarized sentiment, we use a text-based sentiment analysis model based on the best performing text-based model in (Zadeh et al., 2017) trained on the CMU-MOSI dataset and sentences specifically annotated for this task. We use this model to detect the most polarized videos which we crawled, and sent these for annotation on Amazon Mechanical Turk. We also use 500 polarized videos of POM dataset (Park et al., 2014) which have manual sentiment annotations and extract their sentences. A total of 16,000 sentences are extracted using guided crawl.

3.3 MOSEI Final Statistics

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 is discussed in Section 3.6 using facial feature extraction confidence and forced alignment confidence.

3.3.1 Channel IDs

There are a total number of 734 unique YouTube channels from which male and female videos are extracted. We use the channel id as a heuristic to approximate the number of speakers as well as their gender. Each channel gives at most two identities: one male and one female. Using this information, 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.

3.3.2 Video Topics

The list of crawling search terms had a total of 1962 terms. However, only around 250 resulted in acceptable videos. Since the search terms are related to the metadata of the videos provided by the uploader, we can make the assumption that they are highly related to the general topic of video. Using these statistics, the 5 most frequent topics are: reviews (16.2%), debate (2.9%), consulting (1.8%), financial (1.8%) and speech (1.6%). The remaining topics are almost uniformly distributed at around 0.5%-1.5% each. 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.

3.4 Post-processing

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.

3.5 Annotation

Annotation of CMU-MOSEI follows closely the annotation of CMU-MOSI (Zadeh et al., 2016) and

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: CMU-MOSEI dataset summary of statistics.

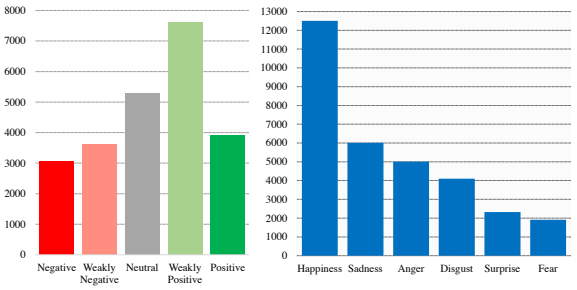

Figure 3: Distribution of sentiment and emotions in the CMU-MOSEI mega corpus. The distribution shows a natural skew towards more frequently used emotions. However, the least frequent emotion, fear, still has 1,900 data points which is an acceptable number for machine learning studies.

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.

Figure 3 shows the distribution of sentiment and emotions in CMU-MOSEI dataset. The distribution 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. 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.

3.5.1 Crowdsourced Annotations

CMU-MOSEI is designed to capture the crowd's perception of a speaker's sentiment and emotions. We rely on minimal training for the annotations to limit the potential bias training may cause. Modeling the crowd's raw perception of sentiment and emotions is vital to creating real-world applications that model the thought processes of the general population. This is in contrast to datasets in the same domain of sentiment and emotion recognition which rely on experts' opinions, which may not agree with the general population's opinion. We prioritize the general population's perception over the psychological definitions of sentiment and emotions, which can only be inferred by experts.

All the monologue sentences in the dataset are annotated using Amazon Mechanical Turk (AMT). Each sentence is annotated thrice by different annotators. Only master annotators with an acceptance rate of over 98% were allowed to annotate the dataset. The following question is asked to the MTurk workers for annotations: "Watch the video clip and rate the sentiment and emotions of the speaker. Please note that you may or may not agree with what the speaker says. It is important that you only rate the sentiment and emotions of the speaker, not yourself."

3.5.2 Sentiment and Emotions Definition

Due to the vast number of similar tasks on AMT, annotators are relatively familiar with broad definition of sentiment and emotions. However, through a five minute training video, we define *emotions* as the speaker's expression of state of mind and feeling while uttering the sentence. *Sentiment* is defined as the speakers attitude towards the topic of his/her discussion. The annotators were asked to annotate sentiment on a seven-step Likert scale of [-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, 1: weakly positive, 2: positive, 3: highly positive]. The Emotions selected are the six basic Ekman emotions (Ekman et al., 1980) of {happiness, sadness, anger, fear, disgust, surprise}. Each of the emotions is annotated at a four-step Likert scale for the presence of an emotion x: [0: no evidence of x, 1: weakly x, 2: x, 3: highly

Figure 4: Annotation user interface for sentiment (left) and emotion (right) labeling.

	CMU-MOSEI Krippendorf alpha
Sentiment	0.53
Happiness	0.41
Anger	0.18
Sadness	0.12
Disgust	0.21
Fear	0.02
Surprise	0.09

Table 3: Agreement Krippendorf alpha values for annotations in the CMU-MOSEI dataset.

x]. The annotators are also asked to determine the speaker's gender.

3.5.3 Annotation User Interface

Figure 4 shows the sample annotation interface that AMT workers see when performing annotations. Each worker must finish watching the training video before starting their annotations. Furthermore, at any time during the annotation session workers can rewatch the training video to refresh their memory for specific instructions.

3.5.4 Annotator Agreement

After all sentences were annotated, we computed Krippendorf's Alpha as a measure of the agreement scores across individual annotators. Table 3 shows these agreement scores. While CMU-MOSEI is annotated by crowdsourced workers in a fairly subjective manner – by asking their opinion about sentiment and emotion of the speaker with minimal training – the overall agreement scores are comparable with other datasets annotated by experts outlined in the submitted paper. Furthermore, a lower agreement would be expected from a dataset such as CMU-MOSEI due to its diversity of topics and speakers, and inherent variance in wild data. These factors impact the agreement since they increase the subjectivity of the task.

3.6 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.

3.7 Modality Alignment

To reach the temporal alignment between different modalities we choose the granularity of the input to be at the level of words. Words are aligned with audio using P2FA (Yuan and Liberman, 2008) to get their utterance times. The visual and acoustic modalities follow the same granularity. We use expected values across the word for visual and acoustic features since they are extracted at a higher frequencies (30 Hz and 100 Hz respectively).

4 Multimodal Data SDK

Dealing with data from multiple sources can be frustrating and intimidating for the community, specifically as the nature of the data from each modality becomes increasingly complex. This problem is compounded by the fact that time series data from multiple modalities must be aligned for best performance (Zadeh et al., 2018b). For the purpose of efficient and reliable data loading, we have built a pipeline in Python that allows users to load any of the multimodal datasets used in this paper. The SDK also allows the users to align the information from multiple sources with various frequencies and returns tensors that are ready to be used in common deep learning frameworks such as TensorFlow (Abadi et al., 2016), Theano (Theano Development Team, 2016) and PyTorch. Additionally, the SDK offers downloading capabilities for each of the datasets (except IEMOCAP which users have to ask for permission from University of Southern California) (Busso et al., 2008). The complete package is available at `https://github.com/A2Zadeh/CMU-MultimodalDataSDK`.

5 Multimodal Fusion Model

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 in multimodal discriminative tasks, many have not developed new insights on how fusion is performed between different modalities.

To understand the fusion process one must first understand the multimodal dynamics (Zadeh et al., 2017). Multimodal 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 multimodal dynamics as a hierarchical process and propose a new fusion model

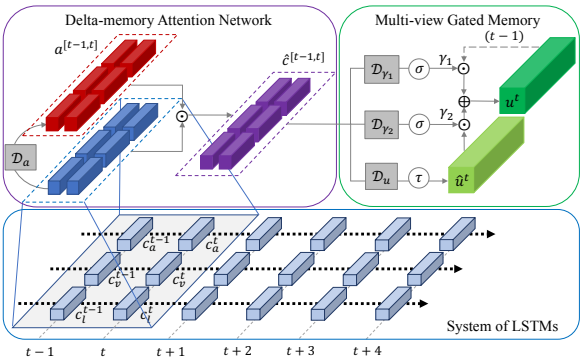

Figure 5: Overview figure of Memory Fusion Network (MFN) (Zadeh et al., 2018a). σ denotes the $sigmoid$ activation function, τ the $tanh$ activation function, $\odot$ the Hadamard product and $\oplus$ element wise addition. Each LSTM encodes information from one view such as language (l), video (v) or audio (a).

called the Dynamic Fusion Graph (DFG). DFG is easily interpretable by analyzing the strengths of 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 (GMFN). 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, GMFN also outperforms previously proposed state-of-the-art models for sentiment analysis and emotion recognition on the CMU-MOSEI. In the following subsections, we first provide a review of the MFN model before a detailed explanation of the Dynamic Fusion Graph and the Graph Memory Fusion Network model.

