Words Can Shift:
Dynamically Adjusting Word Representations Using Nonverbal Behaviours

Yansen Wang\textsuperscript{1}, Ying Shen\textsuperscript{2}, Zhun Liu\textsuperscript{2}, Paul Pu Liang\textsuperscript{2}, Amir Zadeh\textsuperscript{2}, Louis-Philippe Morency\textsuperscript{2}
\textsuperscript{1}Department of Computer Science, Tsinghua University
\textsuperscript{2}School of Computer Science, Carnegie Mellon University
ys-wang15@mails.tsinghua.edu.cn; \{yshen2,zhun1,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. Alongside speaker intentions, this form of communication also reveals the sentiment and emotional state of the speaker, as vocal patterns and facial expressions are associated with different sentiment polarities and emotions. Studying these multimodal signals is a new challenge to the field of natural language processing. In this paper, we model the nonverbal aspects of human communication as a shift of its verbal representation. We assume that the intent of words during communication shifts depending on the accompanying nonverbal behaviors. Based on the magnitude and direction of this nonverbal shift vector, the word representation is translated into a different point in the representation space. These shift vectors are learned using a hierarchical recurrent neural network called the Recurrent Attended Variation Embedding Network (RAVEN). By modeling nonverbal behaviors as variations of language representations, RAVEN achieves competitive performance on two publicly available datasets for multimodal sentiment analysis and emotion recognition.

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 intonations. However, the meaning of words and sentences uttered by the speaker often varies dynamically in different nonverbal contexts. Such dynamic behavior can come from different sources such as cultural shift or different political background (Bamler and Mandt, 2017). Yet more importantly, in human multimodal language, it is often intertwined with their nonverbal contexts. Intentions conveyed through uttering a sentence can display drastic shifts in intensity and direction, leading to the phenomena that the uttered words exhibit highly 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 totally different sentiments and remarks by showing opposing facial expressions and intonations. These subtle nonverbal cues contained in the span of the uttered words, including phonemes, facial landmarks, and high-level acoustic features, are crucial towards determining the exact intent of 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 certain visual and acoustic background is called for.

However, modeling the nonverbal contexts concurrent to an uttered word requires fine-grained analysis. This is be-
cause the visual and acoustic representation units often have a much higher temporal frequency than words, leading to the fact that each uttered word will have a sequence of accompanying visual and acoustic “subword” units. The structure of such subword sequences are especially important towards the representation of nonverbal dynamics. As a matter of fact, 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, much of the previous works in understanding and modeling multimodal language has ignored the role of subword analysis. Therefore, most previous works summarize the subword information during each word span using the very simple averaging strategy (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.

Therefore, 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. By achieving excellent performance on multiple metrics on these tasks, we demonstrate the modeling capacity of our model. In addition, our visualizations of the shifted word representations allow us to gain a better understanding of the impact of subword modeling and dynamic shifts on modeling word meaning. Our ablation studies show that both subword analysis and dynamic shifts are important in achieving improved performance. The shifted embeddings learned by RAVEN do indeed exhibit meaningful distributional patterns with respect to the sentiment expressed by the speaker.

**Recurrent Attended Variation Embedding Network (RAVEN)**

The goal of our work is to better model the multimodal human language by considering the subword structure of nonverbal behaviors and learning the multimodal-shifted word representations conditioned on the occurring nonverbal behaviors. To achieve this goal, we propose a Recurrent Attended Variation Embedding Network (RAVEN) which builds the multimodal-shifted word representations through shifting the original word representation in certain magnitude and direction depending on the co-occurring nonverbal behaviors.

An overview of the proposed RAVEN model is given in Figure 2, our model consists of three major components: (1) **Nonverbal Sub-networks** models 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 denotes how far and in which direction has the intent of the word shifts 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, Nonverbal Sub-networks component operates on the visual/acoustic subword units carried alongside each word and yields the visual/acoustic embedding as output as shown in Figure 2.

