Published in final edited form as:

Proc ACM Int Conf Multimed. 2018 October ; 2018: 537–545. doi:10.1145/3240508.3240714.

Human Conversation Analysis Using Attentive Multimodal Networks with Hierarchical Encoder-Decoder

Yue Gu*,

Rutgers University

Xinyu Li*,

Amazon Inc., Rutgers University

Kaixiang Huang,

MTlab, Meitu Inc., Rutgers University

Shiyu Fu,

Rutger University

Kangning Yang,

Rutgers University

Shuhong Chen,

Rutgers University

Moliang Zhou,

Amazon Inc.

Ivan Marsic

Rutgers University

Abstract

Human conversation analysis is challenging because the meaning can be expressed through words, intonation, or even body language and facial expression. We introduce a hierarchical encoder-decoder structure with attention mechanism for conversation analysis. The hierarchical encoder learns word-level features from video, audio, and text data that are then formulated into conversation-level features. The corresponding hierarchical decoder is able to predict different attributes at given time instances. To integrate multiple sensory inputs, we introduce a novel fusion strategy with modality attention. We evaluated our system on published emotion recognition, sentiment analysis, and speaker trait analysis datasets. Our system outperformed previous state-of-the-art approaches in both classification and regressions tasks on three datasets. We also outperformed previous approaches in generalization tests on two commonly used datasets. We achieved comparable performance in predicting co-existing labels using the proposed model

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

yue.guapp@rutgers.edu.

*Equally Contribution

instead of multiple individual models. In addition, the easily-visualized modality and temporal attention demonstrated that the proposed attention mechanism helps feature selection and improves model interpretability.

Keywords

Human Conversation Analysis; Sensor Fusion; Hierarchical Encoder-Decoder Structure; Attention Mechanism

1 INTRODUCTION

Human conversation analysis, including emotion recognition, sentiment analysis and speaker trait detection, is useful in many real world applications such as medical support, activity recognition, chatbots, etc. Aside from challenges in sensor engineering and speech recognition, conversation understanding is still difficult because: (1) *Meaning can be expressed through different media*. A positive attitude can be expressed by words, facial expressions, and intonation, which are often captured by different sensors, requiring a feature fusion mechanism. Furthermore, we shall consider strategies that synchronize the input at the word-level because word is an important basic unit of meaning. (2) *Different sensors may indicate contradicting meanings*. For example, one person can pretend to be happy by saying happy words but with a sad face. Simply merging the features extracted from different modalities may confuse the system. The correct prediction can only be made by selecting a representative input modality and observing the context. (3) *The emotion, sentiment and traits during conversation may or may not change over time*. Most traditional conversation understanding strategies only make a single prediction per conversation, which is inadequate for real-world applications. Furthermore, as conversations may contain multiple attributes (such as both happy and exciting), we need a flexible model that is able to perform classification, regression, and multi-label classification with only a slight modification.

To synchronize different sensor inputs, we introduce a feature extraction strategy that first aligns the raw text, audio, and video at the word level. Unlike sentence-level feature extraction and synchronization [21, 24, 38], our word-level feature extraction breaks down the features to a finer granularity with more details. Unlike direct feature fusion [10, 24], we proposed a fusion strategy with learned modality attention. The modality attention first identifies the importance of each input modality, then extends the importance of each modality to each feature dimension within that modality. Finally, to build a system that is both accurate and flexible, we designed a hierarchical encoder-decoder structure. The hierarchical structure first encodes the multimodal data into word-level features. The conversation-level encoder further selects important information from word-level features with temporal attention and represents all the conversation-level features as a vector. Considering that emotion and sentiment may change over a conversation and that multiple traits may be present simultaneously, our hierarchical decoder structure first decodes features at each time instance. Then, the attribute decoder will further decode the feature vector at each time instance into attributes at that time.

