Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment

Yue Gu, Kangning Yang[*], Shiyu Fu[*], Shuhong Chen, Xinyu Li and Ivan Marsic
Multimedia Image Processing Lab
Electrical and Computer Engineering Department
Rutgers University, Piscataway, NJ, USA
{yue.guapp, ky189, sf568, sc1624, Xinyu.li1118, marsic}@rutgers.edu

Abstract

Multimodal affective computing, learning to recognize and interpret human affect and subjective information from multiple data sources, is still challenging because: (i) it is hard to extract informative features to represent human affects from heterogeneous inputs; (ii) current fusion strategies only fuse different modalities at abstract levels, ignoring time-dependent interactions between modalities. Addressing such issues, we introduce a hierarchical multimodal architecture with attention and word-level fusion to classify utterance-level sentiment and emotion from text and audio data. Our introduced model outperforms state-of-the-art approaches on published datasets, and we demonstrate that our model's synchronized attention over modalities offers visual interpretability.

1 Introduction

With the recent rapid advancements in social media technology, affective computing is now a popular task in human-computer interaction. Sentiment analysis and emotion recognition, both of which require applying subjective human concepts for detection, can be treated as two affective computing subtasks on different levels (Poria et al., 2017a). A variety of data sources, including voice, facial expression, gesture, and linguistic content have been employed in sentiment analysis and emotion recognition. In this paper, we focus on a multimodal structure to leverage the advantages of each data source. Specifically, given an utterance, we consider the linguistic content and acoustic characteristics together to recognize the opinion or emotion. Our work is important and useful because speech is the most basic and commonly used form of human expression.

A basic challenge in sentiment analysis and emotion recognition is filling the gap between extracted features and the actual affective states (Zhang et al., 2017). The lack of high-level feature associations is a limitation of traditional approaches using low-level handcrafted features as representations (Seppi et al., 2008; Rozgic et al., 2012). Recently, deep learning structures such as CNNs and LSTMs have been used to extract high-level features from text and audio (Eyben et al., 2010a; Poria et al., 2015). However, not all parts of the text and vocal signals contribute equally to the predictions. A specific word may change the entire sentimental state of text; a different vocal delivery may indicate inverse emotions despite having the same linguistic content. Recent approaches introduce attention mechanisms to focus the models on informative words (Yang et al., 2016) and attentive audio frames (Mirsamadi et al., 2017) for each individual modality. However, to our knowledge, there is no common multimodal structure with attention for utterance-level sentiment and emotion classification. To address such issue, we design a deep hierarchical multimodal architecture with an attention mechanism to classify utterance-level sentiments and emotions. It extracts high-level informative textual and acoustic features through individual bidirectional gated recurrent units (GRU) and uses a multi-level attention mechanism to select the informative features in both the text and audio module.

Another challenge is the fusion of cues from heterogeneous data. Most previous works focused on combining multimodal information at a holistic level, such as integrating independent predictions of each modality via algebraic rules (Wöllmer et al., 2013) or fusing the extracted modality-specific features from entire utterances

[*] Equally Contribution

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2225–2235
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

(Poria et al., 2016). They extract word-level features in a text branch, but process audio at the frame-level or utterance-level. These methods fail to properly learn the time-dependent interactions across modalities and restrict feature integration at timestamps due to the different time scales and formats of features of diverse modalities (Poria et al., 2017a). However, to determine human meaning, it is critical to consider both the linguistic content of the word and how it is uttered. A loud pitch on different words may convey inverse emotions, such as the emphasis on "hell" for anger but indicating happy on "great". Synchronized attentive information across text and audio would then intuitively help recognize the sentiments and emotions. Therefore, we compute a forced alignment between text and audio for each word and propose three fusion approaches (horizontal, vertical, and fine-tuning attention fusion) to integrate both the feature representations and attention at the word-level.

We evaluated our model on four published sentiment and emotion datasets. Experimental results show that the proposed architecture outperforms state-of-the-art approaches. Our methods also allow for attention visualization, which can be used for interpreting the internal attention distribution for both single- and multi-modal systems. The contributions of this paper are: (i) a hierarchical multimodal structure with attention mechanism to learn informative features and high-level associations from both text and audio; (ii) three word-level fusion strategies to combine features and learn correlations in a common time scale across different modalities; (iii) word-level attention visualization to help human interpretation.

The paper is organized as follows: We list related work in section 2. Section 3 describes the proposed structure in detail. We present the experiments in section 4 and provide the result analysis in section 5. We discuss the limitations in section 6 and conclude with section 7.