5.1 Memory Fusion Network

The Memory Fusion Network (MFN) (Zadeh et al., 2018a) 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 special attention mechanism that designed for multimodal fusion by assigning coefficients to highlight cross-modal dynamics. 3) Multi-view Gated Memory is a component that stores the output of multimodal fusion over time. The MFN is shown in Figure 5.

System of LSTMs: The input to GMFN is a

multimodal sequence with the set of M modalities each of and length T. For example sequences can consist of language, video, and audio for $M = \{l, v, a\}$. The input of the mth modality is denoted as: $\mathbf{x}^m = [x_t^m : t \leq T, x_t^m \in \mathbb{R}^{d_{x^m}}]$ where d_{x^m} is the input dimensionality of mth modality $\mathbf{x}_m$.

For each modality, a Long-Short Term Memory (LSTM) encodes the intra-modal interactions over time. For the mth modality, the memory of assigned LSTM is denoted as $\mathbf{c}^m = \{c_t^m : t \leq T, c_t^m \in \mathbb{R}^{d_{c^m}}\}$ and the output is defined as $\mathbf{h}^m = \{h_t^m : t \leq T, h_t^m \in \mathbb{R}^{d_{c^m}}\}$ with d_{c^m} denoting the dimensionality of mth LSTM memory $\mathbf{c}^m$. The following update rules are defined for the mth LSTM (Hochreiter and Schmidhuber, 1997):

$$i_t^m = \sigma(W_i^m\, x_t^m + U_i^m\, h_{t-1}^m + b_i^m) \quad (1)$$
$$f_t^m = \sigma(W_f^m\, x_t^m + U_f^m\, h_{t-1}^m + b_f^m) \quad (2)$$
$$o_t^m = \sigma(W_o^m\, x_t^m + U_o^m\, h_{t-1}^m + b_o^m) \quad (3)$$
$$\hat{c}_t^m = W_{\hat{c}}^m\, x_t^m + U_{\hat{c}}^m\, h_{t-1}^m + b_{\hat{c}}^m \quad (4)$$
$$c_t^m = f_t^m \odot c_{t-1}^m + i_t^m \odot \hat{c}_t^m \quad (5)$$
$$h_t^m = o_t^m \odot tanh(c_t^m) \quad (6)$$

The parameters include the two affine transformations $W_*^m \in \mathbb{R}^{d_{x^m} \times d_{c^m}}$ and $U_*^m \in \mathbb{R}^{d_{c^m} \times d_{c^m}}$. i^m, f^m, o^m are the input, forget and output gates of the mth LSTM respectively, $\hat{c}_t^m$ is the proposed memory update of mth LSTM for time t, $\odot$ denotes the Hadamard product (element-wise product), σ is the sigmoid activation function.

Delta Memory Attention Network: The goal of the Delta Memory Attention Network (DMAN) is to outline the cross-modal interactions at timestep t between different view memories in the system of LSTMs. A coefficient assignment technique is applied on the concatenation of LSTM memories $c_{[t-1,t]}$ at times $t-1$ and t to track changes in modality features. High coefficients are assigned to the dimensions jointly form a cross-modal interaction and low coefficients to the other dimensions. Memories $c_{[t-1,t]}$ are passed to a neural network $\mathcal{D}_a : \mathbb{R}^{2d_c} \mapsto \mathbb{R}^{2d_c}, d_c = \sum_n d_{c_n}$ to obtain the attention coefficients.

$$a_{[t-1,t]} = \mathcal{D}_a(c_{[t-1,t]}) \quad (7)$$

$a_{[t-1,t]}$ are softmax activated scores for each LSTM memory at time $t-1$ and t. The output of the DMAN, $\hat{c}$ is defined as:

$$\hat{c}_{[t-1,t]} = c_{[t-1,t]} \odot a_{[t-1,t]} \quad (8)$$

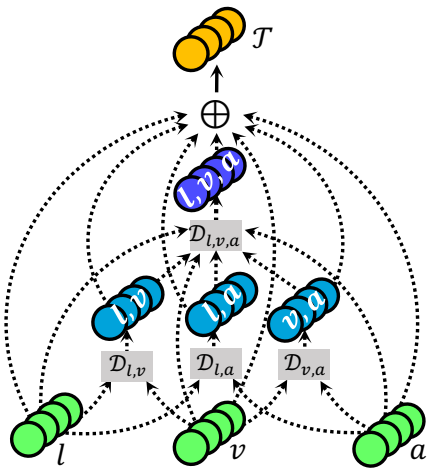

Figure 6: The structure of Dynamic Fusion Graph (DFG) for three modalities of $\{(l)anguage, (v)ision, (a)coustic\}$. Dashed lines in DFG show the dynamic connections between vertices controlled by the efficacies (α).

$\hat{c}_{[t-1,t]}$ are the attended memories of the LSTMs.

Multi-view Gated Memory: u is the neural component that acts as a unifying memory for the system of LSTMs. The output of DMAN, $\hat{c}_{[t-1,t]}$ is directly passed to the Multi-view Gated Memory to signal what dimensions in the system of LSTM memories constitute a cross-view interaction. $\hat{c}_{[t-1,t]}$ is first used as input to a neural network $\mathcal{D}_u : \mathbb{R}^{2\times d_c} \mapsto \mathbb{R}^{d_{mem}}$ to generate a cross-view update proposal $\hat{u}_t$ for u. d_{mem} is the dimensionality of the memory.

$$\hat{u}_t = \mathcal{D}_u(\hat{c}_{[t-1,t]}) \quad (9)$$

This update proposes changes to u based on observations about cross-view interactions at time t.

The Multi-view Gated Memory is controlled using two gates: $\gamma^{(1)}$, $\gamma^{(2)}$ are the retain and update gates respectively. At each timestep t, $\gamma^{(1)}$ assigns how much of the current state of the Multi-view Gated Memory to remember and $\gamma^{(2)}$ assigns how much of the memory to update based on the update proposal $\hat{u}_t$. $\gamma^{(1)}$ and $\gamma^{(2)}$ are each controlled by a neural network. $\mathcal{D}_{\gamma^{(1)}}, \mathcal{D}_{\gamma^{(2)}} : \mathbb{R}^{2\times d_c} \mapsto \mathbb{R}^{d_{mem}}$ control part of the gating mechanism of Multi-view Gated Memory using $\hat{c}_{[t-1,t]}$ as input:

$$\gamma_t^{(1)} = \mathcal{D}_{\gamma^{(1)}}(\hat{c}_{[t-1,t]}),\ \gamma_t^{(2)} = \mathcal{D}_{\gamma^{(2)}}(\hat{c}_{[t-1,t]}) \quad (10)$$

At each time-step of GMFN recursion, u is updated using $\gamma^{(1)}$ and $\gamma^{(2)}$, as well as the current update proposal $\hat{u}_t$ with the following formulation:

$$u_t = \gamma_t^{(1)} \odot u_{t-1} + \gamma_t^{(2)} \odot tanh(\hat{u}_t) \quad (11)$$

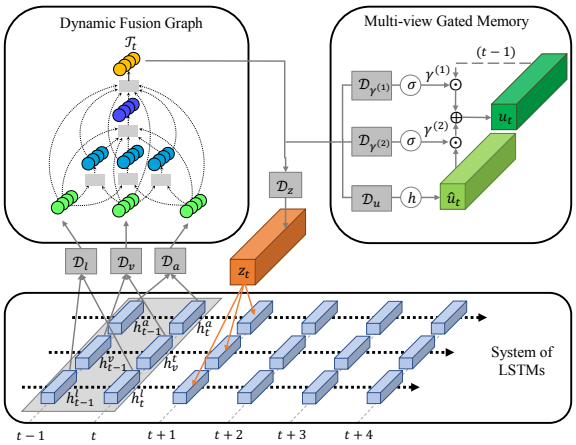

Figure 7: The overview of Graph Memory Fusion Network (GMFN) pipeline. GMFN replaces the fusion block in MFN with a Dynamic Fusion Graph (DFG).