Formally, we begin with a segment of multimodal language $L$ denoting the sequence of uttered words. For the span of the $i$th word denoted as $L^{(i)}$, we have two accompanying sequences from the visual and acoustic modalities: $V^{(i)} = [v_1^{(i)}; v_2^{(i)}; \ldots; v_{t_v}^{(i)}]$, $A^{(i)} = [a_1^{(i)}; a_2^{(i)}; \ldots; a_{t_a}^{(i)}]$. 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 $L^{(i)}$, $i = 1, \ldots, n$. For the $i$-th word $L^{(i)}$ in the language modality, two different LSTMs are applied separately for its underlying visual and acoustic sequences:

\[
h_v^{(i)} = \text{LSTM}_v(V^{(i)})
\]

\[
h_a^{(i)} = \text{LSTM}_a(A^{(i)})
\]

where $h_v^{(i)}$ and $h_a^{(i)}$ 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 through 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 high activation of facial muscles showing shock while the tone in speech may
not be informative. Therefore we propose a gating mechanism controlling the importance of the visual and acoustic embedding.

In order for the model to control how strong a modality’s influence is, we use the modal-specific influence gates to model the intensity of the influence. To be more concrete, for word \( L^{(i)} \), the visual and acoustic influence gates \( w_v^{(i)} \) and \( w_a^{(i)} \) are computed as follows:

\[
    w_v^{(i)} = \sigma(W_v h_v x_{h_v} + b_v) \tag{3}
\]
\[
    w_a^{(i)} = \sigma(W_a h_a x_{h_a} + b_a) \tag{4}
\]

where \( W_v \) and \( W_a \) are the weight vectors for the visual and acoustic gates and \( b_v \) and \( b_a \) are the scalar bias. The sigmoid function \( \sigma(x) \) is defined as \( \sigma(x) = \frac{1}{1+e^{-x}} \), \( x \in \mathbb{R} \).

We use the concatenation of the visual embedding \( h_v^{(i)} \) and the original word representation \( e^{(i)} \) as the input vector \( x_{h_v} \) for the visual gate. Similarly, we also define the concatenated input vector \( x_{h_a} \) for the acoustic influence gates.

The visual and acoustic influence gates are then used to make adjustments to visual and acoustic embedding to cater to the original word representation. The multimodal shift vector is calculated by fusing the visual and acoustic shift vectors with the corresponding influence gates applied. To be more concrete, for a word \( L^{(i)} \), the multimodal shift vector \( h_m^{(i)} \) is calculated as follows:

\[
    h_m^{(i)} = w_v^{(i)} \cdot (W_v h_v^{(i)}) + w_a^{(i)} \cdot (W_a h_a^{(i)}) + b_h^{(i)} \tag{5}
\]

where \( W_v \) and \( W_a \) are the weight matrix for the visual and acoustic embedding, \( b_h^{(i)} \) is the bias vector.

**Multimodal Shifting**

The Multimodal Shifting component learns to dynamically shift the word representations by integrating the nonverbal shift vector to the original word embedding. To be more concrete, the multimodal-shifted word representation for word \( L^{(i)} \) is computed as follows:

\[
    e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)} \tag{6}
\]
\[
    \alpha = \min \left( \frac{\|e^{(i)}\|_2}{\|h_m^{(i)}\|_2}, 1 \right) \tag{7}
\]

where \( \beta \) is a threshold hyper-parameter that can be determined by cross-validation on a validation set.

In order to ensure the magnitude of the nonverbal shift vector \( h_m^{(i)} \) is not too large as compared to the original word embedding \( e^{(i)} \) and to avoid distorting the original embedding space in unexpected ways, we apply a scaling factor \( \alpha \) to constrain the magnitude of the nonverbal shift vector to
be on a desirable level while maintaining the direction of the shift vector.

By applying the same method for every word in \( \mathbf{L} \), we can transform the original sequence triplet into one sequence of multimodal-shifted representations \( \mathbf{E} = [ \mathbf{e}_m^{(1)}, \mathbf{e}_m^{(2)}, \ldots, \mathbf{e}_m^{(n)} ] \). The new sequence \( \mathbf{E} \) now corresponds to the a shifted version of the original sequence of word representations \( \mathbf{L} \) based on the 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 representation obtained from the first level of the hierarchy into a multimodal representation \( \mathbf{h} \) that can be used for downstream tasks:

\[
\mathbf{h} = \text{LSTM}_h(\mathbf{E}) \tag{8}
\]

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. The codes will be shared with the community.

