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 Multimodal Opinion Sentiment and Emotion Intensity (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 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 (Hazariika et al., 2018; Poria et al., 2017a; Chen et al., 2017) and multimodal machine learning (Zadeh et al., 2018a; 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 generalize across different domains. Models trained on only a few topics generalize poorly as the learnt language and non-verbal 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 to 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 solution 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., 2016a) 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 Multimodal Opinion Sentiment and Emotion Intensity (MOSEI). 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 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 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 MOSEI to an extensive pool of datasets for sentiment analysis and emotion recognition. The following datasets include a combination of language, visual and acoustic modalities as their input data.

2.1.1 Multimodal Datasets

CMU-MOSI (Zadeh et al., 2016b) is a collection of 2199 opinion video clips each annotated with sentiment in the range [-3,3]. The ICT-MMMO
Table 1: Comparison between the 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.

(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 MOSEI to the datasets in this section is presented in Table 1. MOSEI has a longer total duration as well as a larger number of total data points. Furthermore, 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, 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; Zadeh et al., 2016b; Perez-Rosas et al., 2013; Park et al., 2014), Random Forests (Breiman, 2001) or Neural Networks (Nojavanaghari 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 segment. Furthermore, the representation over long periods of time become less informative and therefore 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., 2004) 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 Contextual-LSTM (Poria et al., 2017b) leverages additional contextual information from entire videos to obtain predictions for a single segment. More advanced recurrent models were also proposed to model both intra-modal and cross-modal interactions. 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. The Multi-attention Recurrent Network (Zadeh et al., 2018b) models both view-specific and cross-view dynamics using multiple attention coefficient assignments. Intra-modal and cross-modal interactions are stored in a hybrid LSTM memory component. 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), or modeling inter and intra modal interactions by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions (Zadeh et al., 2017).

3 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 Multi-modal Opinion Sentiment and Emotion Intensity (MOSEI). MOSEI is the largest to date with both attributes annotated. In the following subsections, we first explain the details of the 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 MOSEI. The diversity of the video topics brings the following generalizability advantages: (1) models trained on 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 MOSEI a one-of-a-kind dataset for sentiment analysis and emotion recognition.

Figure 1: The diversity of topics of videos in 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.

3.1.2 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.

3.2 Dataset Splits

The 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, 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 MOSEI Natural Split, and then discuss how we
acquired videos with more polarized sentiment and emotion in order to balance the complete MOSEI dataset.

3.2.1 MOSEI Natural Split

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

### 3.5 Annotation

Annotation of MOSEI follows closely the annotation of CMU-MOSI (Zadeh et al., 2016a) and Stanford Sentiment Treebank (Socher et al., 2013). Each sentence is annotated for sentiment on a [-3,3] Likert scale of: [-3: highly negative, -2 negative, -1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 highly positive]. Ekman emotions (Ekman et al., 1980) of {happiness, sadness, anger, fear, disgust, surprise} are annotated on a [0,3] Likert scale for presence of emotion $x$: [0: no evidence of $x$, 1: weakly $x$, 2: $x$, 3: highly $x$]. The annotation was carried out by 3 crowdsourced judges from Amazon Mechanical Turk platform. To avert implicitly biasing the judges and to capture the raw perception of the crowd, we avoided extreme annotation training and instead provided the judges with a 5 minute 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 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

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 $x$]. 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 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 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 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 (Tajman 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 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 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 $m$th modality is denoted as: $x^m = [x^m_t : t \leq T, x^m_t \in \mathbb{R}^{d_m}]$ where $d_m$ is the dimension input of $m$th modality $x_m$.

For each modality, a Long-Short Term Memory (LSTM) encodes the intra-modal interactions over time. For the $m$th modality, the memory of assigned LSTM is denoted as $c^m = \{c^m_t : t \leq T, c^m_t \in \mathbb{R}^{d_m}\}$ and the output is defined as $h^m = \{h^m_t : t \leq T, h^m_t \in \mathbb{R}^{d_m}\}$ with $d_m$ denoting the dimensionality of $m$th LSTM memory $c^m$.