$\hat{u}_t$ is activated using $tanh$ squashing function to improve model stability by avoiding drastic changes to the memory.

5.2 Dynamic Fusion Graph

In this section we discuss the internal structure of the proposed Dynamic Fusion Graph (DFG) neural model. DFG in illustrated in Figure 6. DFG has the following properties that makes it suitable for multimodal fusion: (1) it explicitly models the multimodal interactions by capturing unimodal, bimodal and trimodal interactions, (2) it does so with an efficient number of parameters (as opposed to previous approaches such as Tensor Fusion (Zadeh et al., 2017)), and (3) it can dynamically alter its structure and choose the ideal fusion graph based on the importance of individual multimodal dynamics during fusion.

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 are denoted as $\{l, v\}, \{v, a\}, \{l, a\}$, and the trimodal dynamics are denoted as $\{l, v, a\}$. These dynamics are in the form of latent representations that have been learnt from our multimodal data. Each of these latent representations are considered as vertices inside a graph $G = (V, E)$ with V denoting the set of vertices and E denoting 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}. The set of all edges E consists of all such e_{ij} satisfying this definition.

We define an efficacy for each edge e_{ij} as as weight α_{ij} that determines the importance of that edge in multimodal fusion. Specifically, each α is a sigmoid activated probability neuron which indicates how strong or weak the connection is between v_i and v_j. During fusion, stronger interactions will be emphasized more while the effect of weaker interactions will be minimized. This set of efficacies α 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 optimal use of efficacies. For each multimodal input, different edges will be activated depending on the strength of each interaction. As a result, the DFG is able to dynamically control the structure of the graph during multimodal fusion.

With the singleton vertices in V (l, v, and a), efficacies for each edge α_{ij} and the graph G in place, we can now detail the steps for multimodal fusion. Multimodal fusion happens in a hierarchical manner. Each vertex v_i is multiplied by α_{ij} before being used as input to $\mathcal{D}_j$, a neural network that performs local fusion of several weighted vertices. In total, $\mathcal{D}_j$ takes as input all v_i that satisfy the neural connection formula: $v_i \subset v_j$. The output of $\mathcal{D}_j$ is the result of local multimodal fusion and is passed to the resulting vertex v_j. This process is repeated in a hierarchical manner until all cross-modal interactions across unimodal, bimodal, and trimodal interactions are discovered.

The overall structure of the vertices, edges and respective efficacies is shown in Figure 6. There are a total of 8 vertices (including the output vertex), 19 edges, and therefore 19 efficacies. Singleton vertices l, v, and a are the inputs to the DFG. All vertices are connected to the output vertex $\mathcal{T}_t$ of the network via edges scaled by their respective efficacy. The final output of the DFG is the output vertex $\mathcal{T}_t$, which is a summarization of multimodal interactions up to time t.

5.3 Graph Memory Fusion Network

To test the performance of DFG, we use a recurrent architecture similar to MFN. We replace the Delta-memory Attention Network with DFG and refer to the modified model as Graph Memory Fu-

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$^{\#}$
GMFN	**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**
Δ_{SOTA}	↑ **0.5**	↑ **0.6**	↑ **0.3**	↑ **0.3**	↓ **0.01**	↑ **0.02**	↑ **2.1**	↑ **0.8**	↑ **2.1**	↑ **3.4**	↓ **2.2**	**0.0**	↑ **4.8**	↑ **4.9**	↓ **1.6**	↓ **2.3**	↓ **0.6**	↑ **0.1**

Table 4: Results for multimodal sentiment analysis and multimodal emotion recognition on the newly introduced MOSEI dataset. GMFN sets new state of the art results for multiple metrics. SOTA1 and SOTA2 refer to the previous best and second best state-of-the-art respectively. Symbols depict the model which the baseline result came from: # *MFN*, ■ *MARN*, ⋆ *TFN*, ⋄ *MV-LSTM*, § *EF-LSTM*, † *BC-LSTM*, ♣ *C-MKL*, ♭ *DF*, ♡ *SVM*, • *RF*, ∪ *CNN-LSTM*, *RNTN*, ×: *DynamicCNN*, ▷ *DAN*, ≀ *DHN*, ◁ *RHN*, △: *Adieu-Net*, ∩: *SER-LSTM*. For a detailed table with all baseline results, please refer to the supplementary material. The best results are highlighted in bold and Δ_{SOTA} shows the performance improvement of GMFN over SOTA1, previous state-of-the-art. Improvements are highlighted in green. The GMFN outperforms the SOTA across all datasets and metrics, except the Δ_{SOTA} entries highlighted in gray. These results show that sentiment prediction and emotion recognition on the CMU-MOSEI dataset is non-trivial.

sion Network (GMFN). Figure 7 shows the overall architecture of the GMFN.

Similar to MFN, GMFN 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. This allows the model 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 $\hat{u}_t$. $\gamma^{(1)}$ and $\gamma^{(2)}$ are the Multi-view Gated Memory's retain and update gates respectively and are learned using networks $D_{\gamma^{(1)}}$ and $D_{\gamma^{(2)}}$. Finally, a network D_z transforms $\mathcal{T}_t$ into a multimodal representation z_t to update the system of LSTMs in the hybrid manner as described in (Zadeh et al., 2018b): using a system of Long Short-term Hybrid Memory Networks.

The outputs of the GMFN are the final state of the Multi-view Gated Memory u_T and the outputs of each of the m LSTMs:

$$\mathbf{h}_T = \bigoplus_{m \in M} h^m_T$$

representing individual sequence information. $\oplus$ denotes vector concatenation. This output is subsequently connected to a classification or regression layer for final prediction (for sentiment and emotion recognition).

6 Experiments and Discussion

6.1 Results on MOSEI

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

Table 4 shows the results on CMU-MOSEI. Accuracy is reported as A^C where C is the number of 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. For all metrics, higher values indicate better performance except for MAE where lower values indicate better performance. GMFN shows superior performance in sentiment analysis and competitive performance

Dataset	CMU-MOSI					ICT-MMMO		YouTube		MOUD		IEMOCAP Emotions							
Task	Sentiment					Sentiment		Sentiment		Sentiment		Happy		Sad		Angry		Neutral	
Metric	A^2	F1	A^7	MAE	r	A^2	F1	A^3	F1	A^2	F1	A^2	F1	A^2	F1	A^2	F1	A^2	F1
SOTA2	77.1$^{\blacksquare}$	77.0$^{\blacksquare}$	34.1$^{\#}$	0.97$^{\blacksquare}$	0.63$^{\#}$	72.5*	72.6*	48.3$^{\blacksquare}$	45.0$^{\#}$	81.1$^{\#}$	81.0$^{\#}$	86.0§	83.6*	83.2†	81.7†	83.5†	84.3$^{\diamond}$	67.5*	66.7$^{\diamond}$
SOTA1	77.4$^{\#}$	77.3$^{\#}$	34.7$^{\blacksquare}$	0.97$^{\#}$	0.63§	73.8$^{\#}$	73.1$^{\#}$	51.7$^{\#}$	51.6$^{\#}$	81.1$^{\blacksquare}$	81.2$^{\blacksquare}$	86.0$^{\flat}$	**84.2**§	83.4*	82.8*	85.2§	84.5§	68.8§	68.5§
GMFN no W	71.3	71.3	27.0	1.14	0.54	71.3	71.1	45.0	40.2	71.3	71.1	85.6	79.0	80.8	78.2	82.6	81.3	66.0	63.8
GMFN no E	73.2	73.0	33.2	1.04	0.60	67.5	66.7	48.3	48.2	67.5	66.7	86.0	81.6	82.8	81.4	83.2	81.5	67.6	66.2
GMFN no G	75.5	75.5	30.5	1.00	0.61	68.8	68.9	46.7	46.9	77.4	77.0	86.5	84.0	83.2	82.1	84.0	83.0	66.6	66.0
GMFN no M	63.4	63.4	23.5	1.31	0.41	58.8	59.5	45.0	35.0	70.8	66.9	85.6	83.0	81.3	79.1	82.0	79.3	66.2	66.2
GMFN no S	75.1	75.0	33.7	0.998	0.634	72.5	71.3	53.3	45.8	81.1	80.5	86.4	82.6	82.6	81.2	83.9	83.4	68.4	66.8
GMFN	**77.7**	**77.7**	**35.6**	**0.96**	**0.66**	**76.3**	**76.2**	**55.0**	**53.5**	**81.1**	**80.9**	**86.8**	**84.2**	**83.8**	**83.0**	**85.8**	**85.5**	**69.4**	**68.9**
Δ_{SOTA}	↑ **0.3**	↑ **0.4**	↑ **0.9**	↓ **0.01**	↑ **0.03**	↑ **2.5**	↑ **3.1**	↑ **9.2**	↑ **8.4**	↑ **8.5**	↑ **8.0**	↑ **0.8**	**0.0**	↑ **0.4**	↑ **0.2**	↑ **0.6**	↑ **1.0**	↑ **0.6**	↑ **0.4**

