The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors

Yansen Wang,[1] Ying Shen,[2] Zhun Liu,[2] Paul Pu Liang,[2] Amir Zadeh,[2] Louis-Philippe Morency[2]
[1]Department of Computer Science, Tsinghua University
[2]School of Computer Science, Carnegie Mellon University
ys-wang15@mails.tsinghua.edu.cn; {yshen2,zhunl,pliang,abagherz,morency}@cs.cmu.edu

Abstract

Humans convey their intentions through the usage of both verbal and nonverbal behaviors during face-to-face communication. Speaker intentions often vary dynamically depending on different nonverbal contexts, such as vocal patterns and facial expressions. As a result, when modeling human language, it is essential to not only consider the literal meaning of the words but also the nonverbal contexts in which these words appear. To better model human language, we first model expressive nonverbal representations by analyzing the fine-grained visual and acoustic patterns that occur during word segments. In addition, we seek to capture the dynamic nature of nonverbal intents by shifting word representations based on the accompanying nonverbal behaviors. To this end, we propose the Recurrent Attended Variation Embedding Network (RAVEN) that models the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues. Our proposed model achieves competitive performance on two publicly available datasets for multimodal sentiment analysis and emotion recognition. We also visualize the shifted word representations in different nonverbal contexts and summarize common patterns regarding multimodal variations of word representations.

Introduction

Multimodal language communication happens through both verbal and nonverbal channels. The verbal channel of communication conveys intentions through words and sentences while the nonverbal aspect uses gestures and vocal intonations. However, the meaning of words and sentences uttered by the speaker often varies dynamically in different nonverbal contexts. These dynamic behaviors can arise from different sources such as cultural shift or different political backgrounds (Bamler and Mandt 2017). In human multimodal language, these dynamic behaviors are often intertwined with their nonverbal contexts (Burgoon, Guerrero, and Floyd 2016). Intentions conveyed through uttering a sentence can display drastic shifts in intensity and direction, leading to the phenomena that the uttered words exhibit dynamic meanings depending on different nonverbal contexts.

Previous work in modeling human language often utilizes word embeddings pretrained on a large textual corpus to represent the meaning of language. However, these methods are not sufficient for modeling highly dynamic human multimodal language. The example in Figure 1 demonstrates how the same underlying word can vary in sentiment when paired with different nonverbal cues. Although the two speakers are using the same adjective "sick" to describe movies, they are conveying different sentiments and remarks by showing opposing facial expressions and intonations. These subtle nonverbal cues contained in the span of the uttered words, including facial expressions, facial landmarks, and acoustic features, are crucial towards determining the exact intent displayed in verbal language. We hypothesize that this "exact intent" can often be derived from the representation of the uttered words combined with a shift in the embedding space introduced by the accompanying nonverbal cues. In this regard, a dynamic representation for words in particular visual and acoustic background is required.

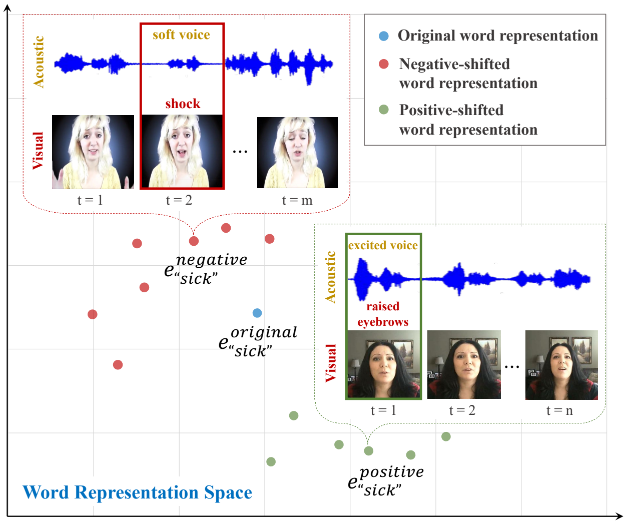

Figure 1: Conceptual figure demonstrating that the word representation of the same underlying word "sick" can vary conditioned on different co-occurring nonverbal behaviors. The nonverbal context during a word segment is depicted by the sequence of facial expressions and intonations. The speaker who has a relatively soft voice and frowning behaviors at the second time step displays negative sentiment.

Modeling the nonverbal contexts concurrent to an uttered word requires fine-grained analysis. This is because the visual and acoustic behaviors often have a much higher temporal frequency than words, leading to a sequence of accompanying visual and acoustic "subword" units for each uttered word. The structure of these subword sequences is especially important towards the representation of nonverbal dynamics. In addition, modeling subword information has become essential for various tasks in natural language processing (Faruqui et al. 2017), including language modeling (Labeau and Allauzen 2017; Kim et al. 2016), learning word representations for different languages (Peters et al. 2018; Oh et al. 2018; Bojanowski et al. 2016), and machine translation (Kudo 2018; Sennrich, Haddow, and Birch 2015). However, many of these previous works in understanding and modeling multimodal language has ignored the role of subword analysis. Instead, they summarize the subword information during each word span using the simple averaging strategies (Liang et al. 2018; Liu et al. 2018; Zadeh et al. 2018b). While average behaviors may be helpful in modeling global characteristics, it is lacking in its representation capacity to accurately model the structure of nonverbal behaviors at the subword level. This motivates the design of a more expressive model that can accurately capture the fine-grained visual and acoustic patterns that occur in the duration of each word.

