Memory Fusion Network for Multi-view Sequential Learning

Amir Zadeh[1], Paul Pu Liang[1], Navonil Mazumder[2], Soujanya Poria[3], Erik Cambria[3], Louis-Philippe Morency[1]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[3]Nanyang Technological University, Singapore

{abagherz,pliang,morency}@cs.cmu.edu, navonil@sentic.net, {sporia,cambria}@ntu.edu.sg

Abstract

Multi-view sequential learning is a fundamental problem in machine learning dealing with multi-view sequences. In a multi-view sequence, there exists two forms of interactions between different views: view-specific interactions and cross-view interactions. In this paper, we present a new neural architecture for multi-view sequential learning called the Memory Fusion Network (MFN) that explicitly accounts for both interactions in a neural architecture and continuously models them through time. The first component of the MFN is called the System of LSTMs, where view-specific interactions are learned in isolation through assigning an LSTM function to each view. The cross-view interactions are then identified using a special attention mechanism called the Delta-memory Attention Network (DMAN) and summarized through time with a Multi-view Gated Memory. Through extensive experimentation, MFN is compared to various proposed approaches for multi-view sequential learning on multiple publicly available benchmark datasets. MFN outperforms all the multi-view approaches. Furthermore, MFN outperforms all current state of the art models, setting new state of the art results for all the multi-view datasets.

Introduction

In many natural scenarios, data is collected from diverse domains and exhibits heterogeneous properties: each of these domains have different dynamics and present a different view of the same data. Such forms of data are known as multi-view data. In a multi-view setting, each view of the data may contain some knowledge that other views do not have access to. Therefore, multiple views must be employed together in order to describe the data comprehensively and accurately. Multi-view learning has been an active area of machine learning research focused on modeling multi-view data (Xu, Tao, and Xu 2013). By exploring the consistency and complementary properties of different views, multi-view learning is considered more effective, more promising, and has better generalization ability than single-view learning.

Multi-view sequential learning extends the definition of multi-view learning to manage with different views all in the form of sequential data, i.e. data that comes in the form of sequences. For example, a video clip of a speech can be partitioned into three sequential views – text of the speech, video of the speaker, and tone of the speaker's voice. In multi-view sequential learning, two main forms of interactions exist. The first form is called view-specific interactions; interactions that involve only one view. One example involves learning the sentiment of a sentence based only on the sequence of words in that sentence. More importantly, the second form of interactions are defined across different views. These are known as cross-view interactions. Cross-view interactions span across different views and times – some examples include a listener's backchannel response or the delayed rumble of distant lightning in the video and audio views. Modeling these view-specific and cross-view interactions lies at the core of multi-view sequential learning.

This paper introduces a novel neural model for multi-view sequence learning called the Memory Fusion Network (MFN). The MFN encodes each view independently using a component called the System of LSTMs. In this System of LSTMs, each view is assigned one LSTM function to model the dynamics in that particular view. These LSTM functions have no connections between one another and only encode the input of their own view. The second component of MFN is called the Delta-memory Attention Network (DMAN) which searches for cross-view interactions across memories of the System of LSTMs. Specifically, the DMAN explicitly outlines the cross-view interactions by associating a relevance score to the memory dimensions of each LSTM. The DMAN synthesizes memories at the current and previous time-steps of all the different LSTMs to decide which cross-view interaction needs to be outlined. The third component of the MFN stores the outlined cross-view information over time in the Multi-view Gated Memory. This memory updates its contents based on the outputs of the DMAN and its previously stored contents, acting as a dynamic memory module for learning crucial cross-view interactions throughout the sequential data.

We perform extensive experimentation to benchmark the performance of MFN on 6 publicly available multi-view sequential datasets. Throughout, we compare to the state-of-the-art approaches in multi-view sequential learning. In all the benchmarks, MFN is able to outperform the baselines, setting new state-of-the-art results across all the datasets.

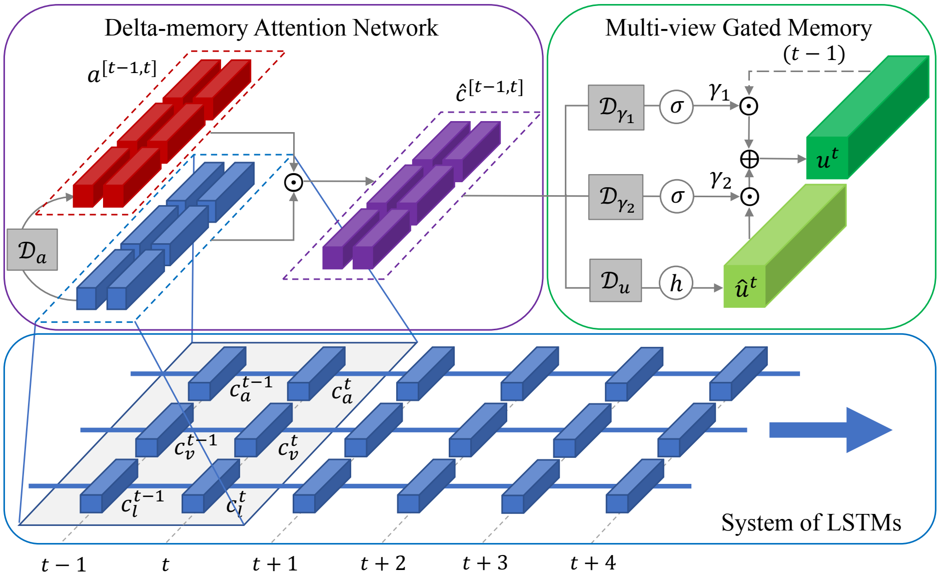

Figure 1: Overview figure of Memory Fusion Network (MFN) pipeline. σ denotes the $sigmoid$ activation function, h the $tanh$ activation function, $\odot$ the Hadamard product and $\oplus$ element wise addition. Each LSTM encodes information from one view such as language (l), video (v) or audio (a).