Dataset	POM Speaker Personality Traits															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	A^7	A^7	A^7	A^7	A^7	A^7	A^7	A^7	A^5	A^5	A^5	A^5	A^5	A^5	A^7	A^5
SOTA2	26.6$^{\bullet}$	27.6§	32.0$^{\heartsuit}$	35.0$^{\heartsuit}$	26.1$^{\flat}$	32.0$^{\flat}$	27.6*	29.6$^{\flat}$	34.0$^{\heartsuit}$	53.2$^{\bullet}$	49.8$^{\heartsuit}$	39.4$^{\flat}$	42.4§	42.4$^{\flat}$	27.6*	36.5†
SOTA1	26.6$^{\bullet}$	31.0*	33.0$^{\flat}$	35.0$^{\heartsuit}$	27.6†	36.5†	30.5†	31.5$^{\heartsuit}$	34.0$^{\heartsuit}$	53.7$^{\flat}$	50.7$^{\diamond}$	42.9$^{\heartsuit}$	45.8†	42.4$^{\flat}$	28.1$^{\heartsuit}$	40.4$^{\bullet}$
GMFN no W	30.0	31.5	33.5	36.0	29.6	31.0	27.6	35.5	31.0	53.2	50.7	39.9	41.9	42.4	28.1	40.9
GMFN no E	26.1	31.5	33.5	33.0	28.6	31.0	27.1	29.6	31.0	53.2	50.7	39.9	41.9	42.4	27.1	40.9
GMFN no G	29.1	31.5	33.5	33.5	29.6	34.0	31.0	30.5	31.5	53.2	50.7	42.9	42.9	43.8	31.5	40.9
GMFN no M	24.1	31.5	28.6	27.6	27.1	27.6	26.6	32.0	31.0	34.5	37.4	39.9	41.4	42.4	23.6	40.9
GMFN no S	32.5	31.5	34.0	36.9	30.5	32.5	34.5	35.0	33.5	53.7	50.7	42.4	45.8	45.3	31.5	42.4
GMFN	**35.5**	**35.0**	**36.5**	**41.9**	**32.0**	**36.9**	**37.4**	**36.9**	**37.4**	**55.2**	**53.7**	**45.8**	**48.3**	**46.3**	**34.5**	**43.8**
Δ_{SOTA}	↑ **8.9**	↑ **4.0**	↑ **3.5**	↑ **6.9**	↑ **4.4**	↑ **0.4**	↑ **6.9**	↑ **5.4**	↑ **3.4**	↑ **1.5**	↑ **3.0**	↑ **2.9**	↑ **2.5**	↑ **3.9**	↑ **6.4**	↑ **3.4**

Table 5: Results for multimodal sentiment analysis on the CMU-MOSI, ICT-MMMO, YouTube and MOUD datasets, multimodal emotion recognition on the IEMOCAP dataset and multimodal speaker traits recognition on the POM dataset. SOTA1 and SOTA2 refer to the previous best and second best state-of-the-art respectively. For a detailed table with all baseline results, please refer to the supplementary material. The best results are highlighted in bold and Δ_{SOTA} shows the performance improvement of GMFN over SOTA1, previous state-of-the-art. Improvements are highlighted in green. GMFN achieves excellent results and is an effective method for multimodal fusion.

in emotion recognition. Therefore, GMFN is an effective method for multimodal fusion.

6.2 Results on External Datasets: Multimodal Sentiment Analysis

We achieve state-of-the-art performance with improvement over all the comparison metrics for three additional English sentiment analysis datasets: CMU-MOSI, ICT-MMMO and MOUD. Table 5 shows the comparison of our GMFN with state-of-the-art approaches for these three dataset. To assess the generalization of the GMFN to speakers communicating in different languages, we also compare with state-of-the-art approaches for sentiment analysis on MOUD, with opinion utterance video clips in Spanish. The final quarter of Table 5 shows these results where we also achieve improvement over state-of-the-art approaches.

6.3 Results on External Datasets: Multimodal Emotion Recognition

Our results for multimodal emotion recognition on IEMOCAP dataset are reported in the bottom half of Table 5. Our approach achieves state-of-the-art performance in emotion recognition: both emotion classification as well as continuous emotion regression except for the case of correlation in dominance which our results are competitive but not state of the art.

6.4 Results on External Datasets: Multimodal Speaker Traits Recognition

We experiment on speaker traits recognition based on observed multimodal communicative behaviors. Table 5 shows the performance of the GMFN on POM dataset, where it achieves state-of-the-art accuracies on 16 speaker trait recognition tasks including confidence, persuasiveness and credibility.

6.5 Subcomponent Ablations

To demonstrate the effects of each component on the overall performance of GMFN, we perform a series of ablation studies which remove individual components. The ablation baselines include the following:

RDFN no W uses only the outputs of LSTMs at time t, instead of both t and $t-1$. This assesses the importance of a convolutional window to track changes in multimodal features for effective fusion.

GMFN no E sets all the efficacies to 1. As a result all intermediate fusion features are therefore assigned the same importance. This effectively

removes the dynamic capability of DFG.

GMFN no G completely removes the DFG and performs multimodal fusion by a simple concatenation of LSTM outputs h_t^l, h_t^a and h_t^v.

RDFN no M removes the MSM component which forces each LSTM memory to carry intermodal dynamics individually.

RDFN no S does not perform synchronization of LSTMs at all. This baseline is simply 3 LSTMs in parallel that encodes the information from each of the 3 modalities before concatenating the features before the final layer.

Table 5 shows the results of these ablation studies on multimodal tasks using the CMU-MOSI, ICT-MMMO, MOUD, IEMOCAP and POM. We observe that each component of our model is indeed necessary for best performance across all datasets.

6.6 Interpretability of Fusion

To better understand the internal fusion mechanism between modalities, we visualize the behavior of the learned DFG efficacies in Figure 8 for various cases (dark red denotes high efficacies and dark blue denotes low efficacies).

Multimodal Fusion is Volatile in Nature: The first observation is that the structure of the DFG is changing across videos and for each video, across 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, implying 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. Consequently the model switches its focus on interactions involving the language and acoustic modalities.

Priors of Multimodal Fusion: Certain efficacies remain unchanged across cases and across time. 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 from the language and acoustic modalities. 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. We believe that these can represent priors from Human Multimodal Language that DFG learns from the diverse videos of human communication in the CMU-MOSEI dataset.

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).

Efficacies to Terminal Vertex: 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 the features from a single modality. Instead, the efficacies from the combined language and acoustic modalities as well as the combined language and visual modalities tend to be higher.

Priors of Human Communication: DFG always prefers to perform fusion between language and acoustic 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 due to the close relationships between text and speech (Yuan and Liberman, 2008). In these cases, we believe DFG has learned some natural priors of human communication between the language and acoustic modalities.

With these observations, we believe that DFG has successfully learned how to dynamically modify its internal structure in order to best model human communication. Not only can it perform dynamic fusion depending on the multimodal input, it has also learned general priors on multimodal fusion and human multimodal language.