To this end, we propose the Recurrent Attended Variation Embedding Network (RAVEN), a model for human multimodal language that considers the fine-grained structure of nonverbal subword sequences and dynamically shifts the word representations based on these nonverbal cues. In order to verify our hypotheses on the importance of subword analysis as well as the dynamic behaviors of word meanings, we conduct experiments on multimodal sentiment analysis and emotion recognition. Our model shows excellent performance on both tasks. We present visualizations of the shifted word representations to better understand the impact of subword modeling and dynamic shifts on modeling word meaning. Finally, we present ablation studies to analyze the effects of subword modeling and dynamic shifting. We discover that the shifted embeddings learned by RAVEN exhibit meaningful distributional patterns with respect to the sentiment expressed by the speaker.

Related Works

Previously, much effort has been devoted to building machine learning models that learn from multiple modalities (Ngiam et al. 2011; Srivastava and Salakhutdinov 2014). However, there has been limited research into modeling the variations of word representations using nonverbal behaviors. To place our work in the context of prior research, we categorize previous works as follows: (1) subword word representations, (2) modeling variations in word representations, and (3) multimodal sentiment and emotion recognition.

Modeling subword information has become crucial for various tasks in natural language processing (Faruqui et al. 2017). Learning the compositional representations from subwords to words allows models to infer representations for words not in the training vocabulary. This has proved especially useful for machine translation (Sennrich, Haddow, and Birch 2015), language modeling (Kim et al. 2016) and word representation learning (Bojanowski et al. 2016). In addition, deep word representations learned via neural models with character convolutions (Zhang, Zhao, and LeCun 2015) have been found to contain highly transferable language information for downstream tasks such as question answering, textual entailment, sentiment analysis, and natural language inference (Peters et al. 2018).

Modeling variations in word representations is an important research area since many words have different meanings when they appear in different contexts. Li and Jurafsky (2015) propose a probabilistic method based on Bayesian Nonparametric models to learn different word representations for each sense of a word, Nguyen et al. (2017) use a Gaussian Mixture Model (Reynolds 2009) and Athiwaratkun, Wilson, and Anandkumar (2018) extend FastText word representations (Bojanowski et al. 2016) with a Gaussian Mixture Model representation for each word.

Prior work in *multimodal sentiment and emotion recognition* has tackled the problem via multiple approaches: the early fusion method refers to concatenating multimodal data at the input level. While these methods are able to outperform unimodal models (Zadeh et al. 2016) and learn robust representations (Wang et al. 2016), they have limited capabilities in learning modality-specific interactions and tend to overfit (Xu, Tao, and Xu 2013). The late fusion method integrates different modalities at the prediction level. These models are highly modular, and one can build a multimodal model from individual pre-trained unimodal models and fine-tuning on the output layer (Poria et al. 2017). While such models can also outperform unimodal models (Pham et al. 2018), they focus mostly on modeling modality-specific interactions rather than cross-modal interactions. Finally, multi-view learning refers to a broader class of methods that perform fusion between the input and prediction levels. Such methods usually perform fusion throughout the multimodal sequence (Rajagopalan et al. 2016; Liang, Zadeh, and Morency 2018), leading to explicit modeling of both modality-specific and cross-modal interactions at every time step. Currently, the best results are achieved by augmenting this class of models with attention mechanisms (Liang et al. 2018), word-level alignment (Tsai et al. 2018), and more expressive fusion methods (Liu et al. 2018).

These previous studies have explored integrating nonverbal behaviors or building word representations with different variations from purely textual data. However, these works do not consider the temporal interactions between the nonverbal modalities that accompany the language modality at the subword level, as well as the contribution of non-verbal behaviors towards the meaning of underlying words. Our proposed method models the nonverbal temporal interactions between the subword units. This is performed by word-level fusion with nonverbal features introducing variations to word representations. In addition, our work can also be seen as an extension of the research performed in modeling multi-sense word representations. We use the accompanying nonverbal behaviors to learn variation vectors that either (1) disambiguate or (2) emphasize the existing word representations for multimodal prediction tasks.