Related Work

Researchers dealing with multi-view sequential data have largely focused on three major thrusts:

The first category of models have relied on concatenation of all multiple views into a single view to simplify the learning setting. These approaches then use this single view as input to a learning model. Hidden Markov Models (HMMs) (Baum and Petrie 1966; Morency, Mihalcea, and Doshi 2011), Support Vector Machines (SVMs) (Cortes and Vapnik 1995; Zadeh et al. 2016), Hidden Conditional Random Fields (HCRFs) (Quattoni et al. 2007) and their variants (Morency, Quattoni, and Darrell 2007) have been successfully used for structured prediction. More recently, with the advent of deep learning, Recurrent Neural Networks, specially Long-short Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), have gained momentum for sequence modeling. Some degree of success for modeling multi-view problems is achieved using this concatenation. However, this concatenation causes over-fitting in the case of a small size training sample and is not physically meaningful because each view has a specific statistical property (Xu, Tao, and Xu 2013) which is ignored in these simplified approaches.

The second category of models introduce multi-view variants to the structured learning approaches of the first category. Multi-view variations of these models have been proposed including Multi-view HCRFs where the potentials of the HCRF are changed to facilitate multiple views (Song, Morency, and Davis 2012; 2013). Recently, multi-view LSTM models have been proposed for multimodal setups where the LSTM memory is partitioned into different components for different views (Rajagopalan et al. 2016).

The third category of models rely on collapsing the time dimension from sequences by learning a temporal representation for each of the different views. Such methods have used average feature values over time (Poria, Cambria, and Gelbukh 2015). Essentially these models apply conventional multi-view learning approaches, such as Multiple Kernel Learning (Poria, Cambria, and Gelbukh 2015), subspace learning or co-training (Xu, Tao, and Xu 2013) to the multi-view representations. Other approaches have trained different models for each view and combined the models using decision voting (Nojavanasghari et al. 2016), tensor products (Zadeh et al. 2017) or deep neural networks (Poria et al. 2017). While these approaches are able to learn the relations between the views to some extent, the lack of the temporal dimension limits these learned representations, eventually affect their performance. Such is the case for long sequences where the learned representations do not sufficiently reflect all the temporal information in each view.

The proposed model in this paper is different from the first category models since it assigns one LSTM to each view instead of concatenating the information from different views. MFN is also different from the second category models since it considers each views in isolation to learn view-specific interactions. It then uses an explicitly designed attention mechanism and memory to find and store cross-view interactions over time. MFN is different from the third category models since view-specific and cross-view interactions are modeled over time.

Memory Fusion Network (MFN)

The Memory Fusion Network (MFN) is a recurrent model for multi-view sequential learning that consists of three main components: 1) **System of LSTMs** consists of multiple Long-short Term Memory (LSTM) networks, one for each of the views. Each LSTM encodes its sequence independently over time. 2) **Delta-memory Attention Network** is a special attention mechanism designed to discover cross-view interactions across different dimensions of memories in the System of LSTMs. 3) **Multi-view Gated Memory** is an unifying memory that stores the cross-view interactions over time. Figure 1 shows the overview of MFN pipeline.

The input to MFN is a multi-view sequence with the set of N views (for example sequences, related to text, video, and audio for $N = \{t, v, a\}$) and length T. The input data of the nth view is denoted as: $\mathbf{x}_n = [x_n^{(t)} : t \leq T, x_n^{(t)} \in \mathbb{R}^{d_{x_n}}]$ where d_{x_n} is the input dimensionality of nth view input $\mathbf{x}_n$.

System of LSTMs

For each view sequence, a Long-Short Term Memory (LSTM), encodes the view-specific interactions over time. At each input timestamp t, information from each view is input to the assigned LSTM. For the nth view, the memory of assigned LSTM is denoted as $\mathbf{c}_n = \{c_n^{(t)} : t \leq T, c_n^{(t)} \in \mathbb{R}^{d_{c_n}}\}$ and the output of each track is defined as $\mathbf{h}_n = \{h_n^{(t)} : t \leq T, h_n^{(t)} \in \mathbb{R}^{d_{c_n}}\}$ with d_{c_n} denoting the dimensionality of nth track memory $\mathbf{c}_n$. Note that the System of LSTMs allows different sequences to have different input and memory (thus output) shapes. The following update rules are defined for the nth LSTM (Hochreiter and Schmidhuber 1997):

$$\begin{pmatrix} i_n^{(t)} \\ f_n^{(t)} \\ o_n^{(t)} \\ m_n^{(t)} \end{pmatrix} = \begin{pmatrix} sigmoid \\ sigmoid \\ sigmoid \\ tanh \end{pmatrix} U_n \begin{pmatrix} x_n^{(t)} W_n \\ h_n^{(t-1)} \end{pmatrix}$$

$$c_n^{(t)} = f_n^{(t)} \odot c_n^{(t-1)} + i_n^t \odot m_n^{(t)}$$

$$h_n^{(t)} = o_n^{(t)} \odot tanh(c_n^{(t)})$$

In the above equations, the trainable parameters are the two affine transformations $W_n \in \mathbb{R}^{d_{x_n} \times d_{c_n}}$ and $U_n \in \mathbb{R}^{d_{c_n} \times d_{c_n}}$. i_n, f_n, o_n are the input, forget and output gates of the n track respectively and $\odot$ denotes the Hadamard product (element-wise product). m_n is the proposed update to the memory of nth LSTM for time t.