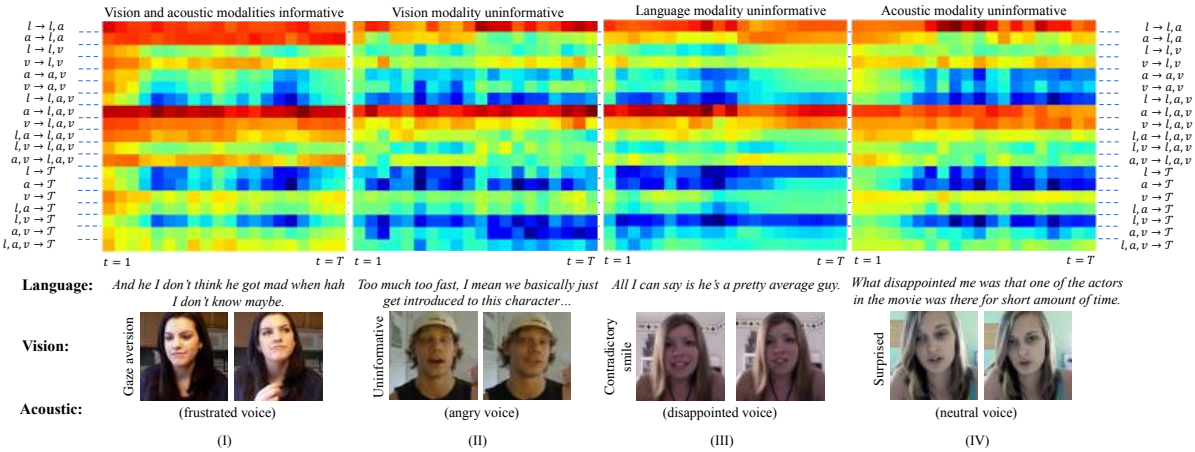

Figure 8: Visualization of DFG efficacies across time. Dark red denotes high efficacies and dark blue denotes low efficacies. The efficacies (thus the DFG structure) change over time as DFG is exposed to new information. DFG is able choose which multimodal dynamics to rely on. It also learns priors about human communication since certain efficacies (thus edges in DFG) remain unchanged across time and across data points.

7 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. We analyzed the structure of multimodal fusion in sentiment analysis and emotion recognition using a novel interpretable fusion mechanism, the Graph Memory Fusion Network (GMFN), which combines Dynamic Fusion Graph with the Memory Fusion Network. The interpretable nature of the hierarchical fusion in DFG gave us the opportunity to investigate the behavior of modalities during fusion. We found that the DFG is able to dynamically select important modalities during fusion and was able to learn interesting priors in human multimodal language. Furthermore, GMFN showed superior performance in multimodal sentiment analysis and emotion recognition, demonstrating that GMFN is highly suitable for human multimodal language analysis. We believe that the CMU-MOSEI dataset and GMFN fusion model will significantly expand the horizons of NLP and encourage further research in human multimodal language analysis.

8 Acknowledgements

I collaborated with Amir Zadeh throughout this project. The original idea of building MOSEI came from Amir and Prof. Morency. Prof. Morency and Amir advised throughout the project. The data crawling mechanism was implemented by Edmund Tong. The annotation UI was built by Edmund and Jonathan Vanbriessen. The MTURK annotation was performed by Jonathan except the last batch which was performed by me. Multiple discussions on data collection and annotation strategies were held between me, Amir, Edmund and Jonathan. The Multimodal SDK was built by Prateek Vij. I was the first person to use the SDK to load data. I debugged it, added features and made version 2.0 which is currently being maintained with help from Zhun Liu and Amir. Upon collection of the data, I performed all the feature extraction, except visual features (FACET) which were extracted by Amir. I loaded and cleaned all the data. I performed all the data analysis described in this paper and collected all the statistics. The idea for the DFG model came from Amir and it was my idea to incorporate DFG with the MFN model. I implemented the MFN model, the GMFN model and all the baselines. I ran all the experiments in the paper and analyzed all the results, including the ablation studies and baselines. Amir and Prof. Morency advised on the experiments and analysis. I made all the tables and the figure on interpretation of multimodal fusion. The DFG and GMFN figures were made by Amir. I wrote all sections of the DAP report.

9 Appendix: CMU-MOSEI Full Results

The full set of results using various approaches on CMU-MOSEI is divided into multimodal and

unimodal and shown in Tables 6 and 7 respectively.

10 Appendix: Full Results on Other Datasets

The full set of results for the GMFN and baselines across all datasets are presented in Tables 8, 9, 10 and 11.

Dataset	MOSEI																		MOSEI Natural									
Task	Sentiment						Anger		Disgust		Fear		Happiness		Sadness		Surprise		Sentiment				Anger		Happiness		Sadness	
Method	A^2	F1	A^5	A^7	MAE	Corr	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	MAE	Corr	WA^2	F1	WA^2	F1	WA^2	F1
MULTIMODAL																												
RF	60.2	56.5	39.6	38.4	-	-	51.0	72.0	50.9	73.2	50.1	89.9	57.0	48.6	50.6	61.8	50.1	85.4	54.6	48.6	-	-	50.5	75.7	58.8	48.5	49.6	57.2
SVM-MD	73.9	73.6	39.9	39.9	0.77	0.46	56.0	33.9	57.1	42.3	60.0	61.8	50.3	36.9	53.3	27.7	49.7	61.4	50.9	56.5	0.81	0.24	53.5	41.7	51.3	50.5	54.8	35.7
DF	74.0	72.5	43.6	43.5	0.72	0.51	49.7	15.2	67.0	71.4	54.0	26.2	52.2	41.7	55.2	47.4	51.0	74.4	57.7	75.4	0.76	0.44	**58.7**	54.4	55.5	57.6	52.0	33.0
EF-LSTM	73.0	72.7	41.9	41.7	0.76	0.46	50.4	11.5	58.5	49.4	47.3	68.5	55.1	49.0	51.5	48.9	48.7	75.0	55.2	73.4	0.76	0.40	57.5	61.5	54.0	55.9	50.6	17.6
EF-SLSTM	74.4	74.3	43.0	42.9	0.73	0.48	50.1	54.2	57.2	57.5	52.9	77.8	57.8	50.7	50.0	51.6	50.2	51.4	55.5	74.5	0.75	0.41	49.8	17.1	54.9	57.1	50.7	20.2
EF-BLSTM	71.6	74.1	42.1	42.0	0.75	0.46	51.5	48.1	62.8	62.3	56.7	56.2	57.8	54.6	59.2	57.3	47.2	49.7	51.4	71.0	0.80	0.29	53.4	46.7	52.8	53.8	48.1	53.3
EF-SBLSTM	73.2	72.5	43.3	43.0	0.74	0.47	50.1	7.5	48.4	12.8	50.6	37.7	51.1	39.2	50.1	17.0	50.5	59.7	57.5	74.2	0.78	0.38	56.4	60.8	53.7	54.8	**59.9**	53.5
EF-RHN	73.5	73.3	42.4	42.4	0.78	0.44	52.7	46.1	56.0	48.8	49.6	56.6	51.9	42.2	52.6	45.8	49.7	57.0	50.0	69.6	0.83	0.21	46.6	58.8	50.2	48.3	49.5	59.1
MV-LSTM	76.4	76.4	43.8	43.5	0.72	0.52	56.0	39.0	38.3	58.3	47.9	36.0	50.3	36.7	55.6	40.1	49.8	74.0	57.4	75.5	0.74	0.45	50.0	71.7	50.0	48.0	50.6	36.7
BC-LSTM	75.8	75.5	44.7	44.6	0.72	0.51	50.4	71.4	47.3	42.0	50.7	10.9	51.4	51.4	53.9	38.4	50.1	51.2	62.6	78.6	**0.72**	**0.49**	50.0	4.9	50.0	48.0	50.0	14.0
TFN	74.8	75.4	44.8	44.7	0.72	0.52	60.5	55.7	64.6	59.8	54.4	65.8	66.5	66.6	58.9	51.2	52.2	56.2	66.8	79.6	0.78	0.46	53.8	**75.6**	54.3	56.2	49.6	41.9
MARN	75.9	75.8	43.3	43.2	0.73	0.51	50.0	6.7	50.0	35.8	50.0	11.9	50.0	71.0	50.0	35.8	50.0	60.8	-	-	-	-	-	-	-	-	-	-
MFN	76.0	76.0	43.3	43.2	0.72	0.52	50.0	64.1	65.2	65.1	50.2	60.8	51.2	52.8	51.5	41.4	53.3	62.2	-	-	-	-	-	-	-	-	-	-
GMFN	**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**	**67.4**	**80.1**	**0.72**	**0.49**	**58.7**	75.4	**63.2**	**66.9**	56.1	**61.3**
Δ_{SOTA}	↑ 0.5	↑ 0.6	↑ 0.3	↑ 0.3	↓ 0.01	↑ 0.02	↑ 2.1	↑ 0.8	↑ 2.1	↑ 3.4	↓ 2.2	0.0	↑ 4.8	↑ 4.9	↓ 1.6	↓ 2.3	↓ 0.6	↑ 0.1	↑ 0.6	↑ 0.5	0.00	0.00	0.0	↓ 0.2	↑ 4.4	↑ 9.3	↓ 3.8	↑ 2.2