2 Related Work

Despite the large body of research on audio-visual affective analysis, there is relatively little work on combining text data. Early work combined human transcribed lexical features and low-level hand-crafted acoustic features using feature-level fusion (Forbes-Riley and Litman, 2004; Litman and Forbes-Riley, 2004). Others used SVMs fed bag of words (BoW) and part of speech (POS) features in addition to low-level acoustic features (Seppi et al., 2008; Rozgic et al., 2012; Savran et al., 2012; Rosas et al., 2013; Jin et al., 2015). All of the above extracted low-level features from each modality separately. More recently, deep learning was used to extract higher-level multimodal features. Bidirectional LSTMs were used to learn long-range dependencies from low-level acoustic descriptors and derivations (LLDs) and visual features (Eyben et al., 2010a; Wöllmer et al., 2013). CNNs can extract both textual (Poria et al., 2015) and visual features (Poria et al., 2016) for multiple kernel learning of feature-fusion. Later, hierarchical LSTMs were used (Poria et al., 2017b). A deep neural network was used for feature-level fusion in (Gu et al., 2018) and (Zadeh et al., 2017) introduced a tensor fusion network to further improve the performance. A very recent work using word-level fusion was provided by (Chen et al., 2017). The key differences between this work and the proposed architecture are: (i) we design a fine-tunable hierarchical attention structure to extract word-level features for each individual modality, rather than simply using the initialized textual embedding and extracted LLDs from COVAREP (Degottex et al., 2014); (ii) we propose diverse representation fusion strategies to combine both the word-level representations and attention weights, instead of using only word-level fusion; (iii) our model allows visualizing the attention distribution at both the individual modality and at fusion to help model interpretability.

Our architecture is inspired by the document classification hierarchical attention structure that works at both the sentence and word level (Yang et al., 2016). For audio, an attention-based BLSTM and CNN were applied to discovering emotion from frames (Huang and Narayanan, 2016; Neumann and Vu, 2017). Frame-level weighted-pooling with local attention was shown to outperform frame-wise, final-frame, and frame-level mean-pooling for speech emotion recognition (Mirsamadi et al., 2017).

3 Method

We introduce a multimodal hierarchical attention structure with word-level alignment for sentiment analysis and emotion recognition (Figure 1). The model consists of three major parts: text attention module, audio attention module, and word-level fusion module. We first make a forced alignment between the text and audio during preprocessing. Then, the text attention module and audio attention module extract the features from the corresponding inputs (shown in Algorithm 1). The word-level fusion module fuses the extracted feature vectors and makes the final prediction via a shared representation (shown in Algorithm 2).

3.1 Forced Alignment and Preprocessing

The forced alignment between the audio and text on the word-level prepares the different data for feature extraction. We align the data at the word-level because words are the basic unit in English for human speech comprehension. We used *aeneas*[1] to determine the time interval for each word in the audio file based on the Sakoe-Chiba Band Dynamic Time Warping (DTW) algorithm (Sakoe and Chiba, 1978).

For the text input, we first embedded the words into 300-dimensional vectors by *word2vec* (Mikolov et al., 2013), which gives us the best result compared to GloVe and LexVec. Unknown words were randomly initialized. Given a sentence S with N words, let w_i represent the ith word. We embed the words through the *word2vec* embedding matrix W_e by:

$$T_i = W_e w_i, i \in [1, N] \tag{1}$$

where T_i is the embedded word vector.

For the audio input, we extracted Mel-frequency spectral coefficients (MFSCs) from raw audio signals as acoustic inputs for two reasons. Firstly, MFSCs maintain the locality of the data by preventing new bases of spectral energies resulting from discrete cosine transform in MFCCs extraction (Abdel-Hamid et al., 2014). Secondly, it has more dimensions in the frequency domain that aid learning in deep models (Gu et al., 2017). We used 64 filter banks to extract the MFSCs for each audio frame to form the MFSCs map. To facilitate training, we only used static coefficients. Each word's MFSCs can be represented as a matrix with $64 \times n$ dimensions, where n is the interval for the given word in frames. We zero-pad all intervals to the same length L, the maximum frame numbers of the word in the dataset. We did extract LLD features using OpenSmile (Eyben et al., 2010b) software and combined them with the MFSCs during our training stage. However, we did not find an obvious performance improvement, especially for the sentiment analysis. Considering the training cost of the proposed hierarchical acoustic architecture, we decided the extra features were not worth the tradeoff. The output is a 3D MFSCs map with dimensions $[N, 64, L]$.