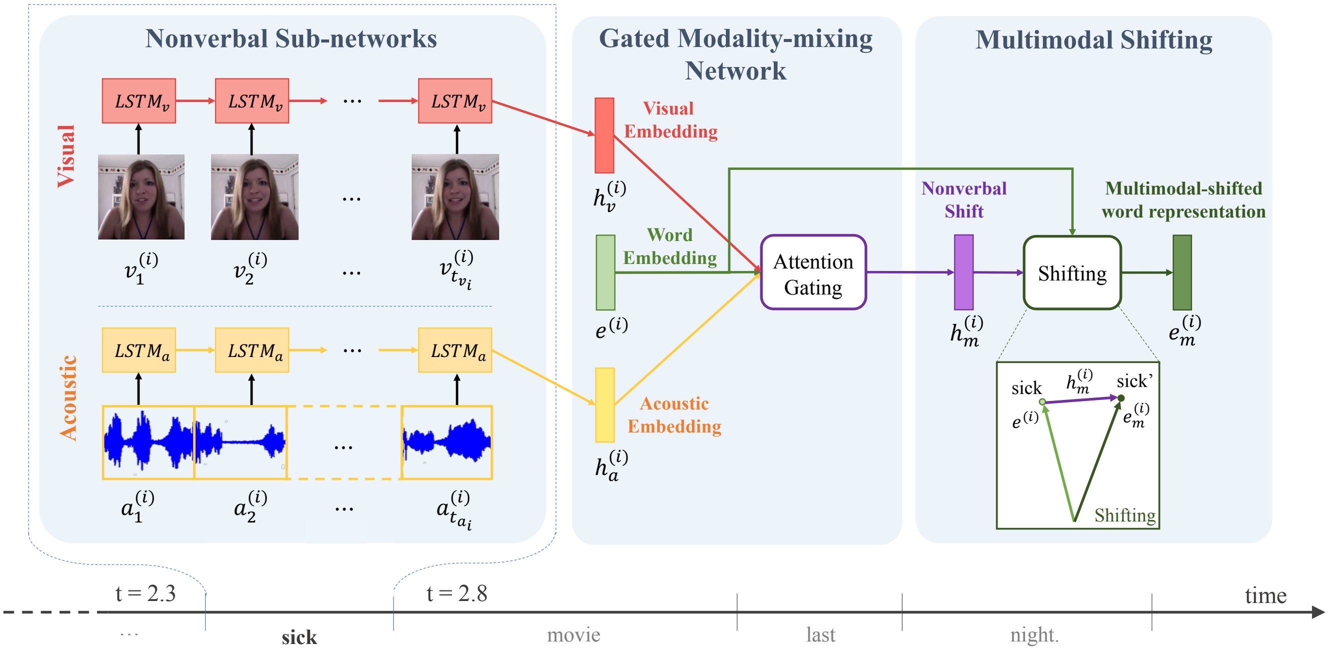

Figure 2: An illustrative example for Recurrent Attended Variation Embedding Network (RAVEN) model: The RAVEN model has three components: (1) Nonverbal Sub-networks, (2) Gated Modality-mixing Network, and (3) Multimodal Shifting. For the given word "sick" in the utterance, the Nonverbal Sub-networks first computes the visual and acoustic embedding through modeling the sequence of visual and acoustic features lying in a word-long segment with separate LSTM network. The Gated Modality-mixing Network module then infers the nonverbal shift vector as the weighted average over the visual and acoustic embedding based on the original word embedding. The Multimodal Shifting finally generates the multimodal-shifted word representation by integrating the nonverbal shift vector to the original word embedding. The multimodal-shifted word representation can be then used in the high-level hierarchy to predict sentiments or emotions expressed in the sentence.

Recurrent Attended Variation Embedding Network (RAVEN)

The goal of our work is to better model multimodal human language by (1) considering subword structure of nonverbal behaviors and (2) learning multimodal-shifted word representations conditioned on the occurring nonverbal behaviors. To achieve this goal, we propose the Recurrent Attended Variation Embedding Network (RAVEN).

An overview of the proposed RAVEN model is given in Figure 2. Our model consists of three major components: (1) *Nonverbal Sub-networks* model the fine-grained structure of nonverbal behaviors at the subword level by using two separate recurrent neural networks to encode a sequence of visual and acoustic patterns within a word-long segment, and outputs the nonverbal embeddings. (2) *Gated Modality-mixing Network* takes as input the original word embedding as well as the visual and acoustic embedding, and uses an attention gating mechanism to yield the nonverbal shift vector which characterizes how far and in which direction has the meaning of the word changed due to nonverbal context. (3) *Multimodal Shifting* computes the multimodal-shifted word representation by integrating the nonverbal shift vector to the original word embedding. The following subsections discuss the details of these three components of our RAVEN model.

Nonverbal Sub-networks

To better model the subword structure of nonverbal behaviors, the proposed Nonverbal Sub-networks operate on the visual and acoustic subword units carried alongside each word. This yields the visual and acoustic embeddings. These output embeddings are illustrated in Figure 2.

Formally, we begin with a segment of multimodal language $\mathbf{L}$ denoting the sequence of uttered words. For the span of the ith word denoted as $\mathbf{L}^{(i)}$, we have two accompanying sequences from the visual and acoustic modalities: $\mathbf{V}^{(i)} = [v_1^{(i)}, v_2^{(i)}, \cdots, v_{t_{v_i}}^{(i)}]$, $\mathbf{A}^{(i)} = [a_1^{(i)}, a_2^{(i)}, \cdots, a_{t_{a_i}}^{(i)}]$. These are temporal sequences of visual and acoustic frames, to which we refer as the visual and acoustic subword units. To model the temporal sequences of sub-word information coming from each modality and compute the nonverbal embeddings, we use Long-short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) networks. LSTMs have been successfully used in modeling temporal data in both computer vision (Ullah et al. 2018) and acoustic signal processing (Hughes and Mierle 2013).