Table 6: Results for sentiment analysis and emotion recognition on the MOSEI Overall and Natural split. GMFN sets new state of the art on all the datasets. The best results are highlighted in bold and Δ_{SOTA} shows the performance improvement of GMFN over previous state-of-the-art. For all metrics, higher values indicate better performance except for MAE where lower values indicate better performance. Improvements are highlighted in green. The GMFN outperforms the SOTA across all datasets and metrics, except the Δ_{SOTA} entries highlighted in gray. These results show that sentiment prediction and emotion recognition on the CMU-MOSEI Mega Corpus is non-trivial.

Dataset	MOSEI																		MOSEI Natural									
Task	Sentiment						Anger		Disgust		Fear		Happiness		Sadness		Surprise		Sentiment				Anger		Happiness		Sadness	
Method	A^2	F1	A^5	A^7	MAE	Corr	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	WA^2	F1	MAE	Corr	WA^2	F1	WA^2	F1	WA^2	F1
LANGUAGE																												
RF	58.0	54.4	38.0	37.9	-	-	50.8	71.8	50.2	72.6	50.0	89.8	52.6	42.4	50.1	61.2	50.0	85.3	57.2	49.8	-	-	51.7	75.8	56.5	45.0	49.5	56.7
SVM	72.8	73.0	42.5	42.5	0.761	0.426	55.8	54.7	58.5	60.1	54.4	76.2	50.0	35.8	50.0	11.9	52.8	69.5	51.1	54.6	0.759	0.417	58.1	47.4	50.0	47.7	50.0	14.7
DAN	74.3	74.1	43.0	42.9	0.743	0.464	49.6	38.4	64.0	67.1	53.6	20.8	53.0	44.1	51.2	35.0	54.3	70.0	62.1	79.1	0.744	0.389	56.0	67.3	52.6	53.0	51.4	30.2
DHN	72.5	71.6	43.1	42.9	0.749	0.460	56.6	52.7	55.1	36.6	54.7	66.7	50.9	38.9	53.8	49.9	54.1	68.8	60.4	78.3	0.739	0.409	56.4	48.6	52.9	52.9	52.9	38.2
LSTM	73.2	73.0	42.8	42.7	0.745	0.469	53.6	40.9	57.3	55.2	50.4	79.7	52.0	41.8	54.0	42.8	52.0	62.1	58.0	77.0	0.749	0.389	54.1	43.4	52.6	53.5	54.6	43.6
SLSTM	73.7	73.6	43.2	42.9	0.745	0.461	52.1	34.1	59.0	59.0	52.7	37.4	53.5	45.4	50.5	25.6	51.6	66.7	58.8	78.1	0.742	0.410	54.3	57.0	52.7	53.1	49.1	37.2
BLSTM	73.6	73.6	42.6	42.3	0.748	0.466	52.1	39.6	54.6	64.4	56.2	56.2	54.0	47.0	52.4	47.4	50.1	49.1	62.5	78.4	0.766	0.388	51.9	49.8	52.3	52.9	50.1	36.9
SBLSTM	74.1	74.1	43.2	43.2	0.746	0.456	52.4	33.9	57.4	54.4	46.5	56.0	52.8	44.7	50.8	30.7	50.2	31.5	59.8	77.7	0.763	0.380	55.8	55.4	51.3	51.2	50.9	37.2
CNN-LSTM	73.8	73.5	42.4	42.4	0.758	0.458	56.0	68.2	50.0	51.2	54.5	57.8	50.0	35.8	50.0	49.7	50.0	39.4	50.0	71.1	0.786	0.291	50.0	5.3	50.0	47.8	50.0	14.7
DynamicCNN	72.6	72.4	42.8	42.8	0.763	0.432	49.9	71.0	50.0	6.1	58.8	45.3	50.0	35.8	49.8	12.3	53.2	64.6	52.4	77.0	0.787	0.328	50.0	5.3	50.0	47.8	50.0	14.7
RHN	63.3	53.4	41.5	41.5	0.825	0.252	50.0	11.9	54.1	67.2	51.1	80.6	50.0	35.8	50.2	57.0	50.1	2.6	50.0	71.1	0.831	0.160	57.0	53.3	50.0	47.7	50.1	57.5
VISUAL																												
RF	57.8	54.0	39.1	39.4	-	-	50.1	71.0	50.0	72.4	50.1	89.8	57.4	49.3	50.0	60.8	50.0	85.4	53.3	43.5	-	-	49.8	73.9	62.1	55.3	49.6	58.8
SVM	72.6	72.5	40.7	40.7	0.791	0.414	53.0	22.7	54.4	26.1	64.2	62.3	50.4	36.6	52.7	25.8	51.3	24.4	49.4	52.8	0.773	0.155	50.5	23.6	50.6	47.5	47.9	26.7
DAN	73.9	73.7	42.5	42.5	0.780	0.385	54.9	44.0	40.1	58.6	44.6	42.7	50.8	39.5	54.3	42.0	46.7	70.2	53.2	73.3	0.767	0.282	53.0	16.9	50.0	46.0	50.2	23.6
DHN	73.4	73.1	42.7	42.7	0.777	0.429	54.4	26.7	60.3	46.0	49.9	2.3	53.4	47.7	52.8	35.3	50.7	14.2	49.5	70.1	0.783	0.173	51.1	69.4	50.0	46.0	52.8	38.0
LSTM	73.2	73.1	41.9	41.9	0.785	0.381	46.5	64.6	32.8	45.7	51.3	78.4	50.5	38.5	54.4	46.5	51.4	81.3	53.1	76.8	0.785	0.187	51.7	42.8	55.6	56.4	52.5	34.9
SLSTM	72.8	72.8	41.6	41.6	0.784	0.398	52.3	19.5	34.2	47.7	49.8	5.0	51.0	39.5	55.0	33.1	49.4	84.2	51.4	75.0	0.785	0.219	49.3	50.7	54.6	54.4	50.9	42.2
BLSTM	73.0	72.9	41.8	41.8	0.782	0.395	58.4	46.4	51.2	14.1	49.6	24.4	51.0	40.8	57.7	44.6	47.5	77.5	55.1	76.7	0.787	0.180	47.6	41.0	54.8	55.2	52.8	42.6
SBLSTM	73.8	73.5	42.4	42.4	0.776	0.406	60.0	56.3	43.6	65.7	50.0	5.3	50.3	37.1	54.0	29.3	51.8	19.0	52.4	76.3	0.748	0.260	50.1	32.5	52.9	51.4	54.0	39.6
RHN	73.2	72.6	42.2	42.2	0.784	0.400	50.0	6.7	49.0	71.5	49.9	1.0	50.0	35.8	49.4	61.5	50.4	42.5	50.0	70.3	0.791	0.146	50.0	5.4	50.0	46.0	50.0	13.0
CNN-LSTM	68.8	63.4	35.6	35.6	0.857	0.072	50.0	5.1	50.0	79.0	50.0	85.4	50.0	43.8	50.0	10.5	50.0	82.8	60.3	78.2	1.001	0.157	60.2	76.6	50.0	59.2	51.4	31.7
3DCNN	71.0	64.1	39.5	39.5	0.790	0.194	50.1	52.7	50.8	51.2	51.1	66.5	50.0	43.8	48.8	41.2	51.7	46.2	60.7	68.8	1.251	0.166	63.1	51.5	43.9	55.2	38.1	26.9
ACOUSTIC																												
RF	58.1	54.3	39.4	39.4	-	-	51.0	71.9	50.0	72.4	50.0	89.8	53.6	43.8	50.3	61.3	50.0	85.4	53.3	42.0	-	-	50.9	77.5	52.1	38.3	51.1	59.1
SVM	73.9	73.6	34.6	34.6	0.794	0.421	54.1	31.3	58.4	41.5	47.4	36.9	50.0	35.8	50.0	12.1	52.2	38.2	50.3	57.0	0.843	0.100	55.7	37.7	50.0	49.7	50.1	14.3
DAN	74.2	73.8	42.1	42.1	0.777	0.428	55.5	43.2	58.9	44.8	58.5	72.9	50.1	36.0	54.8	38.3	47.7	66.5	49.9	72.7	0.843	0.164	50.0	4.4	50.0	49.7	49.9	57.4
DHN	74.2	73.9	41.6	41.6	0.780	0.422	53.2	24.0	47.4	69.9	46.9	42.7	50.0	35.9	53.8	33.5	48.6	69.1	50.7	73.5	0.829	0.171	53.9	27.0	50.0	49.7	54.9	48.3
Adieu-Net	74.2	73.8	42.1	42.1	0.779	0.426	56.4	51.8	54.9	31.1	48.2	24.6	50.0	35.8	53.7	31.9	50.3	34.2	50.0	72.9	0.840	0.129	44.9	51.3	50.0	49.7	49.8	40.9
SER-LSTM	67.1	66.4	46.3	46.3	0.713	0.348	49.9	5.5	53.4	29.1	44.9	43.2	50.0	31.4	57.8	53.1	48.2	50.7	50.5	70.7	0.841	0.069	50.0	5.6	50.0	48.9	50.0	15.1
LSTM	74.0	73.6	41.6	41.6	0.779	0.426	55.2	37.6	60.9	47.8	59.9	66.5	50.1	36.6	54.2	47.8	50.6	8.0	50.6	73.4	0.842	0.121	50.8	21.0	51.2	52.9	52.8	46.6
SLSTM	73.9	73.6	41.2	41.2	0.782	0.427	50.9	11.1	53.9	24.8	62.3	60.0	50.3	36.8	58.8	52.0	53.8	60.9	50.4	73.2	0.832	0.134	54.7	49.6	51.2	52.9	51.5	56.9
BLSTM	73.6	73.3	41.7	41.7	0.783	0.419	52.0	16.0	53.0	21.8	62.7	77.0	50.3	36.6	55.3	36.0	47.6	73.7	49.9	72.7	0.831	0.128	52.0	64.8	50.6	51.2	53.0	50.1
SBLSTM	73.9	73.6	41.2	41.2	0.782	0.427	53.1	34.9	39.5	57.4	53.9	77.7	61.5	61.4	50.5	16.1	51.7	30.3	49.9	72.7	0.837	0.126	53.6	48.3	49.9	49.7	50.7	53.6
RHN	73.1	72.8	41.5	41.5	0.787	0.408	55.5	47.8	70.3	75.9	50.0	89.8	50.0	35.8	58.9	65.9	54.3	72.4	50.0	72.8	0.857	0.149	50.0	76.8	50.0	49.7	50.0	14.0