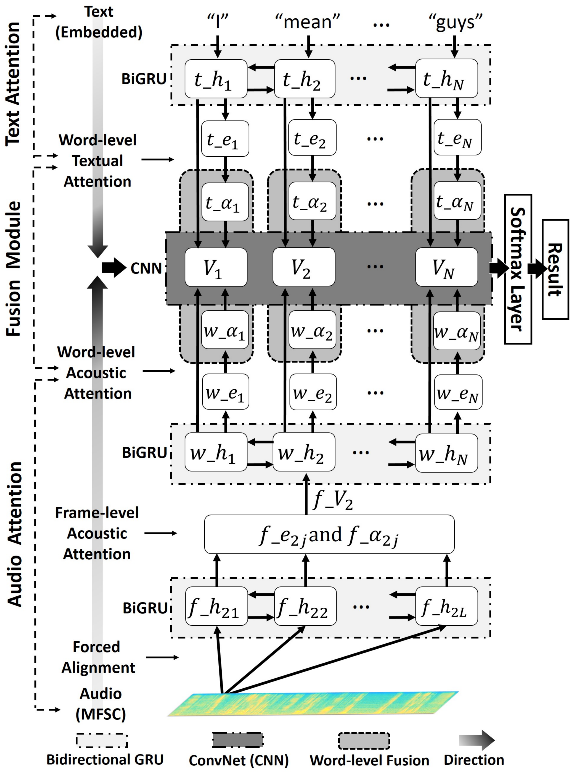

Figure 1: Overall Architecture

3.2 Text Attention Module

To extract features from embedded text input at the word level, we first used bidirectional GRUs, which are able to capture the contextual information between words. It can be represented as:

$$t\_h_i^{\rightarrow}, t\_h_i^{\leftarrow} = bi\_GRU(T_i), i \in [1, N] \tag{2}$$

where bi_GRU is the bidirectional GRU, $t\_h_i^{\rightarrow}$ and $t\_h_i^{\leftarrow}$ denote respectively the forward and backward contextual state of the input text. We combined $t\_h_i^{\rightarrow}$ and $t\_h_i^{\leftarrow}$ as t_h_i to represent the feature vector for the ith word. We choose GRUs instead of LSTMs because our experiments show that LSTMs lead to similar performance (0.07% higher accuracy) with around 25% more trainable parameters.

To create an informative word representation, we adopted a word-level attention strategy that generates a one-dimensional vector denoting the importance for each word in a sequence (Yang et al., 2016). As defined by (Bahdanau et al.,

[1] https://www.readbeyond.it/aeneas/

Algorithm 1 FEATURE EXTRACTION

```
procedure FORCED ALIGNMENT
    Determine time interval of each word
    find $w_i \leftarrow \rightarrow [A_{ij}], j \in [1, L], i \in [1, N]$
end procedure
procedure TEXT BRANCH
    Text Attention Module
    for $i \in [1, N]$ do
        $T_i \leftarrow getEmbedded(w_i)$
        $t\_h_i \leftarrow bi\_GRU(T_i)$
        $t\_e_i \leftarrow getEnergies(t\_h_i)$
        $t\_\alpha_i \leftarrow getDistribution(t\_e_i)$
    end for
    return $t\_h_i, t\_\alpha_i$
end procedure
procedure AUDIO BRANCH
    for $i \in [1, N]$ do
        Frame-Level Attention Module
        for $j \in [1, L]$ do
            $f\_h_{ij} \leftarrow bi\_GRU(A_{ij})$
            $f\_e_{ij} \leftarrow getEnergies(f\_h_{ij})$
            $f\_\alpha_{ij} \leftarrow getDistribution(f\_e_{ij})$
        end for
        $f\_V_i \leftarrow weightedSum(f\_\alpha_{ij}, f\_h_{ij})$
        Word-Level Attention Module
        $w\_h_i \leftarrow bi\_GRU(f\_V_i)$
        $w\_e_i \leftarrow getEnergies(w\_h_i)$
        $w\_\alpha_i \leftarrow getDistribution(w\_e_i)$
    end for
    return $w\_h_i, w\_\alpha_i$
end procedure
```

2014), we compute the textual attentive energies t_e_i and textual attention distribution t_α_i by:

$$t\_e_i = tanh(W_t t\_h_i + b_t), i \in [1, N] \quad (3)$$

$$t\_\alpha_i = \frac{exp(t\_e_i{}^\top v_t)}{\sum_{k=1}^{N} exp(t\_e_k{}^\top v_t)} \quad (4)$$

where W_t and b_t are the trainable parameters and v_t is a randomly-initialized word-level weight vector in the text branch. To learn the word-level interactions across modalities, we directly use the textual attention distribution t_α_i and textual bidirectional contextual state t_h_i as the output to aid word-level fusion, which allows further computations between text and audio branch on both the contextual states and attention distributions.

3.3 Audio Attention Module

We designed a hierarchical attention model with frame-level acoustic attention and word-level attention for acoustic feature extraction.