The following update rules are defined for the $m$th LSTM (Hochreiter and Schmidhuber, 1997):

\[
\begin{align*}
    i^m_t &= \sigma(W^i_m x^m_t + U^i_m h^m_{t-1} + b^i_m) \\
    f^m_t &= \sigma(W^f_m x^m_t + U^f_m h^m_{t-1} + b^f_m) \\
    o^m_t &= \sigma(W^o_m x^m_t + U^o_m h^m_{t-1} + b^o_m) \\
    c^m_t &= f^m_t \odot c^m_{t-1} + i^m_t \odot \tilde{c}^m_t \\
    h^m_t &= o^m_t \odot \text{tanh}(c^m_t)
\end{align*}
\]

The parameters include the two affine transformations $W^m \in \mathbb{R}^{d_m \times d_m}$ and $U^m \in \mathbb{R}^{d_m \times d_m}$. $i^m, f^m, o^m$ are the input, forget and output gates of the $m$th LSTM respectively, $\tilde{c}^m_t$ is the proposed memory update of $m$th LSTM for time $t$, $\odot$ denotes the Hadamard product (element-wise product), $\sigma$ 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 $D_a : \mathbb{R}^{2d_c} \mapsto \mathbb{R}^{2d_c}$, $d_c = \sum_n d_{c_n}$ to obtain the attention coefficients.

\[
a_{[t-1,t]} = D_a(c_{[t-1,t]})
\]

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

\[
\hat{c}_{[t-1,t]} = c_{[t-1,t]} \odot a_{[t-1,t]}
\]

$\hat{c}_{[t-1,t]}$ 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 $D_u : \mathbb{R}^{2d_c} \mapsto \mathbb{R}^{d_{mem}}$ to generate a cross-view update proposal $\hat{u}_t$ for $u$. $d_{mem}$ is the dimensionality of the memory.

\[
\hat{u}_t = D_u(\hat{c}_{[t-1,t]})
\]

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: $\gamma^{(1)}$ and $\gamma^{(2)}$ 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

Figure 6: The structure of Dynamic Fusion Graph (DFG) for three modalities of \{(language, vision, acoustic\}. Dashed lines in DFG show the dynamic connections between vertices controlled by the efficacies ($\alpha$).
At each time-step of GMFN recursion, noted as \( \eta(t) \) in Equation (11), and (3) it can dynamically alter its previous approaches such as Tensor Fusion (Zadeh et al., 2017)), and (2) it does so with a neural network. \( D_{\hat{\gamma}}^{(1)}, D_{\hat{\gamma}}^{(2)} : \mathbb{R}^{2d_{m}} \rightarrow \mathbb{R}^{d_{mem}} \) control part of the gating mechanism of Multi-view Gated Memory using \( \hat{c}_{t} \) as input:

\[
\gamma_{t}^{(1)} = D_{\hat{\gamma}}^{(1)}(\hat{c}_{t-1,t}), \quad \gamma_{t}^{(2)} = D_{\hat{\gamma}}^{(2)}(\hat{c}_{t-1,t})
\]

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

\[
u_{t} = \gamma_{t}^{(1)} \odot u_{t-1} \odot \gamma_{t}^{(2)} \odot \tanh(\hat{u}_{t})
\]

\( \hat{u}_{t} \) 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 make 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 \( M = \{ l, v, a \} \). 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 \( \langle \text{language} \rangle \) and \( \langle \text{language, vision} \rangle \). 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 \( \alpha_{ij} \) that determines the importance of that edge in multimodal fusion. Specifically, each \( \alpha \) 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 \( \alpha \) are the main source of interpretability in DFG. The vector of all \( \alpha \)s is inferred using a deep neural network \( D_{\alpha} \) which takes as input singleton vertices in \( V \) (\( l, v, a \)). We leave it to the supervised training objective to learn parameters of \( D_{\alpha} \) 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 \( V \) (\( l, v, a \)), efficacies for each edge \( \alpha_{ij} \) 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 \( \alpha_{ij} \) before being used as input to \( D_{j} \), a neural network that performs local fusion of several weighted vertices. In total, \( D_{j} \) takes as input all \( v_i \) that satisfy the neural connection formula: \( v_i \subset v_j \). The output of \( D_{j} \) 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 \( l, v, a \) are the inputs to the DFG. All vertices are connected to the output vertex \( T \).
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 Fusion Network (GMFN). Figure 7 shows the overall architecture of the GMFN.