The modality-specific LSTMs are applied to the sub-word sequences for each word $\mathbf{L}^{(i)}, i = 1, \cdots, n$. For the i-th word $\mathbf{L}^{(i)}$ in the language modality, two LSTMs are applied separately for its underlying visual and acoustic sequences:

$$\mathbf{h}_v^{(i)} = \texttt{LSTM}_v(\mathbf{V}^{(i)}) \tag{1}$$

$$\mathbf{h}_a^{(i)} = \texttt{LSTM}_a(\mathbf{A}^{(i)}) \tag{2}$$

where $\mathbf{h}_v^{(i)}$ and $\mathbf{h}_a^{(i)}$ refer to the final states of the visual and acoustic LSTMs. We call these final states visual and acoustic embedding, respectively.

Gated Modality-mixing Network

Our Gated Modality-mixing Network component computes the nonverbal shift vector by learning a non-linear combination between the visual and acoustic embedding using an attention gating mechanism. Our key insight is that depending on the information in visual and acoustic modalities as well as the word that is being uttered, the relative importance of the visual and acoustic embedding may differ. For example, the visual modality may demonstrate a high activation of facial muscles showing shock while the tone in speech may be uninformative. To handle these dynamic dependencies, we propose a gating mechanism that controls the importance of each visual and acoustic embedding.

In order for the model to control how strong a modality's influence is, we use modality-specific influence gates to model the intensity of the influence. To be more concrete, for word $\mathbf{L}^{(i)}$, given the original word representation $\mathbf{e}^{(i)}$, we concatenate $\mathbf{e}^{(i)}$ with the visual and acoustic embedding $\mathbf{h}_v^{(i)}$ and $\mathbf{h}_a^{(i)}$ respectively and then use the concatenated vectors as the inputs of the visual and acoustic gate $w_v^{(i)}$ and $w_a^{(i)}$:

$$w_v^{(i)} = \sigma(\mathbf{W}_{hv}[\mathbf{h}_v^{(i)}; \mathbf{e}^{(i)}] + b_v) \tag{3}$$

$$w_a^{(i)} = \sigma(\mathbf{W}_{ha}[\mathbf{h}_a^{(i)}; \mathbf{e}^{(i)}] + b_a) \tag{4}$$

where $[;]$ denotes the operation of vector concatenation. $\mathbf{W}_{hv}$ and $\mathbf{W}_{ha}$ are weight vectors for the visual and acoustic gates and b_v and b_a are scalar biases. The sigmoid function $\sigma(x)$ is defined as $\sigma(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$.

Then a nonverbal shift vector is calculated by fusing the visual and acoustic embeddings multiplied by the visual and acoustic gates. Specifically, for a word $\mathbf{L}^{(i)}$, the nonverbal shift vector $\mathbf{h}_m^{(i)}$ is calculated as follows:

$$\mathbf{h}_m^{(i)} = w_v^{(i)} \cdot (\mathbf{W}_v \mathbf{h}_v^{(i)}) + w_a^{(i)} \cdot (\mathbf{W}_a \mathbf{h}_a^{(i)}) + \mathbf{b}_h^{(i)} \tag{5}$$

where $\mathbf{W}_v$ and $\mathbf{W}_a$ are weight matrices for the visual and acoustic embedding and $\mathbf{b}_h^{(i)}$ is the bias vector.

Multimodal Shifting

The Multimodal Shifting component learns to dynamically shift the word representations by integrating the nonverbal shift vector $\mathbf{h}_m^{(i)}$ into the original word embedding. Concretely, the multimodal-shifted word representation for word $\mathbf{L}^{(i)}$ is given by:

$$\mathbf{e}_m^{(i)} = \mathbf{e}^{(i)} + \alpha \mathbf{h}_m^{(i)} \tag{6}$$

$$\alpha = \min\left(\frac{\|\mathbf{e}^{(i)}\|_2}{\|\mathbf{h}_m^{(i)}\|_2}\beta, 1\right) \tag{7}$$

where β is a threshold hyper-parameter which can be determined by cross-validation on a validation set.

In order to ensure the magnitude of the nonverbal shift vector $\mathbf{h}_m^{(i)}$ is not too large as compared to the original word embedding $\mathbf{e}^{(i)}$, we apply a scaling factor α to constrain the magnitude of the nonverbal shift vector to be within a desirable range. At the same time, the scaling factor maintains the direction of the shift vector.

By applying the same method for every word in $\mathbf{L}$, we can transform the original sequence triplet $(\mathbf{L}, \mathbf{V}, \mathbf{A})$ into one sequence of multimodal-shifted representations $\mathbf{E} = [\mathbf{e}_m^{(1)}, \mathbf{e}_m^{(2)}, \cdots, \mathbf{e}_m^{(n)}]$. The new sequence $\mathbf{E}$ now corresponds to a shifted version of the original sequence of word representations $\mathbf{L}$ fused with information from its accompanying nonverbal contexts.