Frame-level Attention captures the important MFSC frames from the given word to generate the word-level acoustic vector. Similar to the text attention module, we used a bidirectional GRU:

$$f\_h_{ij}^{\rightarrow}, f\_h_{ij}^{\leftarrow} = bi\_GRU(A_{ij}), j \in [1, L] \quad (5)$$

where $f\_h_{ij}^{\rightarrow}$ and $f\_h_{ij}^{\leftarrow}$ denote the forward and backward contextual states of acoustic frames. A_{ij} denotes the MFSCs of the jth frame from the ith word, $i \in [1, N]$. f_h_{ij} represents the hidden state of the jth frame of the ith word, which consists of $f\_h_{ij}^{\rightarrow}$ and $f\_h_{ij}^{\leftarrow}$. We apply the same attention mechanism used for textual attention module to extract the informative frames using equation 3 and 4. As shown in Figure 1, the input of equation 3 is f_h_{ij} and the output is the frame-level acoustic attentive energies f_e_{ij}. We calculate the frame-level attention distribution f_α_{ij} by using f_e_{ij} as the input for equation 4. We form the word-level acoustic vector f_V_i by taking a weighted sum of bidirectional contextual state f_h_{ij} of the frame and the corresponding frame-level attention distribution f_α_{ij} Specifically,

$$f\_V_i = \sum\nolimits_j f\_\alpha_{ij} f\_h_{ij} \quad (6)$$

Word-level Attention aims to capture the word-level acoustic attention distribution w_α_i based on formed word vector f_V_i. We first used equation 2 to generate the word-level acoustic contextual states w_h_i, where the input is f_V_i and $w\_h_i = (w\_h_i^{\rightarrow}, w\_h_i^{\leftarrow})$. Then, we compute the word-level acoustic attentive energies w_e_i via equation 3 as the input for equation 4. The final output is an acoustic attention distribution w_α_i from equation 4 and acoustic bidirectional contextual state w_h_i.

3.4 Word-level Fusion Module

Fusion is critical to leveraging multimodal features for decision-making. Simple feature concatenation without considering the time scales ignores the associations across modalities. We introduce word-level fusion capable of associating the text and audio at each word. We propose three fusion strategies (Figure 2 and Algorithm 2): horizontal fusion, vertical fusion, and fine-tuning attention fusion. These methods allow easy synchronization between modalities, taking advantage of the attentive associations across text and audio, creating a shared high-level representation.

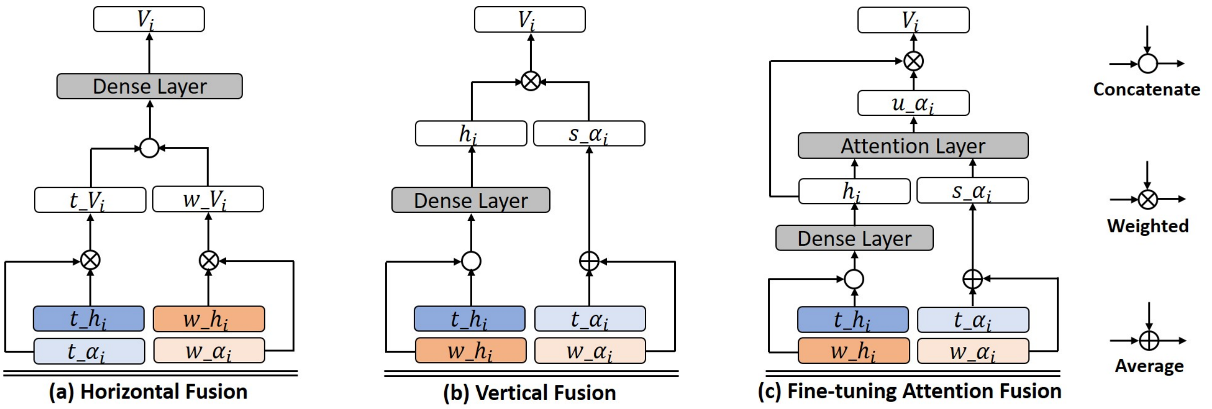

Figure 2: Fusion strategies. t_h_i: word-level textual bidirectional state. t_α_i: word-level textual attention distribution. w_h_i: word-level acoustic bidirectional state. w_α_i: word-level acoustic attention distribution. s_α_i: shared attention distribution. u_α_i: fine-tuning attention distribution. V_i: shared word-level representation.

Algorithm 2 FUSION

```
procedure FUSION BRANCH
    Horizontal Fusion (HF)
    for $i \in [1, N]$ do
        $t\_V_i \leftarrow weighted(t\_\alpha_i, t\_h_i)$
        $w\_V_i \leftarrow weighted(w\_\alpha_i, w\_h_i)$
        $V_i \leftarrow dense([t\_V_i, w\_V_i])$
    end for
    Vertical Fusion (VF)
    for $i \in [1, N]$ do
        $h_i \leftarrow dense([t\_h_i, w\_h_i])$
        $s\_\alpha_i \leftarrow average([t\_\alpha_i, w\_\alpha_i])$
        $V_i \leftarrow weighted(h_i, s\_\alpha_i)$
    end for
    Fine-tuning Attention Fusion (FAF)
    for $i \in [1, N]$ do
        $u\_e_i \leftarrow getEnergies(h_i)$
        $u\_\alpha_i \leftarrow getDistribution(u\_e_i, s\_\alpha_i)$
        $V_i \leftarrow weighted(h_i, u\_\alpha_i)$
    end for
    Decision Making
    $E \leftarrow convNet(V_1, V_2, ..., V_N)$
    return E
end procedure
```