Delta-memory Attention Network

The goal of the Delta-memory Attention Network (DMAN) is to outline the cross-view interactions at timestep t between different view memories in the System of LSTMs. To this end, we use a coefficient assignment technique on the concatenation of LSTM memories $c^{(t)}$ at time t. High coefficients are assigned to the dimensions that belong to a cross-view interaction and low coefficients to the other dimensions. However, coefficient assignment using only memories at time t is not ideal since the same cross-view interactions can happen over multiple time instances if the LSTM memories in those dimensions remain unchanged. This is especially troublesome if the recurring dimensions are assigned high coefficients, in which case they will dominate the coefficient assignment system. To deal with this problem we add the memories c^{t-1} of time $t-1$ so DMAN can have the freedom of leaving unchanged dimensions in the System of LSTMs memories and only assign high coefficient to them if they are about to change. Ideally each cross-view interaction is only assigned high coefficients once before the state of memories in System of LSTMs changes. This can be done by comparing the memories at the two time-steps (hence the name Delta-memory).

The input to the DMAN is the concatenation of memories at time $t-1$ and t, denoted as $c^{[t-1,t]}$. These memories are passed to a deep neural network $\mathcal{D}_a : \mathbb{R}^{2 \times d_c} \mapsto \mathbb{R}^{2 \times d_c}$ to obtain the attention coefficients.

$$a^{[t-1,t]} = \mathcal{D}_a(c^{[t-1,t]}) \tag{1}$$

$a^{[t-1,t]}$ are softmax activated scores for each LSTM memory at time $t-1$ and t. Applying softmax at the output layer of $\mathcal{D}_a$ allows for regularizing high-value coefficients over the $c^{[t-1,t]}$. The output of the DMAN is $\hat{c}$ defined as:

$$\hat{c}^{[t-1,t]} = c^{[t-1,t]} \odot a^{[t-1,t]} \tag{2}$$

$\hat{c}^{[t-1,t]}$ is the attended memories of the LSTMs. Applying this element-wise product amplifies the relevant dimensions of the $c^{[t-1,t]}$ while minimizing the effect of other dimensions. DMAN is also able to find cross-view interactions that do not happen simultaneously since it attends to the memories in the System of LSTMs. These memories can carry information about the observed inputs at different timestamps.

Multi-view Gated Memory

Multi-view Gated Memory u is the neural component that stores a dense history of cross-view interactions over time. It acts as a unifying memory for the memories in 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 LSTMs memories constitute a cross-view interaction. $\hat{c}^{[t-1,t]}$ is first used as input to a deep 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 Multi-view Gated Memory. d_{mem} is the dimensionality of the Multi-view Gated Memory.

$$\hat{u}^{(t)} = \mathcal{D}_u(\hat{c}^{[t-1,t]}) \tag{3}$$

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

The Multi-view Gated Memory is controlled using set of two gates. γ_1, γ_2 are called the retain and update gates respectively. At each timestep t, γ_1 assigns how much of the current state of the Multi-view Gated Memory to remember and γ_2 assigns how much of the Multi-view Gated Memory to update based on the update proposal $\hat{u}^{(t)}$. γ_1 and γ_2 are each controlled by a deep 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:

$$\gamma_1^t = \mathcal{D}_{\gamma_1}(\hat{c}^{[t-1,t]}) \tag{4}$$

$$\gamma_2^t = \mathcal{D}_{\gamma_2}(\hat{c}^{[t-1,t]}) \tag{5}$$

Finally, at each time-step of MFN recursion, u is updated using retain and update gates, γ_1 and γ_2, as well as the current cross-view update proposal $\hat{u}^{(t)}$ with the following formulation:

$$u^{(t)} = \gamma_1^{(t)} \odot u^{(t-1)} + \gamma_2^{(t)} \odot tanh(\hat{u}^{(t)}) \tag{6}$$

$\hat{u}^t$ is activated using $tanh$ squashing function to improve model stability by avoiding drastic changes to the Multi-view Gated Memory.

Output of MFN

The outputs of the MFN is the final state of the Multi-view Gated Memory $u^{(T)}$ and the outputs of each of the n LSTMs, $\mathbf{h}^T = \bigoplus_{n \in N} h_n^{(T)}$ as individual sequence information.

Experimental Setup

In this section we design extensive experiments to evaluate the performance of MFN. We choose 3 multi-view benchmarks – multimodal sentiment analysis, multimodal emotion recognition and multimodal speaker traits analysis. All benchmarks involve three views with completely different natures – language (text), vision (video), and acoustic (audio). The multi-view input signal is the video of a person speaking about a certain topic. Since humans communicate their intentions in a structured manner, there are synchronizations between intention in text, gestures used and tone of speech. These synchronizations constitute the relations between the three views.

Datasets

In all the videos in these datasets described below, only one speaker is present in front of the camera.