Table 7: Results for sentiment analysis and emotion recognition on the MOSEI Overall and Natural split using unimodal language, visual and acoustic methods.

Dataset	CMU-MOSI				
Task	Sentiment				
Method	A^2	F1	A^7	MAE	Corr
Majority	50.2	50.1	17.5	1.864	0.057
RF	56.4	56.3	21.3	-	-
SVM-MD	71.6	72.3	26.5	1.100	0.559
THMM	50.7	45.4	17.8	-	-
SAL-CNN	73.0	-	-	-	-
C-MKL	72.3	72.0	30.2	-	-
EF-HCRF	65.3	65.4	24.6	-	-
EF-LDHCRF	64.0	64.0	24.6	-	-
MV-HCRF	44.8	27.7	22.6	-	-
MV-LDHCRF	64.0	64.0	24.6	-	-
CMV-HCRF	44.8	27.7	22.3	-	-
CMV-LDHCRF	63.6	63.6	24.6	-	-
EF-HSSHCRF	63.3	63.4	24.6	-	-
MV-HSSHCRF	65.6	65.7	24.6	-	-
DF	72.3	72.1	26.8	1.143	0.518
EF-LSTM	74.3	74.3	32.4	1.023	0.622
EF-SLSTM	72.7	72.8	29.3	1.081	0.600
EF-BLSTM	72.0	72.0	28.9	1.080	0.577
EF-SBLSTM	73.3	73.2	26.8	1.037	0.619
MV-LSTM	73.9	74.0	33.2	1.019	0.601
BC-LSTM	73.9	73.9	28.7	1.079	0.581
TFN	74.6	74.5	28.7	1.040	0.587
GME-LSTM(A)	76.5	73.4	-	**0.955**	-
MARN	77.1	77.0	34.7	0.968	0.625
MFN	77.4	77.3	34.1	0.965	0.632
GMFN no W	71.3	71.3	27.0	1.143	0.537
GMFN no E	73.2	73.0	33.2	1.040	0.603
GMFN no G	75.5	75.5	30.5	0.999	0.610
GMFN no M	63.4	63.4	23.5	1.312	0.408
GMFN no S	75.1	75.0	33.7	0.998	0.634
GMFN	**77.7**	**77.7**	**35.6**	**0.955**	**0.664**
Δ_{SOTA}	↑ **0.3**	↑ **0.4**	↑ **0.9**	**0.0**	↑ **0.03**
Human	85.7	87.5	53.9	0.710	0.820

Table 8: Sentiment prediction results on CMU-MOSI test set using multimodal methods. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art. Improvements are highlighted in green. The GMFN significantly outperforms the current state of the art across all evaluation metrics.

Dataset	ICT-MMMO		YouTube		MOUD	
Task	Sentiment		Sentiment		Sentiment	
Method	A^2	F1	A^3	F1	A^2	F1
Majority	40.0	22.9	42.4	25.2	60.4	45.5
RF	70.0	69.8	33.3	32.3	64.2	63.3
SVM	68.8	68.7	42.4	37.9	59.4	45.5
THMM	53.8	53.0	42.4	27.9	61.3	57.0
DF	65.0	58.7	45.8	32.0	67.0	67.1
EF-LSTM	66.3	65.0	44.1	43.6	67.0	64.3
EF-SLSTM	72.5	70.9	40.7	41.2	56.6	51.4
EF-BLSTM	63.8	49.6	42.4	38.1	58.5	58.9
EF-SBLSTM	62.5	49.0	37.3	33.2	63.2	63.3
MV-LSTM	72.5	72.3	45.8	43.3	57.6	48.2
BC-LSTM	70.0	70.1	45.0	45.1	72.6	72.9
TFN	72.5	72.6	45.0	41.0	63.2	61.7
MARN	71.3	70.2	48.3	44.9	**81.1**	**81.2**
MFN	73.8	73.1	51.7	51.6	**81.1**	80.4
GMFN no W	71.3	71.1	45.0	40.2	71.3	71.1
GMFN no E	67.5	66.7	48.3	48.2	67.5	66.7
GMFN no G	68.8	68.9	46.7	46.9	77.4	77.0
GMFN no M	58.8	59.5	45.0	35.0	70.8	66.9
GMFN no S	72.5	71.3	53.3	45.8	81.1	80.5
GMFN	**76.3**	**76.2**	**55.0**	**53.5**	**81.1**	80.9
Δ_{SOTA}	↑ **2.5**	↑ **3.1**	↑ **9.2**	↑ **8.4**	**0.0**	↑ **8.0**

Table 9: Sentiment prediction results on ICT-MMMO, YouTube and MOUD test sets. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art. Improvements are highlighted in green. The GMFN significantly outperforms the current state of the art across all evaluation metrics.