Similar to MFN, GMFN employs a system of LSTMs for modeling individual modalities. $c_l, c_e$, 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 $h_{t-1}^m$, 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 $T_t$ for storage in the Multi-view Gated Memory $u_t$. The Multi-view Gated Memory functions using a network $D_h$ that transforms $T_t$ 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 $T_t$ 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:

$$ h_T = \bigoplus_{m \in M} h_T^m $$

representing individual sequence information. $\bigoplus$ 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 MOSEI. Accuracy is reported as $A^2$ where $C$ is the number of classes as well as F1 measure. For regression we report MAE and correlation ($r$). For emotion recognition
We achieve state-of-the-art performance with state-of-the-art approaches.

The generalization of the GMFN to speakers com-

the-art approaches for these three dataset. To assess

shows the comparison of our GMFN with state-of-

material. The best results are highlighted in bold and

lower values indicate better performance. GMFN

Improvements are highlighted in green. GMFN achieves

We experiment on speaker traits recognition based

which our results are competitive but not state of

the art.

Table 5 shows the performance of the GMFN on

IEMOCAP dataset are reported in the bottom half

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.

6.3 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:

due to the natural imbalances across various emo-
tions, we use weighted accuracy (Tong et al., 2017) and F1 score. 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 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-MMMMO 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CMU-MOSI</th>
<th>ICT-MMMMO</th>
<th>YouTube</th>
<th>MOUD</th>
<th>EMOCAP Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Sentiment</td>
<td>Sentiment</td>
<td>Sentiment</td>
<td>Happy</td>
<td>Sad</td>
</tr>
<tr>
<td>Metric</td>
<td>A2</td>
<td>F1</td>
<td>MAE</td>
<td>r</td>
<td>A2</td>
</tr>
<tr>
<td>SOTA1</td>
<td>77.1</td>
<td>77.0</td>
<td>34.7</td>
<td>0.9</td>
<td>72.5</td>
</tr>
<tr>
<td>SOTA2</td>
<td>77.4</td>
<td>73.7</td>
<td>34.7</td>
<td>0.9</td>
<td>63.6</td>
</tr>
<tr>
<td>GMFN no W</td>
<td>71.3</td>
<td>71.3</td>
<td>27.0</td>
<td>1.1</td>
<td>0.5</td>
</tr>
<tr>
<td>GMFN no E</td>
<td>73.2</td>
<td>73.0</td>
<td>33.2</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>GMFN no G</td>
<td>75.5</td>
<td>75.5</td>
<td>30.5</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>GMFN no M</td>
<td>63.4</td>
<td>63.4</td>
<td>23.5</td>
<td>1.3</td>