Sentiment Analysis The first benchmark in our experiments is multimodal sentiment analysis, where the goal is to identify a speakers sentiment based on the speakers display of intentions. Multimodal sentiment analysis extends the conventional text-based definition of sentiment analysis to a multimodal setup where different views contribute to modeling the sentiment of the speaker. We use four different datasets for English and Spanish sentiment analysis in our experiments. The *CMU-MOSI* dataset (Zadeh et al. 2016) is a collection of 93 opinion videos from online sharing websites. Each video consists of multiple opinion segments and each segment is annotated with sentiment in the range [-3,3]. The *MOUD* dataset (Perez-Rosas, Mihalcea, and Morency 2013) consists of product review videos in Spanish. Each video consists of multiple segments labeled to display positive, negative or neutral sentiment. To maintain consistency with previous works (Poria et al. 2017; Perez-Rosas, Mihalcea, and Morency 2013) we remove segments with the neutral label. The *YouTube* dataset (Morency, Mihalcea, and Doshi 2011) introduced tri-modal sentiment analysis to the research community. Multi-dimensional data from the audio, visual and textual modalities are collected in the form of 47 videos from the social media web site YouTube. The collected videos span a wide range of product reviews and opinion videos. These are annotated at the segment level for sentiment. The *ICT-MMMO* dataset (Wöllmer et al. 2013) consists of online social review videos that encompass a strong diversity in how people express opinions, annotated at the video level for sentiment.

Emotion Recognition The second benchmark in our experiments is multimodal emotion recognition, where the goal is to identify a speakers emotions based on the speakers verbal and nonverbal behaviors. These emotions are categorized as basic emotions (Ekman 1992) and continuous emotions (Gunes 2010). We perform experiments on *IEMOCAP* dataset (Busso et al. 2008). IEMOCAP consists of 151 sessions of recorded dialogues, of which there are 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral) as well as valence, arousal and dominance.

Speaker Traits Analysis The third benchmark in our experiments is speaker trait recognition based on communicative behavior of the speaker. The goal is to identify 16 different speaker traits. The *POM* dataset (Park et al. 2014) contains 1,000 movie review videos. Each video is annotated for various personality and speaker traits, specifically: confident (con), passionate (pas), voice pleasant (voi), dominant (dom), credible (cre), vivid (viv), expertise (exp), entertaining (ent), reserved (res), trusting (tru), relaxed (rel), outgoing (out), thorough (tho), nervous (ner), persuasive (per) and humorous (hum). The short form of these speaker traits is indicated inside the parentheses and used for the rest of this paper.

Sequence Features

The chosen system of sequences are the three modalities: language, visual and acoustic. To get the exact utterance timestamp of each word we perform forced alignment using P2FA which allows us to align the three modalities together. Since words are considered the basic units of language we use the interval duration of each word utterance as a time-step. We calculate the expected video and audio context by taking the expectation of their view feature values over the word utterance time interval (Zadeh et al. 2017). As a result, the first word and its aligned visual and audio features indicates $t = 1$, the second word and its aligned visual and audio features indicates $t = 2$, and so on. Applying this approach at each word time-step t, we obtain a system of 3 sequences each of length T and of dimensionality $d_{x_{text}}$, $d_{x_{video}}$ and $d_{x_{audio}}$. For each of the 3 modalities, we process the information from videos as follows:

Language View For the language view, Glove word embeddings (Pennington, Socher, and Manning 2014) were used to embed a sequence of individual words from video segment transcripts into a sequence of word vectors that represent spoken text. The Glove word embeddings used are 300 dimensional word embedding trained on 840 billion tokens from the common crawl dataset, resulting in a sequence of dimension $T \times d_{x_{text}} = T \times 300$ after alignment. The timing of word utterances is extracted using P2FA forced aligner (Yuan and Liberman 2008). This extraction enables alignment between text, audio and video.

Visual View For the visual view, the library Facet (iMotions 2017) is used to extract a set of visual features including

Dataset	CMU-MOSI	ICT-MMMO	YouTube	MOUD	IEMOCAP	POM
Level	Segment	Video	Segment	Segment	Segment	Video
# Train	52→1284	220	30→169	49→243	5→6373	600
# Valid	10→229	40	5→41	10→37	1→1775	100
# Test	31→686	80	11→59	20→106	1→1807	203

Table 1: Data splits to ensure speaker independent learning.

facial action units, facial landmarks, head pose, gaze tracking and HOG features (Zhu et al. 2006). These visual features are extracted from the full video segment at 30Hz to form a sequence of facial gesture changes throughout time, resulting in a sequence of dimension $T \times d_{x_{video}} = T \times 35$.

Acoustic View For the audio view, the software COVAREP (Degottex et al. 2014) is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features (Drugman and Alwan 2011), glottal source parameters (Childers and Lee 1991; Drugman et al. 2012; Titze and Sundberg 1992; Alku 1992; Alku, Strik, and Vilkman 1997; Alku, Bäckström, and Vilkman 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl 2013). These visual features are extracted from the full audio clip of each segment at 100Hz to form a sequence that represent variations in tone of voice over an audio segment, resulting in a sequence of dimension $T \times d_{x_{audio}} = T \times 74$ after alignment.

Experimental Details

The timestamps in the sequences are chosen based on word utterances. The expected visual and acoustic sequences features are calculated for each word utterance to ensure time alignment between all LSTMs. In all the aforementioned datasets, it is important that the same speaker does not appear in both train and test sets in order to learn generalizable speaker independent features. The training, validation and testing splits are performed at the level of videos so that there is no speaker dependent contamination in our experiments. The full set of videos (and segments for datasets where the annotations are at the resolution of segments) in each split is detailed in Table 1. All baselines were re-trained using these video-level train-test splits of each dataset and with the same set of extracted sequence features. Training is performed on the labeled segments for datasets annotated at the segment level and on the labeled videos otherwise. Upon acceptance of the paper all the code and data required to recreate the reported results will be made available on github (omitted for review).

Baseline Models

We compare the performance of the MFN with current state-of-the-art models for multi-view sequential learning. To perform a more extensive comparison we train all the baseline for all the benchmarks.