**Datasets**

To evaluate our approach, we use two multimodal datasets involving tri-modal 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 in a multimodal sentiment analysis task. 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 two people 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).

**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 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; 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\(^1\).

**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\(^2\) 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 derived from algorithms 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) (Nojavanagasgahri 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 (LSTM).

**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

---

\(^1\)https://nlp.stanford.edu/projects/glove/

\(^2\)https://imotions.com/
We present our results on the multimodal datasets in Tables 1 when compared with state-of-the-art models across multiple emotion recognition tasks. Table 1: Sentiment prediction results on the CMU-MOSI test set using multimodal methods. The best three results are noted with *, † and ‡ successively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>MAE</th>
<th>Corr</th>
<th>Acc-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td></td>
<td>1.864</td>
<td>0.057</td>
<td>50.2</td>
</tr>
<tr>
<td>DF</td>
<td></td>
<td>1.143</td>
<td>0.518</td>
<td>72.3</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td></td>
<td>1.079</td>
<td>0.581</td>
<td>73.9</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td></td>
<td>1.019</td>
<td>0.601</td>
<td>73.9</td>
</tr>
<tr>
<td>MARN</td>
<td></td>
<td>0.968</td>
<td>0.625</td>
<td>77.1</td>
</tr>
<tr>
<td>MFN</td>
<td></td>
<td>0.965</td>
<td>0.632</td>
<td>77.4†</td>
</tr>
<tr>
<td>RMFN</td>
<td></td>
<td>0.923†</td>
<td>0.681†</td>
<td>78.4*</td>
</tr>
<tr>
<td>LMF</td>
<td></td>
<td>0.912†</td>
<td>0.668†</td>
<td>76.4</td>
</tr>
<tr>
<td>RAVEN</td>
<td></td>
<td>0.915†</td>
<td>0.691*</td>
<td>78.0</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Happy</th>
<th>Sad</th>
<th>Angry</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metric</td>
<td>Acc-2</td>
<td>F1</td>
<td>Acc-2</td>
<td>F1</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>86.1</td>
<td>81.5</td>
<td>81.1</td>
<td>78.8</td>
</tr>
<tr>
<td>DF</td>
<td></td>
<td>86.0</td>
<td>81.0</td>
<td>81.8</td>
<td>81.2</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td></td>
<td>84.9</td>
<td>81.7</td>
<td>83.2</td>
<td>81.7</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td></td>
<td>85.9</td>
<td>81.3</td>
<td>80.4</td>
<td>74.0</td>
</tr>
<tr>
<td>MARN</td>
<td></td>
<td>86.7</td>
<td>83.6</td>
<td>82.0</td>
<td>81.2</td>
</tr>
<tr>
<td>MFN</td>
<td></td>
<td>86.5</td>
<td>84.0</td>
<td>83.5†</td>
<td>82.1</td>
</tr>
<tr>
<td>RMFN</td>
<td></td>
<td>87.5†</td>
<td>85.8†</td>
<td>82.9</td>
<td>85.1†</td>
</tr>
<tr>
<td>LMF</td>
<td></td>
<td>87.3†</td>
<td>85.8†</td>
<td>86.2†</td>
<td>85.9†</td>
</tr>
<tr>
<td>RAVEN</td>
<td></td>
<td>87.3†</td>
<td>85.8†</td>
<td>83.4†</td>
<td>83.1†</td>
</tr>
</tbody>
</table>

**Results and Discussion**

In this section, we present results for the aforementioned experiments and compare our performance with other state-of-the-art models. We also visualize the multimodal-shifted representations and show that they form interpretable patterns. Finally, in order 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 Sub-networks and Multimodal Shifting from our full model, and find that the presence of both modules is critical to performance.