<td>0.1</td>
</tr>
<tr>
<td>GMFN no S</td>
<td>75.1</td>
<td>75.0</td>
<td>33.7</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>GMFN</td>
<td>77.7</td>
<td>77.7</td>
<td>35.6</td>
<td>0.9</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Con</th>
<th>Pas</th>
<th>Voi</th>
<th>Dom</th>
<th>Cre</th>
<th>Viv</th>
<th>Exp</th>
<th>Ent</th>
<th>Res</th>
<th>Tru</th>
<th>Rel</th>
<th>Out</th>
<th>Tho</th>
<th>Ner</th>
<th>Per</th>
<th>Hum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
</tr>
<tr>
<td>SOTA1</td>
<td>26.6</td>
<td>27.6</td>
<td>32.0</td>
<td>35.0</td>
<td>26.1</td>
<td>32.0</td>
<td>27.6</td>
<td>29.6</td>
<td>34.0</td>
<td>35.3</td>
<td>52.2</td>
<td>49.8</td>
<td>39.4</td>
<td>42.4</td>
<td>42.4</td>
<td>27.6</td>
<td>35.4</td>
</tr>
<tr>
<td>SOTA2</td>
<td>26.0</td>
<td>31.0</td>
<td>33.0</td>
<td>35.0</td>
<td>27.6</td>
<td>36.5</td>
<td>30.5</td>
<td>31.5</td>
<td>34.0</td>
<td>33.7</td>
<td>50.7</td>
<td>42.9</td>
<td>45.8</td>
<td>45.8</td>
<td>42.8</td>
<td>28.1</td>
<td>40.4</td>
</tr>
<tr>
<td>GMFN no W</td>
<td>30.0</td>
<td>31.5</td>
<td>33.5</td>
<td>36.0</td>
<td>29.6</td>
<td>31.0</td>
<td>27.6</td>
<td>35.5</td>
<td>31.0</td>
<td>53.2</td>
<td>50.7</td>
<td>39.9</td>
<td>41.9</td>
<td>42.4</td>
<td>28.1</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>GMFN no E</td>
<td>26.1</td>
<td>31.5</td>
<td>33.5</td>
<td>33.0</td>
<td>28.6</td>
<td>31.0</td>
<td>27.1</td>
<td>29.6</td>
<td>31.0</td>
<td>53.2</td>
<td>50.7</td>
<td>39.9</td>
<td>41.9</td>
<td>42.4</td>
<td>27.1</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>GMFN no G</td>
<td>29.1</td>
<td>31.5</td>
<td>33.5</td>
<td>33.5</td>
<td>29.6</td>
<td>34.0</td>
<td>31.0</td>
<td>30.5</td>
<td>31.5</td>
<td>53.2</td>
<td>50.7</td>
<td>42.9</td>
<td>42.9</td>
<td>43.8</td>
<td>43.8</td>
<td>41.9</td>
<td></td>
</tr>
<tr>
<td>GMFN no M</td>
<td>24.1</td>
<td>31.5</td>
<td>28.6</td>
<td>27.6</td>
<td>27.6</td>
<td>28.6</td>
<td>26.6</td>
<td>32.0</td>
<td>31.0</td>
<td>34.5</td>
<td>37.4</td>
<td>39.9</td>
<td>41.4</td>
<td>42.4</td>
<td>23.6</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>GMFN no S</td>
<td>32.5</td>
<td>31.5</td>
<td>34.0</td>
<td>36.9</td>
<td>30.5</td>
<td>32.5</td>
<td>34.5</td>
<td>35.0</td>
<td>33.5</td>
<td>53.7</td>
<td>50.7</td>
<td>42.4</td>
<td>45.8</td>
<td>45.3</td>
<td>31.5</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>GMFN</td>
<td>35.5</td>
<td>35.0</td>
<td>36.5</td>
<td>41.9</td>
<td>36.9</td>
<td>37.4</td>
<td>36.9</td>
<td>37.4</td>
<td>37.4</td>
<td>55.2</td>
<td>53.7</td>
<td>45.8</td>
<td>48.3</td>
<td>46.3</td>
<td>34.5</td>
<td>43.8</td>
<td></td>
</tr>
</tbody>
</table>