First Category: View Concatenation in Sequential Learning Models

Song2013: This is a layered model that uses CRFs with latent variables to learn hidden spatio-temporal dynamics. For each layer an abstract feature representation is learned through non-linear gate functions. This procedure is repeated to obtain a hierarchical sequence summary (HSS) representation (Song, Morency, and Davis 2013).

Morency2011: Hidden Markov Model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states (Baum and Petrie 1966). We follow the implementation in (Morency, Mihalcea, and Doshi 2011) for tri-modal data.

Quattoni2007: Concatenated features are used as input to a Hidden Conditional Random Field (HCRF) (Quattoni et al. 2007). HCRF learns a set of latent variables conditioned on the concatenated input at each time step.

Morency2007: Latent Discriminative Hidden Conditional Random Fields (LDHCRFs) are a class of models that learn hidden states in a Conditional Random Field using a latent code between observed input and hidden output (Morency, Quattoni, and Darrell 2007).

Hochreiter1997: A LSTM with concatenation of data from different views as input (Hochreiter and Schmidhuber 1997). Stacked, Bidirectional and Stacked Bidirectional LSTMs are also trained in a similar fashion for stronger baselines.

Second Category: Multi-view Sequence Learning Models

Rajagopalan2016: Multi-view (MV) LSTM (Rajagopalan et al. 2016) aims to extract information from multiple sequences by modeling sequence-specific and cross-sequence interactions over time and output. MV-LSTM is a strong tool for synchronizing a system of multi-dimensional data sequences.

Song2012: MV-HCRF (Song, Morency, and Davis 2012) is an extension of the HCRF for Multi-view data. Instead of view concatenation, view-shared and view specific substructures are explicitly learned to capture the interaction between views. We also implement the topological variations - linked, coupled and linked-couple that differ in the types of interactions between the modeled views. *Song2012LD*: is a variation of this model that uses LDHCRF instead of HCRF.

Song2013MV: MV-HSSHCRF is an extension of *Song2013* that performs Multi-view hierarchical sequence summary representation.

Third Category: Multi-view Learning by Sequence Representation Learning

Poria2015: Multiple Kernel Learning (Bach, Lanckriet, and Jordan 2004) classifiers have been widely applied to problems involving multi-view data. Our implementation follows a previously proposed model for multimodal sentiment analysis (Poria, Cambria, and Gelbukh 2015).

Nojavanasghari2016: Deep Fusion Approach (Nojavanasghari et al. 2016) trains single neural networks for each view's input and combine the views with a joint neural network. This baseline is current state of the art in POM dataset.

Zadeh2016: Support Vector Machine (Cortes and Vapnik 1995) is a widely used classifier. This baseline is closely implemented similar to a previous work in multimodal sentiment analysis (Zadeh et al. 2016).

Ho1998: To compare to another non-neural strong classifier, we also introduce another baseline using a Random Forest (Ho 1998).

Dataset	CMU-MOSI					ICT-MMMO				YouTube		MOUD	
Task	Sentiment					Sentiment				Sentiment		Sentiment	
Metric	BA	F1	MA(7)	MAE	r	BA	F1	MAE	r	MA(3)	F1	BA	F1
SOTA2	$73.9^{\dagger}$	$74.0^{\diamond}$	$32.4^{\S}$	$1.023^{\S}$	$0.601^{\diamond}$	$81.3^{\#}$	$79.6^{\#}$	$0.968^{\flat}$	$0.499^{\flat}$	$49.2^{\bullet}$	$49.2^{\bullet}$	$72.6^{\dagger}$	$72.9^{\dagger}$
SOTA1	74.6^{*}	74.5^{*}	$33.2^{\diamond}$	$1.019^{\diamond}$	$0.622^{\S}$	$81.3^{\blacksquare}$	$79.6^{\blacksquare}$	$0.842^{\S}$	$0.588^{\S}$	$50.2^{\clubsuit}$	$50.8^{\clubsuit}$	$74.0^{\clubsuit}$	$74.7^{\clubsuit}$
MFN l	73.2	73.0	32.9	1.012	0.607	60.0	55.3	1.144	0.042	50.9	49.1	69.8	69.9
MFN a	53.1	47.5	15.0	1.446	0.186	80.0	79.3	1.089	0.462	39.0	27.0	60.4	47.1
MFN v	55.4	54.7	15.0	1.446	0.155	58.8	58.6	1.204	0.402	42.4	35.7	61.3	47.6
MFN (no Δ)	75.5	75.2	34.5	0.980	0.626	76.3	75.8	0.890	0.577	55.9	55.4	71.7	70.6
MFN (no mem)	76.5	76.5	30.8	0.998	0.582	82.5	82.4	0.883	0.597	47.5	42.8	75.5	72.9
MFN	**77.4**	**77.3**	**34.1**	**0.965**	**0.632**	**87.5**	**87.1**	**0.739**	**0.696**	**61.0**	**60.7**	**81.1**	**80.4**
Δ_{SOTA}	**↑ 2.8**	**↑ 2.8**	**↑ 0.9**	**↓ 0.054**	**↑ 0.010**	**↑ 6.2**	**↑ 7.5**	**↓ 0.103**	**↑ 0.108**	**↑ 10.8**	**↑ 9.9**	**↑ 7.1**	**↑ 5.7**