Horizontal Fusion (HF) provides the shared representation that contains both the textual and acoustic information for a given word (Figure 2 (a)). The HF has two steps: (i) combining the bidirectional contextual states (t_h_i and w_h_i in Figure 1) and attention distributions for each branch (t_α_i and w_α_i in Figure 1) independently to form the word-level textual and acoustic representations. As shown in Figure 2, given the input (t_α_i, t_h_i) and (w_α_i, w_h_i), we first weighed each input branch by:

$$t\_V_i = t\_\alpha_i t\_h_i \tag{7}$$

$$w\_V_i = w\_\alpha_i w\_h_i \tag{8}$$

where t_V_i and w_V_i are word-level representations for text and audio branches, respectively; (ii) concatenating them into a single space and further applying a dense layer to create the shared context vector V_i, and $V_i = (t\_V_i, w\_V_i)$. The HF combines the unimodal contextual states and attention weights; there is no attention interaction between the text modality and audio modality. The shared vectors retain the most significant characteristics from respective branches and encourages the decision making to focus on local informative features.

Vertical Fusion (VF) combines textual attentions and acoustic attentions at the word-level, using a shared attention distribution over both modalities instead of focusing on local informative representations (Figure 2 (b)). The VF is computed in three steps: (i) using a dense layer after the concatenation of the word-level textual (t_h_i) and acoustic (w_h_i) bidirectional contextual states to form the shared contextual state h_i; (ii) averaging the textual (t_α_i) and acoustic (w_α_i) attentions for each word as the shared attention distribution s_α_i; (iii) computing the weight of h_i and s_α_i as final shared context vectors V_i, where $V_i = h_i s\_\alpha_i$. Because the shared attention distribution (s_α_i) is based on averages of unimodal attentions, it is a joint attention of both textual and acoustic attentive information.

Fine-tuning Attention Fusion (FAF) preserves the original unimodal attentions and provides a fine-tuning attention for the final prediction (Figure2 (c)). The averaging of attention weights in vertical fusion potentially limits the representational power. Addressing such issue, we propose a trainable attention layer to tune the shared attention in three steps: (i) computing the shared attention distribution s_α_i and shared bidirectional contextual states h_i separately using the same approach as in vertical fusion; (ii) applying attention fine-tuning:

$$u\_e_i = tanh(W_u h_i + b_u) \tag{9}$$

$$u\_\alpha_i = \frac{exp(u\_e_i{}^\top v_u)}{\sum_{k=1}^{N} exp(u\_e_k{}^\top v_u)} + s\_\alpha_i \tag{10}$$

where W_u, b_u, and v_u are additional trainable parameters. The u_α_i can be understood as the sum of the fine-tuning score and the original shared attention distribution s_α_i; (iii) calculating the weight of u_α_i and h_i to form the final shared context vector V_i.

3.5 Decision Making

The output of the fusion layer V_i is the ith shared word-level vectors. To further make use of the combined features for classification, we applied a CNN structure with one convolutional layer and one max-pooling layer to extract the final representation from shared word-level vectors (Poria et al., 2016; Wang et al., 2016). We set up various widths for the convolutional filters (Kim, 2014) and generated a feature map c_k by:

$$f_i = tanh(W_c V_{i:i+k-1} + b_c) \tag{11}$$

$$c_k = max\{f_1, f_2, ..., f_N\} \tag{12}$$

where k is the width of the convolutional filters, f_i represents the features from window i to $i + k - 1$. W_c and b_c are the trainable weights and biases. We get the final representation c by concatenating all the feature maps. A softmax function is used for the final classification.

4 Experiments

4.1 Datasets

We evaluated our model on four published datasets: two multimodal sentiment datasets (MOSI and YouTube) and two multimodal emotion recognition datasets (IEMOCAP and EmotiW).

MOSI dataset is a multimodal sentiment intensity and subjectivity dataset consisting of 93 reviews with 2199 utterance segments (Zadeh et al., 2016). Each segment was labeled by five individual annotators between -3 (strong negative) to +3 (strong positive). We used binary labels based on the sign of the annotations' average.

YouTube dataset is an English multimodal dataset that contains 262 positive, 212 negative, and 133 neutral utterance-level clips provided by (Morency et al., 2011). We only consider the positive and negative labels during our experiments.