Δ SOTA

| 1.8 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 5: Results for multimodal sentiment analysis on the CMU-MOSI, ICT-MMMMO, 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 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.
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 inter-modal 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 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 $T$ is low, implying that $T$ 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 mod-
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 to 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.

ify 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.

7 Conclusion

In this paper, we presented the largest dataset of multimodal sentiment analysis and emotion recognition called Multimodal Opinion Sentiment and Emotion Intensity (MOSEI). 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 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 Zhuo 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.
Appendix: MOSEI Full Results

The full set of results using various approaches on MOSEI is divided into multimodal and unimodal and shown in Tables 6 and 7 respectively.

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.
Table 7: Results of sentiment analysis and emotion recognition on the MOSEI Overall and Natural split.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentiment</th>
<th>FE W</th>
<th>A</th>
<th>MAE</th>
<th>Corr</th>
<th>W</th>
<th>A</th>
<th>FE MAE Corr W</th>
<th>A</th>
<th>FE MAE Corr W</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMFN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Results for sentiment analysis and emotion recognition on the MOSEI Overall and Natural split. The best results are highlighted in bold and the SOTA entries highlighted in gray. These results show that sentiment prediction and emotion recognition on the MOSEI Mega Corpus is non-trivial.
### Table 8: Sentiment prediction results on CMU-MOSI test set using multimodal methods. The best results are highlighted in bold and $\Delta_{SOTA}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>A$^2$</th>
<th>F1</th>
<th>A$^3$</th>
<th>MAE</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>50.2</td>
<td>50.1</td>
<td>17.5</td>
<td>1.864</td>
<td>0.057</td>
</tr>
<tr>
<td>RF</td>
<td>56.4</td>
<td>56.3</td>
<td>21.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>71.6</td>
<td>72.3</td>
<td>26.5</td>
<td>1.100</td>
<td>0.559</td>
</tr>
<tr>
<td>THMM</td>
<td>50.7</td>
<td>45.4</td>
<td>17.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SAL-CNN</td>
<td>73.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-MKL</td>
<td>72.3</td>
<td>72.0</td>
<td>30.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF-HCRF</td>
<td>65.3</td>
<td>65.4</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF-LDHCRF</td>
<td>64.0</td>
<td>64.0</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>44.8</td>
<td>27.7</td>
<td>22.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV-LDHCRF</td>
<td>64.0</td>
<td>64.0</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CMV-HCRF</td>
<td>44.8</td>
<td>27.7</td>
<td>22.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CMV-LDHCRF</td>
<td>63.6</td>
<td>63.6</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF-HSSHCRF</td>
<td>63.3</td>
<td>63.4</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV-HSSHCRF</td>
<td>65.6</td>
<td>65.7</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DF</td>
<td>72.3</td>
<td>72.1</td>
<td>26.8</td>
<td>1.143</td>
<td>0.518</td>
</tr>
<tr>
<td>EF-LSTM</td>
<td>74.3</td>
<td>74.3</td>
<td>32.4</td>
<td>1.023</td>
<td>0.622</td>
</tr>
<tr>
<td>EF-SLSTM</td>
<td>72.7</td>
<td>72.8</td>
<td>29.3</td>
<td>1.081</td>
<td>0.600</td>
</tr>
<tr>
<td>EF-BLSTM</td>
<td>72.0</td>
<td>72.0</td>
<td>28.9</td>
<td>1.090</td>
<td>0.577</td>
</tr>
<tr>
<td>EF-SBLSTM</td>
<td>73.3</td>
<td>73.2</td>
<td>26.8</td>
<td>1.037</td>
<td>0.619</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>73.9</td>
<td>74.0</td>
<td>33.2</td>
<td>1.019</td>
<td>0.601</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>73.9</td>
<td>73.9</td>
<td>28.7</td>
<td>1.079</td>
<td>0.581</td>
</tr>
<tr>
<td>TFN</td>
<td>74.6</td>
<td>74.5</td>
<td>28.7</td>
<td>1.040</td>
<td>0.587</td>
</tr>
<tr>
<td>GME-LSTM(A)</td>
<td>76.5</td>
<td>73.4</td>
<td>-</td>
<td>0.955</td>
<td>-</td>
</tr>
<tr>
<td>MARN</td>
<td>77.1</td>
<td>77.0</td>
<td>34.7</td>
<td>0.968</td>
<td>0.625</td>
</tr>
<tr>
<td>MFN</td>
<td>77.4</td>
<td>77.3</td>
<td>34.1</td>
<td>0.965</td>
<td>0.632</td>
</tr>
<tr>
<td>GMFN no W</td>
<td>71.3</td>
<td>71.3</td>
<td>27.0</td>
<td>1.143</td>
<td>0.537</td>
</tr>
<tr>
<td>GMFN no E</td>
<td>73.2</td>
<td>73.0</td>
<td>33.2</td>
<td>1.040</td>
<td>0.603</td>
</tr>
<tr>
<td>GMFN no G</td>
<td>75.5</td>
<td>75.5</td>
<td>30.5</td>
<td>0.999</td>
<td>0.610</td>
</tr>
<tr>
<td>GMFN no M</td>
<td>63.4</td>
<td>63.4</td>
<td>23.5</td>
<td>1.312</td>
<td>0.408</td>
</tr>
<tr>
<td>GMFN no S</td>
<td>75.1</td>
<td>75.0</td>
<td>33.7</td>
<td>0.998</td>
<td>0.634</td>
</tr>
<tr>
<td>GMFN</td>
<td>77.7</td>
<td>77.7</td>
<td>35.6</td>
<td>0.955</td>
<td>0.664</td>
</tr>
<tr>
<td>$\Delta_{SOTA}$</td>
<td>0.3</td>
<td>0.4</td>
<td>1.9</td>
<td>0.0</td>
<td>1.03</td>
</tr>
</tbody>
</table>