We tested our model on five published datasets. Our model outperformed the most recent state-of-the-art systems: emotion recognition with IEMOCAP [4] on classification; sentiment analysis with MOSI [38] on both classification and regression; and trait analysis with POM [20] on classification. In addition, our hierarchical encoder-decoder system was able to make multi-label predictions (predict multiple traits at once) and achieved performance comparable with most recent research which used 11 individual models. We further tested the generalizability of our system by training it on IEMOCAP and MOSI and testing on EmotiW and MOUD, respectively. Our system outperformed most recent state-of-the-art systems [22] on the same transfer learning task by 6.8% accuracy. We further visualized our modality attention mechanism for modality fusion and temporal attention for encoding, demonstrating that the introduced modality attention model is able to select representative input modalities for sensor fusion. Our contributions include:

1. A word-level feature extraction strategy that is able to synchronize and extract features from different input modalities at the word level
2. A sensor fusion strategy with modality attention that can identify the importance of each modality and the importance of features within each modality.
3. A hierarchical encoder-decoder framework. The encoder encodes features from low level (word level) to high level (conversation level). The decoder first decodes the abstract features into attribute profile at each time instance, and then decodes the attribute profile at each time instance to individual attributes.

The rest of the paper is organized as follows: Section 2 introduces related work. The details of our system are described in section 3. Section 4 presents our experiments and section 5 discusses our system's limitations and possible extensions. A conclusion is in section 6.

2 RELATED WORK

Research on multimodal conversation understanding can be divided into three generations. The first generation used low-level handcrafted features for different modalities, including lexical representations for text, low-level descriptors (LLDs) for audio, and facial characteristic points (FCPs) for video [28, 29, 31]. Instead of simply concatenating the extracted modality-specific features as the final feature representation [28], different shallow fusion strategies were introduced to learn the associations across different modalities, including Bayesian filtering [31] and ensemble SVM trees [29]. Handcrafted features, however, do not generalize well to different application scenarios, and modality-specific fusion strategies cannot effectively model the complex correlations between spatial and temporal information.

The second-generation systems tried address the issues caused by manually crafted features by extracting high-level features using deep learning. Convolutional neural networks (CNNs) were used for visual feature extraction [13] and recurrent neural networks (RNNs) (i.e. gated recurrent units (GRUs) and long short-term memories (LSTMs) [10, 21, 24]) were used for audio and text feature extraction. RNNs were also used for learning long-term dependencies for the fusion of audio and video data [34]. Deep feature extractors can automatically learn features that are general and representative compared with manually-

crafted features. Combining features learned from different input modalities with different time scales, however, remains a challenge [21, 24]. Researchers tried to address modality fusion by applying decision-level fusion such as voting [34], but this approach failed to learn correlations between features extracted from different modalities over time.

Recent systems focus on model-level fusion, generating a shared representation of different modalities [5, 22, 36, 37]. Previous work directly used gated multimodal embedded fusion structures to fuse the raw features, ignoring the temporal associations of individual modalities [5]. To avoid this problem, our system uses individual LSTMs to extract modality-specific features and learns their local temporal associations before fusion. Some researchers used only CNN and openSMILE to extract features [9, 22], but we introduce an attention-based LSTM structure to select informative word-level features. Instead of directly combining sentence-level features and ignoring the temporal associations [36], our hierarchical encoder decoder LSTM learns the temporal associations at the word-level and conversation-level. We also applied temporal attention to select informative shared word-level representations to further improve the system's performance. The most recent research introduced a multi-attention recurrent network to fuse the modality features and learn the temporal associations [37], which achieved state-of-the-art performance on published multimodal datasets. However, this structure cannot learn the correlations across different modalities. Different modalities may have different importance during feature fusion; for example, acoustic features play a more important role in emotion recognition, but less in sentiment analysis. To model this information, we designed a modality fusion strategy that dynamically assigns the importance weights for input modalities.

3 METHODOLOGY

3.1 System Overview

We designed our system to be generalizable and flexible. To achieve generalizability, our system extracts features directly from raw data (instead of using pre-extracted features [7, 8]). Our system consists of two modules (Figure 1): (1) the hierarchical encoder learns features at word-level and conversation-level; (2) the hierarchical decoder makes continuous multi-label predictions at each time instance.