IEMOCAP is a multimodal emotion dataset including visual, audio, and text data (Busso et al., 2008). For each sentence, we used the label agreed on by the majority (at least two of the three annotators). In this study, we evaluate both the 4-catgeory (*happy+excited*, *sad*, *anger*, and *neutral*) and 5-catgeory(*happy+excited*, *sad*, *anger*, *neutral*, and *frustration*) emotion classification problems. The final dataset consists of 586 *happy*, 1005 *excited*, 1054 *sad*, 1076 *anger*, 1677 *neutral*, and 1806 *frustration*.

EmotiW[2] is an audio-visual multimodal utterance-level emotion recognition dataset consist of video clips. To keep the consistency with the IEMOCAP dataset, we used four emotion categories as the final dataset including 150 *happy*, 117 *sad*, 133 *anger*, and 144 *neutral*. We used IBM Watson[3] speech to text software to transcribe the audio data into text.

4.2 Baselines

We compared the proposed architecture to published models. Because our model focuses on extracting sentiment and emotions from human speech, we only considered the audio and text branch applied in the previous studies.

4.2.1 Sentiment Analysis Baselines

BL-SVM extracts a bag-of-words as textual features and low-level descriptors as acoustic features. An SVM structure is used to classify the sentiments (Rosas et al., 2013).

LSTM-SVM uses LLDs as acoustic features and bag-of-n-grams (BoNGs) as textual features. The final estimate is based on decision-level fusion of text and audio predictions (Wöllmer et al., 2013).

[2] https://cs.anu.edu.au/few/ChallengeDetails.html

[3] https://www.ibm.com/watson/developercloud/speech-to-text/api/v1/

Sentiment Analysis (MOSI)					Emotion Recognition (IEMOCAP)				
Approach	Category	WA(%)	UA(%)	Weighted-F1	Approach	Category	WA(%)	UA(%)	Weighted-F1
BL-SVM*	2-class	70.4	70.6	0.668	SVM Trees	4-class	67.4	67.4	-
LSTM-SVM*	2-class	72.1	72.1	0.674	GSV-e Vector	4-class	63.2	62.3	-
C-MKL$_1$	2-class	73.6	-	0.752	C-MKL$_2$	4-class	65.5	65.0	-
TFN	2-class	75.2	-	0.760	H-DMS	5-class	60.4	60.2	0.594
LSTM(A)	2-class	73.5	-	0.703	UL-Fusion*	4-class	66.5	66.8	0.663
UL-Fusion*	2-class	72.5	72.5	0.730	DL-Fusion*	4-class	65.8	65.7	0.665
DL-Fusion*	2-class	71.8	71.8	0.720	Ours-HF	4-class	70.0	69.7	0.695
Ours-HF	2-class	74.1	74.4	0.744	Ours-VF	4-class	71.8	71.8	0.713
Ours-VF	2-class	75.3	75.3	0.755	Ours-FAF	4-class	**72.7**	**72.7**	**0.726**
Ours-FAF	2-class	**76.4**	**76.5**	**0.768**	Ours-FAF	5-class	**64.6**	**63.4**	**0.644**

Table 1: Comparison of models. *WA* = weighted accuracy. *UA* = unweighted accuracy. * denotes that we duplicated the method from cited research with the corresponding dataset in our experiment.

C-MKL$_1$ uses a CNN structure to capture the textual features and fuses them via multiple kernel learning for sentiment analysis (Poria et al., 2015).

TFN uses a tensor fusion network to extract interactions between different modality-specific features (Zadeh et al., 2017).

LSTM(A) introduces a word-level LSTM with temporal attention structure to predict sentiments on MOSI dataset (Chen et al., 2017).

4.2.2 Emotion Recognition Baselines

SVM Trees extracts LLDs and handcrafted bag-of-words as features. The model automatically generates an ensemble of SVM trees for emotion classification (Rozgic et al., 2012).

GSV-eVector generates new acoustic representations from selected LLDs using Gaussian Supervectors and extracts a set of weighed handcrafted textual features as an eVector. A linear kernel SVM is used as the final classifier (Jin et al., 2015).

C-MKL$_2$ extracts textual features using a CNN and uses openSMILE to extract 6373 acoustic features. Multiple kernel learning is used as the final classifier (Poria et al., 2016).

H-DMS uses a hybrid deep multimodal structure to extract both the text and audio emotional features. A deep neural network is used for feature-level fusion (Gu et al., 2018).

4.2.3 Fusion Baselines

Utterance-level Fusion (UL-Fusion) focuses on fusing text and audio features from an entire utterance (Gu et al., 2017). We simply concatenate the textual and acoustic representations into a joint feature representation. A softmax function is used for sentiment and emotion classification.

Decision-level Fusion (DL-Fusion) Inspired by (Wöllmer et al., 2013), we extract textual and acoustic sentence representations individually and infer the results via two softmax classifiers, respectively. As suggested by Wöllmer, we calculate a weighted sum of the text (1.2) result and audio (0.8) result as the final prediction.