This sequence of multimodal-shifted word representations is then used in the high-level hierarchy to predict sentiments or emotions expressed in the utterance. We can use a simple word-level LSTM to encode a sequence of the multimodal-shifted word representations into an utterance-level multimodal representation $\mathbf{h}$. This multimodal representation can then be used for downstream tasks:

$$\mathbf{h} = \texttt{LSTM}_e(\mathbf{E}) \tag{8}$$

For concrete tasks, the representation $\mathbf{h}$ is passed into a fully-connected layer to produce an output that fits the task. The various components of RAVEN are trained end-to-end together using gradient descent.

Experiments

In this section, we describe the experiments designed to evaluate our RAVEN model. We start by introducing the tasks and datasets and then move on to the feature extraction scheme. [1]

Datasets

To evaluate our approach, we use two multimodal datasets involving tri-modal human communications: CMU-MOSI (Zadeh et al. 2016) and IEMOCAP (Busso et al. 2008), for multimodal sentiment analysis and emotion recognition tasks, respectively.

Multimodal Sentiment Analysis: we first evaluate our approach for multimodal sentiment analysis. For this task, we choose the CMU-MOSI dataset. It comprises 2199 short video segments excerpted from 93 Youtube movie review videos and has real-valued sentiment intensity annotations from $[-3, +3]$. Negative values indicate negative sentiments and vice versa.

Multimodal Emotion Recognition: we investigate the performance of our model under a different, dyadic conversational environment for emotion recognition. The IEMOCAP dataset we use for this task contains 151 videos about dyadic interactions, where professional actors are required to perform scripted scenes that elicit specific emotions. Annotations for 9 different emotions are present (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral).

[1] The codes are available at https://github.com/victorywys/RAVEN.

Evaluation Metrics: since the multimodal sentiment analysis task can be formulated as a regression problem, we evaluate the performance in terms of Mean-absolute Error (MAE) as well as the correlation of model predictions with true labels. On top of that, we also follow the convention of the CMU-MOSI dataset, and threshold the regression values to obtain a categorical output and evaluate the performance in terms of classification accuracy. As for the multimodal emotion recognition, the labels for every emotion are binary so we evaluate it in terms of accuracy and F1 score.

Unimodal Feature Representations

Following prior practice (Liu et al. 2018; Liang et al. 2018; Gu et al. 2018), we adopted the same feature extraction scheme for language, visual and acoustic modalities.

Language Features: we use the GloVe vectors from (Pennington, Socher, and Manning 2014). In our experiments, we used the 300-dimensional version trained on 840B tokens[2].

Visual Features: given that the two multimodal tasks all include a video clip with the speakers' facial expressions, we employ the facial expression analysis toolkit FACET[3] as our visual feature extractor. It extracts features including facial landmarks, action units, gaze tracking, head pose and HOG features at the frequency of 30Hz.

Acoustic Features: we use the COVAREP (Degottex et al. 2014) acoustic analysis framework for feature extraction. It includes 74 features for pitch tracking, speech polarity, glottal closure instants, spectral envelope. These features are extracted at the frequency of 100Hz.

Baseline Models

Our proposed Recurrent Attended Variation Embedding Network (RAVEN) is compared to the following baselines and state-of-the-art models in multimodal sentiment analysis and emotion recognition.

Support Vector Machines (SVMs) (Cortes and Vapnik 1995) are widely used non-neural classifiers. This baseline is trained on the concatenated multimodal features for classification or regression tasks (Pérez-Rosas, Mihalcea, and Morency 2013; Park et al. 2014; Zadeh et al. 2016).

Deep Fusion (DF) (Nojavanasghari et al. 2016) performs late fusion by training one deep neural model for each modality and then combining the output of each modality network with a joint neural network.

Bidirectional Contextual LSTM (BC-LSTM) (Poria et al. 2017) performs context-dependent fusion of multimodal data.

Multi-View LSTM (MV-LSTM) (Rajagopalan et al. 2016) partitions the memory cell and the gates inside an LSTM corresponding to multiple modalities in order to capture both modality-specific and cross-modal interactions.

Multi-attention Recurrent Network (MARN) (Zadeh et al. 2018b) explicitly models interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory called the Long-short Term Hybrid Memory (LSTHM).

Dataset	**CMU-MOSI**		
Metric	MAE	Corr	Acc-2
SVM	1.864	0.057	50.2
DF	1.143	0.518	72.3
BC-LSTM	1.079	0.581	73.9
MV-LSTM	1.019	0.601	73.9
MARN	0.968	0.625	77.1
MFN	0.965	0.632	$77.4^{\ddagger}$
RMFN	$0.922^{\ddagger}$	$0.681^{\dagger}$	$78.4^{\star}$
LMF	$0.912^{\star}$	$0.668^{\ddagger}$	76.4
RAVEN	$0.915^{\dagger}$	$0.691^{\star}$	$78.0^{\dagger}$

Table 1: Sentiment prediction results on the CMU-MOSI test set using multimodal methods. The best three results are noted with *, † and ‡ successively.

Memory Fusion Network (MFN) (Zadeh et al. 2018a) continuously models the view-specific and cross-view interactions through time with a special attention mechanism and summarized through time with a Multi-view Gated Memory.

Recurrent Multistage Fusion Network (RMFN) (Liang et al. 2018) decomposes the fusion problem into multiple stages to model temporal, intra-modal and cross-modal interactions.