Dataset	IEMOCAP							
Task	Discrete Emotion		Valence		Arousal		Dominance	
Metric	MA(9)	F1	MAE	r	MAE	r	MAE	r
SOTA2	$35.9^{\dagger}$	$34.1^{\dagger}$	$0.248^{\dagger}$	$0.065^{\dagger}$	0.521^{*}	$0.617^{\S}$	0.671^{*}	$0.479^{\S}$
SOTA1	36.0^{*}	34.5^{*}	$0.244^{\S}$	$0.088^{\S}$	$0.513^{\diamond}$	$0.620^{\diamond}$	$0.668^{\diamond}$	**0.519**$^{\diamond}$
MFN l	25.8	16.1	0.250	-0.022	1.566	0.105	1.599	0.162
MFN a	22.5	11.6	0.279	0.034	1.924	0.447	1.848	0.417
MFN v	21.5	10.5	0.248	-0.014	2.073	0.155	2.059	0.083
MFN (no Δ)	34.8	33.1	0.243	0.098	0.500	0.590	0.629	0.466
MFN (no mem)	31.2	28.0	0.246	0.089	0.509	0.634	0.679	0.441
MFN	**36.5**	**34.9**	**0.236**	**0.111**	**0.482**	**0.645**	**0.612**	0.509
Δ_{SOTA}	**↑ 0.5**	**↑ 0.4**	**↓ 0.008**	**↑ 0.023**	**↓ 0.031**	**↑ 0.025**	**↓ 0.056**	**↓ 0.010**

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)
SOTA2	$26.6^{\bullet}$	$27.6^{\S}$	$32.0^{\heartsuit}$	$35.0^{\heartsuit}$	$26.1^{\flat}$	$32.0^{\flat}$	27.6^{*}	$29.6^{\flat}$	$34.0^{\heartsuit}$	$53.2^{\bullet}$	$49.8^{\heartsuit}$	$39.4^{\flat}$	$42.4^{\S}$	$42.4^{\flat}$	27.6^{*}	$36.5^{\dagger}$
SOTA1	$26.6^{\bullet}$	31.0^{*}	$33.0^{\flat}$	$35.0^{\heartsuit}$	$27.6^{\dagger}$	$36.5^{\dagger}$	$30.5^{\dagger}$	$31.5^{\heartsuit}$	$34.0^{\heartsuit}$	$53.7^{\flat}$	$50.7^{\diamond}$	$42.9^{\heartsuit}$	$45.8^{\dagger}$	$42.4^{\flat}$	$28.1^{\heartsuit}$	$40.4^{\bullet}$
MFN l	26.6	31.5	21.7	34.0	25.6	28.6	26.6	30.5	29.1	34.5	39.9	31.5	30.5	34.0	24.1	42.4
MFN a	27.1	26.1	29.6	34.5	24.6	29.6	26.6	31.0	32.5	35.0	45.8	37.4	35.0	40.4	28.1	36.5
MFN v	25.6	23.6	26.6	31.5	25.1	28.6	25.6	26.6	32.5	48.3	43.3	36.9	42.4	33.5	24.1	37.4
MFN (no Δ)	28.1	32.0	34.5	36.0	32.0	33.0	29.6	33.5	33.0	56.2	51.2	42.9	44.3	43.8	31.5	42.9
MFN (no mem)	26.1	27.1	34.5	35.5	28.1	31.0	27.1	30.0	32.0	55.2	50.7	39.4	42.9	42.4	29.1	33.5
MFN	**34.5**	**35.5**	**37.4**	**41.9**	**34.5**	**36.9**	**36.0**	**37.9**	**38.4**	**57.1**	**53.2**	**46.8**	**47.3**	**47.8**	**34.0**	**47.3**
Δ_{SOTA}	**↑ 7.9**	**↑ 4.5**	**↑ 4.4**	**↑ 6.9**	**↑ 6.9**	**↑ 0.4**	**↑ 5.5**	**↑ 6.4**	**↑ 4.4**	**↑ 3.4**	**↑ 2.5**	**↑ 3.9**	**↑ 1.5**	**↑ 5.4**	**↑ 5.9**	**↑ 6.9**

Dataset	POM															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	MAE															
SOTA2	$1.033^{\flat}$	$1.067^{\S}$	$0.911^{\S}$	0.864^{*}	$1.022^{\S}$	$0.981^{\S}$	$0.990^{\S}$	$0.967^{\flat}$	$0.884^{\flat}$	$0.556^{\S}$	$0.594^{\S}$	$0.700^{\S}$	$0.712^{\S}$	$0.705^{\dagger}$	$1.084^{\S}$	$0.768^{\flat}$
SOTA1	$1.016^{\dagger}$	$1.008^{\dagger}$	$0.899^{\flat}$	$0.859^{\dagger}$	$0.942^{\dagger}$	**0.905**†	$0.906^{\dagger}$	$0.927^{\dagger}$	$0.877^{\diamond}$	$0.523^{\diamond}$	$0.591^{\flat}$	$0.698^{\flat}$	$0.680^{\dagger}$	$0.687^{\diamond}$	$1.025^{\dagger}$	$0.767^{\dagger}$
MFN l	1.065	1.152	1.033	0.875	1.074	1.111	1.135	0.994	0.915	0.591	0.612	0.792	0.753	0.722	1.134	0.838
MFN a	1.086	1.147	0.937	0.887	1.104	1.028	1.075	1.009	0.882	0.589	0.611	0.719	0.759	0.697	1.159	0.783
MFN v	1.083	1.153	1.009	0.931	1.085	1.073	1.135	1.028	0.929	0.664	0.682	0.771	0.770	0.773	1.138	0.793
MFN (no Δ)	1.015	1.061	0.891	0.859	0.994	0.958	1.000	0.955	0.875	0.527	0.583	0.691	0.711	0.691	1.052	0.750
MFN (no mem)	1.018	1.077	0.887	0.865	1.014	0.995	1.012	0.959	0.877	0.530	0.581	0.701	0.719	0.694	1.063	0.764
MFN	**0.952**	**0.993**	**0.882**	**0.835**	**0.903**	0.908	**0.886**	**0.913**	**0.821**	**0.521**	**0.566**	**0.679**	**0.665**	**0.654**	**0.981**	**0.727**
Δ_{SOTA}	**↓ 0.064**	**↓ 0.015**	**↓ 0.017**	**↓ 0.024**	**↓ 0.039**	**↑ 0.003**	**↓ 0.020**	**↓ 0.014**	**↓ 0.056**	**↓ 0.002**	**↓ 0.025**	**↓ 0.019**	**↓ 0.015**	**↓ 0.033**	**↓ 0.044**	**↓ 0.040**