4.3 Model Training

We implemented the model in Keras with Tensorflow as the backend. We set 100 as the dimension for each GRU, meaning the bidirectional GRU dimension is 200. For the decision making, we selected 2, 3, 4, and 5 as the filter width and apply 300 filters for each width. We used the rectified linear unit (ReLU) activation function and set 0.5 as the dropout rate. We also applied batch normalization functions between each layer to overcome internal covariate shift (Ioffe and Szegedy, 2015). We first trained the text attention module and audio attention module individually. Then, we tuned the fusion network based on the word-level representation outputs from each fine-tuning module. For all training procedures, we set the learning rate to 0.001 and used Adam optimization and categorical cross-entropy loss. For all datasets, we considered the speakers independent and used an 80-20 training-testing split. We further separated 20% from the training dataset for validation. We trained the model with 5-fold cross validation and used 8 as the mini batch size. We set the same amount of samples from each class to balance the training dataset during each iteration.

5 Result Analysis

5.1 Comparison with Baselines

The experimental results of different datasets show that our proposed architecture achieves state-of-the-art performance in both sentiment analysis and emotion recognition (Table 1). We re-implemented some published methods (Rosas et al., 2013; Wöllmer et al., 2013) on MOSI to get baselines.

For sentiment analysis, the proposed architecture with FAF strategy achieves 76.4% weighted accuracy, which outperforms all the five baselines (Table 1). The result demonstrates that the proposed hierarchical attention architecture and word-level fusion strategies indeed help improve the performance. There are several findings worth mentioning: (i) our model outperforms the baselines without using the low-level handcrafted acoustic features, indicating the sufficiency of MFSCs; (ii) the proposed approach achieves performance comparable to the model using text, audio, and visual data together (Zadeh et al., 2017). This demonstrates that the visual features do not contribute as much during the fusion and prediction on MOSI; (iii) we notice that (Poria et al., 2017b) reports better accuracy (79.3%) on MOSI, but their model uses a set of utterances instead of a single utterance as input.

For emotion recognition, our model with FAF achieves 72.7% accuracy, outperforming all the baselines. The result shows the proposed model brings a significant accuracy gain to emotion recognition, demonstrating the pros of the fine-tuning attention structure. It also shows that word-level attention indeed helps extract emotional features. Compared to C-MKL$_2$ and SVM Trees that require feature selection before fusion and prediction, our model does not need an additional architecture to select features. We further evaluated our models on 5 emotion categories, including frustration. Our model shows 4.2% performance improvement over H-DMS and achieves 0.644 weighted-F1. As H-DMS only achieves 0.594 F1 and also uses low-level handcrafted features, our model is more robust and efficient.

From Table 1, all the three proposed fusion strategies outperform UL-Fusion and DL-Fusion on both MOSI and IEMOCAP. Unlike utterance-level fusion that ignores the time-scale-sensitive associations across modalities, word-level fusion combines the modality-specific features for each word by aligning text and audio, allowing associative learning between the two modalities, similar to what humans do in natural conversation. The result indicates that the proposed methods improve the model performance by around 6% accuracy. We also notice that the structure with FAF outperforms the HF and VF on both MOSI and IEMOCAP dataset, which demonstrates the effectiveness and importance of the FAF strategy.

Modality	MOSI		IEMOCAP	
	WA	F1	WA	F1
T	75.0	0.748	61.8	0.620
A	60.2	0.604	62.5	0.614
T+A	**76.4**	**0.768**	**72.7**	**0.726**

Table 2: Accuracy (%) and F1 score on text only (T), audio only (A), and multi-modality using FAF (T+A).

Approach	MOSI ↓ YouTube		IEMOCAP ↓ EmotiW	
	WA	F1	WA	F1
Ours-HF	62.9	0.627	59.3	0.584
Ours-VF	64.7	0.643	60.8	0.591
Ours-FAF	**66.2**	**0.665**	**61.4**	**0.608**

Table 3: Accuracy (%) and F1 score for generalization testing.

5.2 Modality and Generalization Analysis

From Table 2, we see that textual information dominates the sentiment prediction on MOSI and there is an only 1.4% accuracy improvement from fusing text and audio. However, on IEMOCAP, audio-only outperforms text-only, but as expected, there is a significant performance improvement by combining textual and audio. The difference in modality performance might because of the more significant role vocal delivery plays in emotional expression than in sentimental expression.

We further tested the generalizability of the proposed model. For sentiment generalization testing, we trained the model on MOSI and tested on the YouTube dataset (Table 3), which achieves 66.2% accuracy and 0.665 F1 scores. For emotion recognition generalization testing, we tested the model (trained on IEMOCAP) on EmotiW and achieves 61.4% accuracy. The potential reasons that may influence the generalization are: (i) the biased labeling for different datasets (five annotators of MOSI vs one annotator of Youtube); (ii) incomplete utterance in YouTube dataset (such as “about”, “he”, etc.); (iii) without enough speech information (EmotiW is a wild audio-visual dataset that focuses on facial expression).