Low-rank Multimodal Fusion (LMF) model (Liu et al. 2018) learns both modality-specific and cross-modal interactions by performing efficient multimodal fusion with modality-specific low-rank factors.

Results and Discussion

In this section, we present results for the aforementioned experiments and compare our performance with state-of-the-art models. We also visualize the multimodal-shifted representations and show that they form interpretable patterns. Finally, to gain a better understanding of the importance of subword analysis and multimodal shift, we perform ablation studies on our model by progressively removing Nonverbal Subnetworks and Multimodal Shifting from our model, and find that the presence of both is critical for good performance.

Comparison with the State of the Art

We present our results on the multimodal datasets in Tables 1 and 2. Our model shows competitive performance when compared with state-of-the-art models across multiple metrics and tasks. Note that our model uses only a simple LSTM for making predictions. This model can easily be enhanced with more advanced modules such as temporal attention.

Multimodal Sentiment Analysis: On the multimodal sentiment prediction task, RAVEN achieves comparable performance to previous state-of-the-art models as shown in Table 1. Note the multiclass accuracy *Acc-7* is calculated by mapping the range of continuous sentiment values into a set of intervals that are used as discrete classes.

Multimodal Emotion Recognition: On the multimodal emotion recognition task, the performance of our model is also competitive compared to previous ones across all emotions on both the accuracy and F1 score.

[2] https://nlp.stanford.edu/projects/glove/

[3] https://imotions.com/

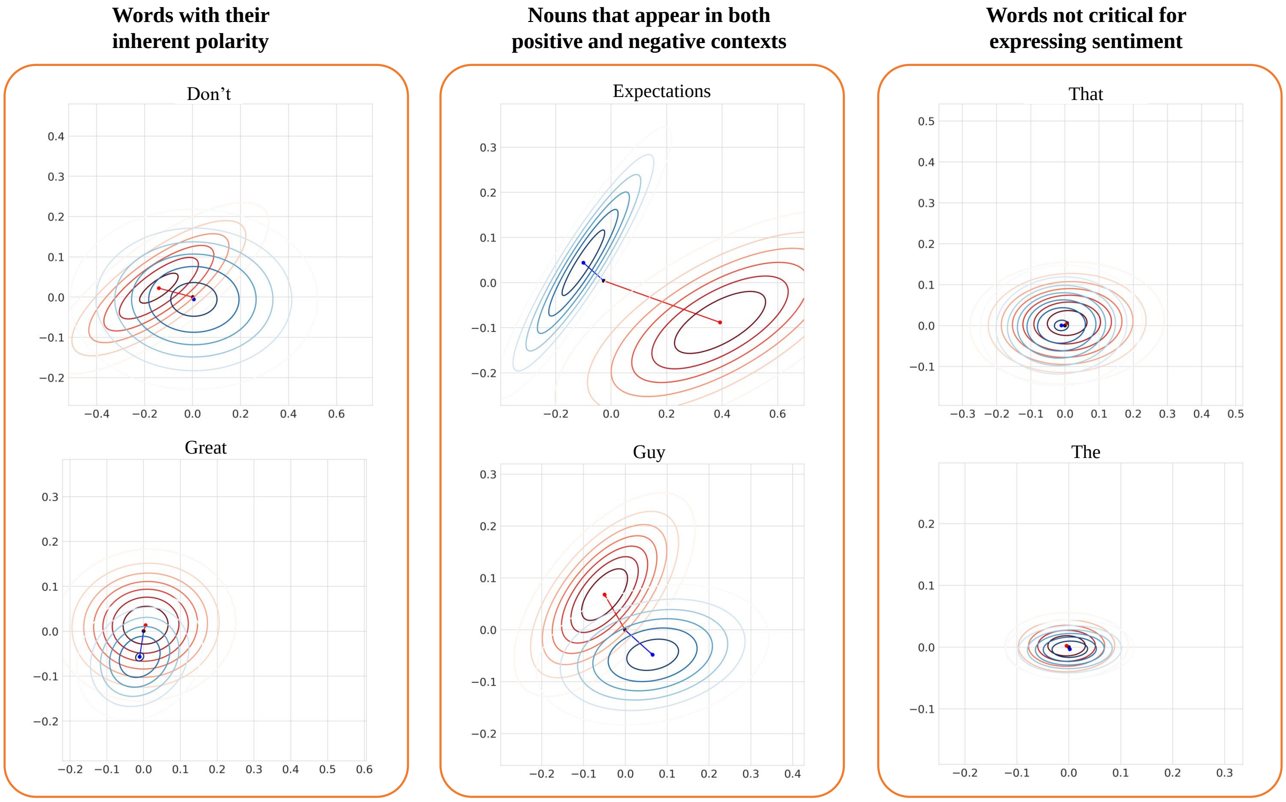

Figure 3: The Gaussian contours of shifted embeddings in 2-dimensional space. Three types of patterns observed in the distribution of all instances of the same word type: words with their inherent polarity will need a drastic variation to convey opposite sentiment; nouns that can appear in both positive and negative contexts will have large variations in both cases; words not critical for expressing sentiment minimal variations in both positive and negative contexts and the distribution of positive/negative instances significantly overlap.