**Comparison with the State of the Art**

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

**Multimodal Sentiment Analysis**: On the multimodal sentiment regression 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.

**Multimodal Representations in Different Nonverbal Contexts**

As our model builds the multimodal-shifted representation of every word by integrating the concurrent nonverbal context, every instance of the same word will have a different multimodal-shifted embedding, depending on the shifts its concurrent non-verbal behaviors suggest. Such shifts in all instances of the same word often exhibit consistent patterns. Here we visualize the distributions of word tokens that belong to the same word in the CMU-MOSI dataset in 2-dimensional space with PCA (Jolliffe, 2011), and then plot Gaussian contours for points in positive-sentiment context and negative-sentiment context respectively. In addition, we plot the centroid of these points as well as the centroids of the subset in positive/negative contexts. To highlight the relative positions of these centroids, we plot blue and red arrows starting from overall centroid and pointing to negative and positive centroids. The visualizations are shown in Figure 3. Empirically we discover that variations on different words exhibit three different patterns depending on their roles in expressing sentiments in a multimodal context:

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 the overall and the positive centroid.

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 multimodal context. For example, the noun “guy” are frequently used for both addressing good actors and bad actors, and RAVEN is able to shift them accordingly in embedding space toward two different directions, as illustrated in Figure 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 centroid, negative centroid, and overall centroid all lying close to each other. Two example words that fall under this category are “that” and “the”. Their resulting centroids are
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.

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 two components in modeling human multimodal language, we conducted ablation studies to examine the impact on the performance of these two components. We do so by starting with our full model and progressively remove different components, culminating in a simple early-fusion model in the end. The different versions of the model are explained as follows:

RAVEN: our complete proposed model that is capable of modeling subword dynamics and dynamically shifting 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. In this case, vector representations of visual and acoustic modalities 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.

Table 3 shows the results of ablation studies of 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 for visual and acoustic modalities where sequences are of higher frequency, the crude averaging method for sub-sampling them to the same frequency of words does hurt performance. Another interesting 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 drastically improves upon the early-fusion model. This implies that our architecture does successfully capture some underlying structures of human multimodal language.

### Related Works

Previously, much effort has been devoted to building machine learning models that learn from multiple modalities. 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 learning word representations (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 (Chinese Restaurant Process) to learn different word representations for each sense of a word, while 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 works in multimodal sentiment and emotion recognition have 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 can learn robust representations (Wang et al., 2016), they have limited capability of learning modality-specific interactions and tend to overfit (Abbari et al., 2016; 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 separately 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 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 temporal 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 this class of models with use of attention mechanisms (Liang et al., 2018; Bahdanau, Cho, and Bengio, 2014), word-level fusion (Tsai et al., 2018; Gu et al., 2018), and expressive fusion methods (Liu et al., 2018).

These previous studies have explored integrating nonverbal behaviors for sentiment analysis, emotion recognition 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 toward the meaning of the underlying words. Our proposed method obtains these word-level nonverbal features by modeling the nonverbal temporal interaction between the subword units. Unlike previous work, we perform word-level fusion where the word-level nonverbal features introduce variations to word representations. 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.

### 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. We also visualize the shifted word representation in different nonverbal contexts and summarize three 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MAE</th>
<th>Corr</th>
<th>Acc-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAVEN</td>
<td>0.915</td>
<td>0.691</td>
<td>77.9</td>
</tr>
<tr>
<td>RAVEN w/o SHIFT</td>
<td>0.954</td>
<td>0.666</td>
<td>77.7</td>
</tr>
<tr>
<td>RAVEN w/o SUB</td>
<td>0.934</td>
<td>0.652</td>
<td>73.9</td>
</tr>
<tr>
<td>RAVEN w/o SUB&amp;SHIFT</td>
<td>1.423</td>
<td>0.116</td>
<td>50.6</td>
</tr>
</tbody>
</table>

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


Csáji, B. C. 2001. Approximation with artificial neural networks. Faculty of Sciences, Eotvos Lorand University, Hungary


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.