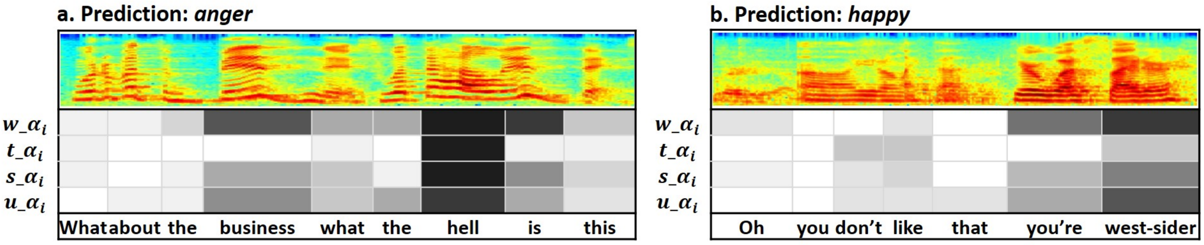

Figure 3: Attention visualization.

5.3 Visualize Attentions

Our model allows us to easily visualize the attention weights of text, audio, and fusion to better understand how the attention mechanism works. We introduce the emotional distribution visualizations for word-level acoustic attention (w_α_i), word-level textual attention (t_α_i), shared attention (s_α_i), and fine-tuning attention based on the FAF structure (u_α_i) for two example sentences (Figure 3). The color gradation represents the importance of the corresponding source data at the word-level.

Based on our visualization, the textual attention distribution (t_α_i) denotes the words that carry the most emotional significance, such as "hell" for *anger* (Figure 3 a). The textual attention shows that "don't", "like", and "west-sider" have similar weights in the *happy* example (Figure 3 b). It is hard to assign this sentence *happy* given only the text attention. However, the acoustic attention focuses on "you're" and "west-sider", removing emphasis from "don't" and "like". The shared attention (s_α_i) and fine-tuning attention (u_α_i) successfully combine both textual and acoustic attentions and assign joint attention to the correct words, which demonstrates that the proposed method can capture emphasis from both modalities at the word-level.

6 Discussion

There are several limitations and potential solutions worth mentioning: (i) the proposed architecture uses both the audio and text data to analyze the sentiments and emotions. However, not all the data sources contain or provide textual information. Many audio-visual emotion clips only have acoustic and visual information. The proposed architecture is more related to spoken language analysis than predicting the sentiments or emotions based on human speech. Automatic speech recognition provides a potential solution for generating the textual information from vocal signals. (ii) The word alignment can be easily applied to human speech. However, it is difficult to align the visual information with text, especially if the text only describes the video or audio. Incorporating visual information into an aligning model like ours would be an interesting research topic. (iii) The limited amount of multimodal sentiment analysis and emotion recognition data is a key issue for current research, especially for deep models that require a large number of samples. Compared large unimodal sentiment analysis and emotion recognition datasets, the MOSI dataset only consists of 2199 sentence-level samples. In our experiments, the EmotiW and MOUD datasets could only be used for generalization analysis due to their small size. Larger and more general datasets are necessary for multimodal sentiment analysis and emotion recognition in the future.

7 Conclusion

In this paper, we proposed a deep multimodal architecture with hierarchical attention for sentiment and emotion classification. Our model aligned the text and audio at the word-level and applied attention distributions on textual word vectors, acoustic frame vectors, and acoustic word vectors. We introduced three fusion strategies with a CNN structure to combine word-level features to classify emotions. Our model outperforms the state-of-the-art methods and provides effective visualization of modality-specific features and fusion feature interpretation.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and feedback. We thank the useful suggestions from Kaixiang Huang. This research was funded by the National Institutes of Health under Award Number R01LM011834.

References

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. 2010a. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1-2):7–19.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010b. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Kate Forbes-Riley and Diane Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Yue Gu, Shuhong Chen, and Ivan Marsic. 2018. Deep multimodal learning for emotion recognition in spoken language. *arXiv preprint arXiv:1802.08332*.

Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech intention classification with multimodal deep learning. In *Canadian Conference on Artificial Intelligence*, pages 260–271. Springer.

Che-Wei Huang and Shrikanth S Narayanan. 2016. Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *INTERSPEECH*, pages 1387–1391.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456.

Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. 2015. Speech emotion recognition with acoustic and lexical features. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4749–4753. IEEE.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diane J Litman and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2227–2231. IEEE.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.

Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma. 2012. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 485–492. ACM.

Dino Seppi, Anton Batliner, Björn Schuller, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Vered Aharonson. 2008. Patterns, prototypes, performance: classifying emotional user states. In *Ninth Annual Conference of the International Speech Communication Association*.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.