Dataset	POM															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	r															
SOTA2	$0.240^{\flat}$	$0.302^{\S}$	$0.031^{\S}$	$0.139^{\flat}$	$0.170^{\S}$	$0.244^{\S}$	$0.265^{\S}$	$0.240^{\S}$	$0.148^{\flat}$	$0.109^{\dagger}$	$0.083^{\S}$	$0.093^{\flat}$	$0.260^{\S}$	$0.136^{\flat}$	$0.217^{\S}$	$0.259^{\flat}$
SOTA1	$0.359^{\dagger}$	$0.425^{\dagger}$	$0.131^{\diamond}$	$0.234^{\dagger}$	$0.358^{\dagger}$	$0.417^{\dagger}$	$0.450^{\dagger}$	$0.361^{\dagger}$	$0.295^{\diamond}$	$0.237^{\diamond}$	$0.119^{\diamond}$	$0.238^{\diamond}$	$0.363^{\dagger}$	$0.258^{\diamond}$	$0.344^{\dagger}$	$0.319^{\dagger}$
MFN l	0.223	0.281	-0.013	0.118	0.141	0.189	0.188	0.227	-0.168	-0.064	0.126	0.095	0.173	0.024	0.183	0.216
MFN a	0.092	0.128	-0.019	0.050	0.021	-0.007	0.035	0.130	0.152	-0.071	0.019	-0.003	-0.019	0.106	0.024	0.064
MFN v	0.146	0.091	-0.077	-0.012	0.019	-0.035	0.012	0.038	-0.004	-0.169	0.030	-0.026	0.047	0.059	0.078	0.159
MFN (no Δ)	0.307	0.373	0.140	0.209	0.272	0.334	0.333	0.305	0.194	0.218	0.160	0.152	0.277	0.182	0.288	0.334
MFN (no mem)	0.259	0.261	0.166	0.109	0.161	0.188	0.209	0.247	0.189	0.059	0.151	0.115	0.161	0.134	0.190	0.231
MFN	**0.395**	**0.428**	**0.193**	**0.313**	**0.367**	**0.431**	**0.452**	**0.395**	**0.333**	**0.296**	**0.255**	**0.259**	**0.381**	**0.318**	**0.377**	**0.386**
Δ_{SOTA}	**↑ 0.036**	**↑ 0.003**	**↑ 0.062**	**↑ 0.079**	**↑ 0.009**	**↑ 0.014**	**↑ 0.002**	**↑ 0.034**	**↑ 0.038**	**↑ 0.059**	**↑ 0.136**	**↑ 0.021**	**↑ 0.018**	**↑ 0.060**	**↑ 0.033**	**↑ 0.067**

Table 2: Results for sentiment analysis on the CMU-MOSI, ICT-MMMO, YouTube and MOUD datasets, emotion recognition on the IEMOCAP dataset and personality trait recognition on the POM dataset. 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: $\#$: *Morency2007*, $\blacksquare$: *Song2012LD*, $\clubsuit$: *Poria2015*, $\heartsuit$: *Zadeh2016*, $\bullet$: *Ho1998*, $\flat$: *Nojavanasghari2016*, §: *Hochreiter1997*, $\diamond$: *Rajagopalan2016*, †: *Poria2017*, $\star$: *Zadeh2017*. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over SOTA1. Improvements are highlighted in green. The MFN significantly outperforms the SOTA across all datasets and metrics, except the Δ_{SOTA} entries highlighted in gray. For a detailed table with all baseline results, please refer to the Supplementary Materials.

Dataset Specific State-of-the-art Baselines

Poria2017: Bidirectional Contextual LSTM (Poria et al. 2017) is a model for context-dependent fusion of multi-sequence data that holds the state-of-the-art for emotion recognition on IEMOCAP dataset and sentiment analysis on MOUD dataset.

Zadeh2017: Tensor Fusion Network (Zadeh et al. 2017) learns explicit uni-view, bi-view and tri-view concepts in multi-view data. It is the current state-of-the-art for sentiment analysis on CMU-MOSI dataset.

Wang2016: Selective Additive Learning Convolutional Neural Network (Wang et al. 2016) is a multimodal sentiment analysis model that attempts to prevent identity-dependent information from being learned so as to improve generalization based only on accurate indicators of sentiment.

MFN Ablation Study Baselines

MFN {l,v,a}: These baselines use only one of the multiple present views – l stands for language, v for visual, and a for acoustic. The DMAN and Multi-view Gated Memory are also removed since only one view is present. This effectively reduces the MFN to one single LSTM which uses input from one view.

MFN (no Δ): This variation of our model shrinks the context to only the current timestamp t in the DMAN. We compare to this model to show the importance of context in learning relations between multi-view sequences.

MFN (no mem): This variation of our model removes the Delta-memory Attention Network and Multi-view Gated Memory from the MFN. Essentially this is equivalent to three disjoint LSTMs. The output of the MFN in this case would only be the outputs of LSTM at the final timestamp T. This baseline is designed to evaluate the importance of spatio-temporal relations between views through time.