Dataset	**IEMOCAP Emotions**							
Task	Happy		Sad		Angry		Neutral	
Metric	Acc-2	F1	Acc-2	F1	Acc-2	F1	Acc-2	F1
SVM	86.1	81.5	81.1	78.8	82.5	82.4	65.2	64.9
DF	86.0	81.0	81.8	81.2	75.8	65.4	59.1	44.0
BC-LSTM	84.9	81.7	83.2	81.7	83.5	84.2	67.5	64.1
MV-LSTM	85.9	81.3	80.4	74.0	85.1‡	84.3‡	67.0	66.7
MARN	86.7	83.6	82.0	81.2	84.6	84.2	66.8	65.9
MFN	86.5	84.0	83.5†	82.1	85.0	83.7	69.6‡	69.2‡
RMFN	87.5*	85.8*	82.9	85.1†	84.6	84.2	69.5	69.1
LMF	87.3†	85.8*	86.2*	85.9*	89.0*	89.0*	72.4*	71.7*
RAVEN	87.3†	85.8*	83.4‡	83.1‡	87.3†	86.7†	69.7†	69.3†

Table 2: Emotion recognition results on IEMOCAP test set using multimodal methods. The best three results are noted with *, † and ‡ successively.

Multimodal Representations in Different Nonverbal Contexts

As our model learns shifted representations by integrating each word with its accompanying nonverbal contexts, every instance of the same word will have a different multimodal-shifted representation. We observe that the shifts across all instances of the same word often exhibit consistent patterns. Using the CMU-MOSI dataset, we visualize the distribution of shifted word representations that belong to the same word. These visualizations are shown in Figure 3. We begin by projecting each word representation into 2-dimensional space using PCA (Jolliffe 2011). For each word, we plot Gaussian contours for the occurrences in positive-sentiment contexts and the occurrences in negative-sentiment contexts individually. Finally, we plot the centroid of all occurrences as well as the centroids of the subset in positive/negative contexts. To highlight the relative positions of these centroids, we add blue and red arrows starting from that overall centroid and pointing towards the positive and negative centroids. We discover that the variations of different words can be categorized into the following three different patterns depending on their roles in expressing sentiment in a multimodal context:

(1) For *words with their inherent polarity*, their instances in the opposite sentiment context often have strong variations that pull them away from the overall centroid. On the other hand, their instances in their default sentiment context usually experience minimal variations and are close to the overall

centroid. In Figure 3, the word "great" has an overall centroid that is very close to its positive centroid, while its negative centroid is quite far from both overall and positive centroids.

(2) For *nouns that appear in both positive and negative contexts*, both of their positive and negative centroids are quite far away from the overall centroid, and their positive and negative instances usually occupy different half-planes. While such nouns often refer to entities without obvious polarity in sentiment, our model learns to "polarize" these representations based on the accompanying multimodal context. For example, the noun "guy" is frequently used for addressing both good and bad actors, and RAVEN is able to shift them accordingly in the word embedding space towards two different directions (Figure 3).

(3) For *words that are not critical in conveying sentiment* (e.g. stop words), their average variations under both positive and negative contexts are minimal. This results in their positive, negative, and overall centroids all lying close to each other. Two example words that fall under this category are "that" and "the" with their centroids shown in Figure 3.

These patterns show that RAVEN is able to learn meaningful and consistent shifts for word representations to capture their dynamically changing meanings.

Ablation studies

RAVEN consists of three main components for performing multimodal fusion: Nonverbal Sub-networks, Gated Modality-mixing Network and Multimodal Shifting. Among these modules, Nonverbal Sub-networks and Multimodal Shifting are explicitly designed to model the subtle structures in non-verbal behaviors and to introduce dynamic variations to the underlying word representations. In order to demonstrate the necessity of these components in modeling multimodal language, we conducted several ablation studies to examine the impact of each component. We start with our full model and progressively remove different components. The different versions of the model are explained as follows:

RAVEN: our proposed model that models subword dynamics and dynamically shifts word embeddings.

RAVEN w/o SUB: our model without the Nonverbal Sub-networks. In this case, the visual and acoustic sequences are averaged into a vector representation, hence the capability of subword modeling is disabled.

RAVEN w/o SHIFT: our model without Multimodal Shifting. Visual and acoustic representations are concatenated with the word embedding before being fed to downstream networks. While this also generates a representation associated with the underlying word, it is closer to a multimodal representation projected into a different space. This does not guarantee that the new representation is a dynamically-varied embedding in the original word embedding space.

RAVEN w/o SUB&SHIFT: our model with both Nonverbal Sub-networks and Multimodal Shifting removed. This leads to a simple early-fusion model where the visual and acoustic sequences are averaged into word-level representations and concatenated with the word embeddings. It loses both the capabilities of modeling subword structures and creating dynamically-adjusted word embeddings.

Dataset	**CMU-MOSI**		
Metric	MAE	Corr	Acc-2
RAVEN	**0.915**	**0.691**	**78.0**
RAVEN w/o SHIFT	0.954	0.666	77.7
RAVEN w/o SUB	0.934	0.652	73.9
RAVEN w/o SUB&SHIFT	1.423	0.116	50.6

Table 3: Ablation studies on CMU-MOSI dataset. The complete RAVEN that models subword dynamics and word shifts works best.