Dataset	IEMOCAP Emotions							
Task	Happy		Sad		Angry		Neutral	
Method	A^2	F1	A^2	F1	A^2	F1	A^2	F1
Majority	85.6	79.0	79.4	70.3	75.8	65.4	59.1	44.0
SVM	86.1	81.5	81.1	78.8	82.5	82.4	65.2	64.9
RF	85.5	80.7	80.1	76.5	81.9	82.0	63.2	57.3
THMM	85.6	79.2	79.5	79.8	79.3	73.0	58.6	46.4
EF-HCRF	85.7	79.2	79.4	70.3	75.8	65.4	59.1	44.0
EF-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
MV-HCRF	15.0	4.9	79.4	70.3	24.2	9.4	59.1	44.0
MV-LDHCRF	85.7	79.2	79.4	70.3	75.8	65.4	59.1	44.0
CMV-HCRF	14.4	3.6	79.4	70.3	24.2	9.4	59.1	44.0
CMV-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
EF-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
MV-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
DF	86.0	81.0	81.8	81.2	75.8	65.4	59.1	44.0
EF-LSTM	85.2	83.3	82.1	81.1	84.5	84.3	68.2	67.1
EF-SLSTM	85.6	79.0	80.7	80.2	82.8	82.2	68.8	68.5
EF-BLSTM	85.0	83.7	81.8	81.6	84.2	83.3	67.1	66.6
EF-SBLSTM	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1
MV-LSTM	85.9	81.3	80.4	74.0	85.1	84.3	67.0	66.7
BC-LSTM	84.9	81.7	83.2	81.7	83.5	84.2	67.5	64.1
TFN	84.8	83.6	83.4	82.8	83.4	84.2	67.5	65.4
GMFN no W	85.6	79.0	80.8	78.2	82.6	81.3	66.0	63.8
GMFN no E	86.0	81.6	82.8	81.4	83.2	81.5	67.6	66.2
GMFN no G	86.5	84.0	83.2	82.1	84.0	83.0	66.6	66.0
GMFN no M	85.6	83.0	81.3	79.1	82.0	79.3	66.2	66.2
GMFN no S	86.4	82.6	82.6	81.2	83.9	83.4	68.4	66.8
GMFN	**86.8**	**84.2**	**83.8**	**83.0**	**85.8**	**85.5**	**69.4**	**68.9**
Δ_{SOTA}	↑ **0.8**	**0.0**	↑ **0.4**	↑ **0.2**	↑ **0.6**	↑ **1.0**	↑ **0.6**	↑ **0.4**

Table 10: Emotion recognition results on IEMOCAP test set using multimodal methods. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art. Improvements are highlighted in green. The GMFN significantly outperforms the current state of the art across all evaluation metrics.

Dataset	POM															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	MA(7)	MA(7)	MA(7)	MA(7)	MA(7)	MA(7)	MA(7)	MA(7)	MA(5)	MA(5)	MA(5)	MA(5)	MA(5)	MA(5)	MA(7)	MA(5)
Majority	19.2	20.2	30.5	18.2	21.7	25.6	26.1	19.7	29.6	44.3	39.4	36.0	31.0	24.1	20.7	6.9
SVM	20.6	20.7	32.0	35.0	25.1	29.1	26.6	31.5	34.0	50.2	49.8	42.9	39.9	41.4	28.1	36.0
RF	26.6	27.1	29.6	26.1	23.2	23.6	26.6	26.1	34.0	53.2	40.9	32.5	37.4	36.0	25.6	40.4
THMM	24.1	15.3	19.2	29.1	27.6	26.1	18.7	12.3	22.7	31.0	31.5	30.0	30.0	27.1	17.2	24.6
DF	25.6	24.1	33.0	34.0	26.1	32.0	26.6	29.6	30.0	53.7	50.2	39.4	37.9	42.4	26.6	34.5
EF-LSTM	20.7	27.6	31.5	35.0	25.1	31.0	25.1	29.1	30.0	48.3	48.3	38.4	42.4	40.4	25.6	36.0
EF-SLSTM	22.2	28.6	30.5	36.9	27.1	32.0	27.6	27.6	32.5	49.3	46.8	40.4	39.9	41.9	22.7	35.0
EF-BLSTM	25.1	26.1	34.0	32.0	29.6	31.0	25.6	33.5	30.0	52.2	46.3	34.0	41.9	42.9	25.6	39.4
EF-SBLSTM	23.2	30.5	29.1	31.0	27.6	32.5	31.0	25.1	33.5	52.7	47.8	38.4	39.4	44.8	25.6	38.9
MV-LSTM	25.6	28.6	28.1	34.5	25.6	32.5	29.6	29.1	33.0	52.2	50.7	38.4	37.9	42.4	26.1	38.9
BC-LSTM	26.6	26.6	31.0	33.0	27.6	36.5	30.5	29.6	33.0	52.2	47.3	37.9	45.8	36.0	27.1	36.5
TFN	24.1	31.0	31.5	34.5	24.6	25.6	27.6	29.1	30.5	38.9	35.5	37.4	33.0	42.4	27.6	33.0
GMFN no W	30.0	31.5	33.5	36.0	29.6	31.0	27.6	35.5	31.0	53.2	50.7	39.9	41.9	42.4	28.1	40.9
GMFN no E	26.1	31.5	33.5	33.0	28.6	31.0	27.1	29.6	31.0	53.2	50.7	39.9	41.9	42.4	27.1	40.9
GMFN no G	29.1	31.5	33.5	33.5	29.6	34.0	31.0	30.5	31.5	53.2	50.7	42.9	42.9	43.8	31.5	40.9
GMFN no M	24.1	31.5	28.6	27.6	27.1	27.6	26.6	32.0	31.0	34.5	37.4	39.9	41.4	42.4	23.6	40.9
GMFN no S	32.5	31.5	34.0	36.9	30.5	32.5	34.5	35.0	33.5	53.7	50.7	42.4	45.8	45.3	31.5	42.4
GMFN	**35.5**	**35.0**	**36.5**	**41.9**	**32.0**	**36.9**	**37.4**	**36.9**	**37.4**	**55.2**	**53.7**	**45.8**	**48.3**	**46.3**	**34.5**	**43.8**
Δ_{SOTA}	↑ **8.9**	↑ **4.0**	↑ **3.5**	↑ **6.9**	↑ **4.4**	↑ **0.4**	↑ **6.9**	↑ **5.4**	↑ **3.4**	↑ **1.5**	↑ **3.0**	↑ **2.9**	↑ **2.5**	↑ **3.9**	↑ **6.4**	↑ **3.4**

Table 11: Results for personality trait recognition on the POM dataset. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over SOTA. Improvements over the SOTA are highlighted in green. The GMFN significantly outperforms the current SOTA across all datasets and evaluation metrics.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. USENIX Association, Berkeley, CA, USA, OSDI'16, pages 265–283. http://dl.acm.org/citation.cfm?id=3026877.3026899.

Harika Abburi, Rajendra Prasath, Manish Shrivastava, and Suryakanth V Gangashetty. 2016. Multimodal sentiment analysis using deep neural networks. In *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, pages 58–65.

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, 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.

Leonard E Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics* 37(6):1554–1563.

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.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

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.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, USA, ICML'11, pages 513–520. http://dl.acm.org/citation.cfm?id=3104482.3104547.

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.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *CoRR* abs/1508.01991.

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

Itseez. 2015. Open source computer vision library. `https://github.com/itseez/opencv`.

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.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pages 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018a. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018b. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI 2018.

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*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2247–2256. http://aclweb.org/anthology/P18-1209.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. Cite arxiv:1603.01354Comment: 10 pages, 3 figures. To appear on ACL 2016. http://arxiv.org/abs/1603.01354.

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.

Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Association for Computational Linguistics, pages 97–102.

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.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In Lise Getoor and Tobias Scheffer, editors, *ICML*. Omnipress, pages 689–696.

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.

Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, Melbourne, Australia, pages 53–63. http://www.aclweb.org/anthology/W18-3308.

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.

Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. 2007. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(10):1848–1852. https://doi.org/10.1109/TPAMI.2007.1124.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016a. *Extending Long Short-Term Memory for Multi-View Structured Learning*, Springer International Publishing, Cham, pages 338–353.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016b. 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.

Yale Song, Louis-Philippe Morency, and Randall Davis. 2012. Multi-view latent variable discriminative models for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pages 2120–2127.

Yale Song, Louis-Philippe Morency, and Randall Davis. 2013. Action recognition by hierarchical sequence summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 3562–3569.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pages 2222–2230. http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf.

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.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* abs/1605.02688. http://arxiv.org/abs/1605.02688.

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.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* .

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. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* .

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* .