Finally, we perform ablation studies by reporting the performance of the MFN without its major 2 components: the Multi-view Gated Memory and the context window of 2.

MFN Results and Discussion

Table 2 summarizes the comparison between MFN and proposed baselines for sentiment analysis, emotion recognition and speaker traits recognition. Different evaluation tasks are performed for different datasets based on the provided labels: binary classification, multi-class classification, and regression. For binary classification we report results in binary accuracy BA and binary F1 score. For multiclass classification we report multiclass accuracy MA(k) where k denotes the number of classes, and multiclass F1 score. For regression we report Mean Absolute Error (MAE) and Pearson's correlation r. Higher values denote better performance for all metrics. The only exception is MAE which lower values indicate better performance. All the baselines are trained for all the benchmarks using the same input data as MFN and best set of hyperparameters are chosen based on a validation set. The best performing baseline for each benchmark is referred to as SOTA1 (SOTA stands for state-of-the-art). SOTA2 is the second best performing model[1]. SOTA models change per different metrics since different models are suitable for different tasks. The superscript on each number indicates what method it belongs to. The performance improvement of our MFN over the SOTA1 model is denoted as Δ_{SOTA}, the raw improvement over the previous models. The results of our experiments can be summarized as follows:

MFN Achieves State of The Art Performance for Multi-view Sequential Modeling: our approach is able to significantly outperform all the proposed baselines in all the benchmarks, setting new state of the art in all the datasets. Furthermore, MFN shows consistent trend for both classification and regression. The same is not true for other baselines as their performance varies based on the dataset and evaluation task.

Ablation Studies: our comparison with variations of our model show a consistent trend:

$$\text{MFN} > \text{MFN (no } \Delta) > \text{MFN (no mem)} > \text{MFN } \{l, v, a\}$$

The comparison between MFN and MFN (no Δ) indicates the crucial role of the memories of time $t-1$. The comparison between MFN and MFN (no mem) shows the essential role of the Multi-view Gated Memory. Furthermore comparison between MFN (no Δ) and MFN (no mem) shows that even a Multi-view Gated Memory without $t-1$ is able to outperform the model that has no Multi-view Gated Memory. The final observation comes from comparing all multi-view variation of MFN with single view MFN $\{l, v, a\}$. This indicates that using multiple views results in better performance.

Increasing The DMAN Input Region Size: we extensively studied increasing the input region of DMAN to cover q steps before t instead of just 1 step. This set of experiments was aimed to find if there is an effect for increasing the context to q memories back in time. Results showed no improvement. [2]

sizes bigger than 2. At the same time a sharp drop in accuracy happens for context size of 1 (MFN no con in our baselines). This indicates that the model is learning We argue that this is due to the fact that a context size of 2 allows the memory to track changes and pull the related information from the memories only if the memory of the LSTMs is overriding it.

Conclusion

This paper introduced a novel approach for multi-view sequential learning called Memory Fusion Network (MFN). The first component of MFN is called System of LSTMs. In System of LSTMs, each view is assigned one LSTM function to model the interactions within the view. The second component of MFN is called Delta-memory Attention Network (DMAN). DMAN outlines the relations between views through time by associating a cross-view relevance score to the memory dimensions of each LSTM. The third component of the MFN unifies the sequences and is called Multi-view Gated Memory. This memory updates its content based on the outputs of DMAN and memories in System of LSTMs. Through extensive experimentation on multiple publicly available datasets, the performance of MFN is compared with various baselines. MFN shows state-of-the-art performance in multi-view sequential learning.

[1] for complete comparison with all models please refer to the supplementary material

[2] please refer to supplementary material for results

References

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

Alku, P.; Strik, H.; and Vilkman, E. 1997. Parabolic spectral parametera new method for quantification of the glottal flow. *Speech Communication* 22(1):67–79.

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

Bach, F. R.; Lanckriet, G. R.; and Jordan, M. I. 2004. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, 6. ACM.

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

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42(4):335–359.

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

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 960–964. IEEE.

Drugman, T., and Alwan, A. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, 1973–1976.

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

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

Gunes, H. 2010. Automatic, dimensional and continuous emotion recognition.

Ho, T. K. 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* 20(8):832–844.

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

iMotions. 2017. Facial expression analysis.

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

Morency, L.-P.; Mihalcea, R.; and Doshi, P. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, 169–176. ACM.

Morency, L.-P.; Quattoni, A.; and Darrell, T. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8. IEEE.

Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; and Morency, L.-P. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, 284–288. New York, NY, USA: ACM.

Park, S.; Shim, H. S.; Chatterjee, M.; Sagae, K.; and Morency, L.-P. 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*, ICMI '14, 50–57. New York, NY, USA: ACM.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation.

Perez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*.

Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *Association for Computational Linguistics*.

Poria, S.; Cambria, E.; and Gelbukh, A. F. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis.

Quattoni, A.; Wang, S.; Morency, L.-P.; Collins, M.; and Darrell, T. 2007. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(10):1848–1852.

Rajagopalan, S. S.; Morency, L.-P.; Baltrušaitis, T.; and Goecke, R. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

Song, Y.; Morency, L.-P.; and Davis, R. 2012. Multi-view latent variable discriminative models for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2120–2127. IEEE.

Song, Y.; Morency, L.-P.; and Davis, R. 2013. Action recognition by hierarchical sequence summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3562–3569.

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

Wang, H.; Meghawat, A.; Morency, L.-P.; and Xing, E. P. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*.

Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; and Morency, L.-P. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28(3):46–53.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

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

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*.

Zhu, Q.; Yeh, M.-C.; Cheng, K.-T.; and Avidan, S. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, 1491–1498. IEEE.