Table 3 shows the results of ablation studies using several different variants of our model. The results show that both Nonverbal Sub-networks and Multimodal Shifting components are necessary for achieving state-of-the-art performance. This further implies that in scenarios where the visual and acoustic modalities are sequences extracted at a higher frequency, the crude averaging method for sub-sampling them to the same frequency of words does hurt performance. Another observation is that given neural networks are universal function approximators (Csáji 2001), the early-fusion model, in theory, is the most flexible model. Yet in practice, our model improves upon the early-fusion model. This implies that our model does successfully capture underlying structures of human multimodal language.

Conclusion

In this paper, we presented the Recurrent Attended Variation Embedding Network (RAVEN). RAVEN models the fine-grained structure of nonverbal behaviors at the subword level and builds multimodal-shifted word representations that dynamically captures the variations in different nonverbal contexts. RAVEN achieves competitive results on well-established tasks in multimodal language including sentiment analysis and emotion recognition. Furthermore, we demonstrate the importance of both subword analysis and dynamic shifts in achieving improved performance via ablation studies on different components of our model. Finally, we also visualize the shifted word representations in different nonverbal contexts and summarize several common patterns regarding multimodal variations of word representations. This illustrates that our model successfully captures meaningful dynamic shifts in the word representation space given nonverbal contexts. For future work, we will explore the effect of dynamic word representations towards other multimodal tasks involving language and speech (prosody), videos with multiple speakers (diarization), and combinations of static and temporal data (i.e. image captioning).

Acknowledgments

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

References

Athiwaratkun, B.; Wilson, A.; and Anandkumar, A. 2018. Probabilistic fasttext for multi-sense word embeddings. In *ACL*.

Bamler, R., and Mandt, S. 2017. Dynamic word embeddings. *arXiv preprint arXiv:1702.08359*.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *CoRR* abs/1607.04606.

Burgoon, J. K.; Guerrero, L. K.; and Floyd, K. 2016. *Nonverbal communication*. Routledge.

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.

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

Csáji, B. C. 2001. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary* 24:48.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. Covarep–a collaborative voice analysis repository for speech technologies. In *IEEE ICASSP*, 960–964. IEEE.

Faruqui, M.; Schuetze, H.; Trancoso, I.; and Yaghoobzadeh, Y., eds. 2017. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. ACL.

Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; and Marsic, I. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL*, 2225–2235. Association for Computational Linguistics.

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

Hughes, T., and Mierle, K. 2013. Recurrent neural networks for voice activity detection. In *IEEE ICASSP*, 7378–7382. IEEE.

Jolliffe, I. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*. Springer. 1094–1096.

Kim, Y.; Jernite, Y.; Sontag, D.; and Rush, A. M. 2016. Character-aware neural language models. In *AAAI'16*, 2741–2749.

Kudo, T. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR* abs/1804.10959.

Labeau, M., and Allauzen, A. 2017. Character and subword-based word representation for neural language modeling prediction. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Association for Computational Linguistics.

Li, J., and Jurafsky, D. 2015. Do multi-sense embeddings improve natural language understanding? *CoRR* abs/1506.01070.

Liang, P. P.; Liu, Z.; Zadeh, A.; and Morency, L.-P. 2018. Multimodal language analysis with recurrent multistage fusion. In *EMNLP*.

Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Multimodal local-global ranking fusion for emotion recognition. In *ICMI*. ACM.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*.

Nguyen, D. Q.; Nguyen, D. Q.; Modi, A.; Thater, S.; and Pinkal, M. 2017. A mixture model for learning multi-sense word embeddings. *CoRR* abs/1706.05111.

Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; and Morency, L.-P. 2016. Deep multimodal fusion for persuasiveness prediction. In *ICMI*, 284–288. ACM.

Oh, A.; Park, S.; Byun, J.; Baek, S.; and Cho, Y. 2018. Subword-level word vector representations for korean. In *ACL*, 2429–2438.

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 *ICMI*. ACM.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543. Association for Computational Linguistics.

Pérez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013. Utterance-level multimodal sentiment analysis. In *ACL*, volume 1, 973–982.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *CoRR* abs/1802.05365.

Pham, H.; Manzini, T.; Liang, P. P.; and Poczos, B. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*. ACL.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 873–883.

Rajagopalan, S. S.; Morency, L.-P.; Baltrusaitis, T.; and Goecke, R. 2016. Extending long short-term memory for multi-view structured learning. In *ECCV*, 338–353.

Reynolds, D. A. 2009. Gaussian mixture models. In Li, S. Z., and Jain, A. K., eds., *Encyclopedia of Biometrics*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *CoRR* abs/1508.07909.

Srivastava, N., and Salakhutdinov, R. 2014. Multimodal learning with deep boltzmann machines. *JMLR* 15.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; and Baik, S. W. 2018. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access* 6:1155–1166.

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

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

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.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI*. AAAI Press.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L. 2018b. Multi-attention recurrent network for human communication comprehension. In *AAAI*. AAAI Press.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *NIPS*.