### Table 9: Sentiment prediction results on ICT-MMO, YouTube and MOUD test sets. The best results are highlighted in bold and $\Delta_{SOTA}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Method</th>
<th>Task</th>
<th>Sentiment</th>
<th>Sentiment</th>
<th>Sentiment</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>85.0</td>
<td>Happy</td>
<td>79.0</td>
<td>79.0</td>
<td>73.5</td>
<td>85.4</td>
</tr>
<tr>
<td>RF</td>
<td>85.5</td>
<td>Sad</td>
<td>80.7</td>
<td>80.1</td>
<td>81.9</td>
<td>82.0</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>85.6</td>
<td>Angry</td>
<td>79.2</td>
<td>79.5</td>
<td>78.9</td>
<td>73.0</td>
</tr>
<tr>
<td>THMM</td>
<td>15.0</td>
<td>Neutral</td>
<td>79.4</td>
<td>79.7</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN no W</td>
<td>14.4</td>
<td>Happy</td>
<td>79.2</td>
<td>79.4</td>
<td>73.5</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN no E</td>
<td>85.8</td>
<td>Sad</td>
<td>79.5</td>
<td>79.4</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN no G</td>
<td>15.0</td>
<td>Angry</td>
<td>79.6</td>
<td>79.7</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN no M</td>
<td>14.4</td>
<td>Neutral</td>
<td>79.9</td>
<td>79.7</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN no S</td>
<td>85.8</td>
<td>Happy</td>
<td>79.5</td>
<td>79.4</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>GMFN</td>
<td>85.8</td>
<td>Sad</td>
<td>79.5</td>
<td>79.4</td>
<td>75.8</td>
<td>65.4</td>
</tr>
<tr>
<td>$\Delta_{SOTA}$</td>
<td>0.3</td>
<td>0.4</td>
<td>1.9</td>
<td>0.0</td>
<td>1.03</td>
<td></td>
</tr>
</tbody>
</table>

### Table 10: Emotion recognition results on IEMOCAP test set using multimodal methods. The best results are highlighted in bold and $\Delta_{SOTA}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Method</th>
<th>Task</th>
<th>Happy</th>
<th>Sad</th>
<th>Angry</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>85.6</td>
<td>Happy</td>
<td>79.0</td>
<td>82.6</td>
<td>81.5</td>
<td>86.0</td>
</tr>
<tr>
<td>RF</td>
<td>85.6</td>
<td>Sad</td>
<td>84.3</td>
<td>83.0</td>
<td>83.2</td>
<td>86.0</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>85.6</td>
<td>Angry</td>
<td>84.0</td>
<td>83.2</td>
<td>83.0</td>
<td>86.0</td>
</tr>
<tr>
<td>THMM</td>
<td>85.6</td>
<td>Neutral</td>
<td>84.3</td>
<td>83.0</td>
<td>83.2</td>
<td>86.0</td>
</tr>
<tr>
<td>GMFN no W</td>
<td>84.3</td>
<td>Happy</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>GMFN no E</td>
<td>84.3</td>
<td>Sad</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>GMFN no G</td>
<td>84.3</td>
<td>Angry</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>GMFN no M</td>
<td>84.3</td>
<td>Neutral</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>GMFN no S</td>
<td>84.3</td>
<td>Happy</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>GMFN</td>
<td>84.3</td>
<td>Sad</td>
<td>83.4</td>
<td>82.8</td>
<td>82.8</td>
<td>86.8</td>
</tr>
<tr>
<td>$\Delta_{SOTA}$</td>
<td>0.3</td>
<td>0.4</td>
<td>1.9</td>
<td>0.0</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>Dataset</td>
<td>Task</td>
<td>Metric</td>
<td>POM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Majority</td>
<td>Con</td>
<td>MA(7)</td>
<td>MA(7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>20.6</td>
<td>20.7</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>26.6</td>
<td>27.1</td>
<td>29.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THMM</td>
<td>24.1</td>
<td>15.3</td>
<td>19.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td>25.6</td>
<td>24.1</td>
<td>33.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF-LSTM</td>
<td>20.7</td>
<td>27.6</td>
<td>31.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF-SLSTM</td>
<td>22.2</td>
<td>28.6</td>
<td>30.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF-BLSTM</td>
<td>25.1</td>
<td>26.1</td>
<td>34.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF-SBLSTM</td>
<td>23.2</td>
<td>30.5</td>
<td>29.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>25.6</td>
<td>28.6</td>
<td>28.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>26.6</td>
<td>26.6</td>
<td>31.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFN</td>
<td>24.1</td>
<td>31.0</td>
<td>31.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN no W</td>
<td>30.0</td>
<td>31.5</td>
<td>33.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN no E</td>
<td>26.1</td>
<td>31.5</td>
<td>33.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN no G</td>
<td>29.1</td>
<td>31.5</td>
<td>33.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN no M</td>
<td>24.1</td>
<td>31.5</td>
<td>28.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN no S</td>
<td>32.5</td>
<td>31.5</td>
<td>34.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMFN</td>
<td>35.5</td>
<td>35.0</td>
<td>36.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Results for personality trait recognition on the POM dataset. The best results are highlighted in bold and $\Delta_{SOTA}$ 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


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


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.