Hierarchical Encoder: (Figure 1, left). Our hierarchical encoder has two levels, the *word-level encoder (WLE)* and *conversation-level encoder (CLE)*. The *word-level encoder* (Figure 1, gray shaded region at left top) synchronizes and combines the features extracted from different sensors, and only selects informative information to form the shared representation. As different sensors have different sampling rates, we perform word-level data synchronization before feature extraction. Compared to multimodal frame-level [39] and sentence-level encoding [22], the *conversation-level encoder* (Figure 1, solid line to the left) combines useful information extracted about each word into a single feature vector. This encoder allows the system to make use of multimodal information extracted over the entire conversation to make the predictions. Because not all words are important, we use a temporal attention mechanism from neural machine translation to select the important word vectors [2].

Hierarchical Decoder: (Figure 1, right). Our hierarchical decoder also has two levels, the *instance-level decoder (ILD)* and *attribute-level decoder (ALD)*. The *instance-level decoder* decodes the features to each time instance. During simple or short conversations, the emotion, sentiment, and traits remain the same, so we can decode all the information into a single instance. In such scenarios, the ILD treats the entire input data as a single time instance, performing a single prediction per case. During complex conversations with changing emotion, sentiment, and traits, the ILD allows us to make continuous predictions within a single conversation. In addition to multi-class classification, some datasets have more than one label per time instance (such as speaker trait analysis datasets). Our *attribute-level decoder* decodes the features at each time instance into multiple co-existing attributes and makes multi-label predictions. This hierarchical decoder structure can be applied to both classification and regression on both the instance-level and attribute-level.

3.2 Word-level Feature Extraction

Word-level feature synchronization and extraction are the foundations of our encoder. Word-level synchronization aligns the features by word, aiding fusion across different modalities. For datasets without word-level timestamps, we synchronized the audio and video to text using aeneas¹ from Sakoe-Chiba Band Dynamic Time Warping (DTW) [30].

Even with word-level timestamps, sensors have different sampling rates that inhibit merging. We considered two options: (1) Down sample all data to a common rate, or (2) Extract features for each modality at their original sampling rate and encode them into a single vector. We adopted the second approach because down sampling loses information. We used different encoding strategies for different modalities (Figure 2):

Text: To capture textual word-level representations, we first embed each word using a pretrained *word2vec* dictionary [16]. We introduced an attention-based bidirectional LSTM (bi-LSTM) to extract word-level representations [35]. The key aspects of our approach are: (1) Instead of directly using the embedded vectors as textual features [5], our LSTM extracts high-level associations. (2) Instead of CNNs with fixed window size, our LSTM fully captures sequential information with varying lengths [21]. (3) The attention mechanism enables selection of informative word-level representations [36]. We set the bi-LSTM dimension equal to 256 and the *word2vec* dimension as 300.

Audio: We introduce a hierarchical attention structure to extract informative acoustic features at both frame-level and word level. Unlike previous research that directly used low-level acoustic descriptors (LLDs), we extracted 100 fps Mel-frequency spectral coefficients (MFSCs), which had been demonstrated effective on deep models due to locality maintenance and higher dimensionality [1, 10]. First, we use an attentive bi-LSTM (same as for the text feature extraction) to select the informative frames. The word-level representation is then a weighted sum of a word's frames. We then apply another attentive bi-LSTM over the word-level representations to learn the associations between representations and select informative representations at word-level. The final outputs are

¹<https://www.readbeyond.it/aeneas/>

word-level acoustic representations with the same dimensionality as the word-level textual representations. As suggested in [11], we used a 64-filter bank to extract MFSCs and initialized the bi-LSTM dimension as 256 at both frame-level and word-level.

Video: This branch captures the facial expression and body posture features from video. Previous work on facial recognition [27] suggested that ResNet [12] performs well at person identification and tracking. We thus built our visual feature extractor with resnet50 for each frame. Attention [2] is applied to the extracted feature vectors for each frame to select the most representative features on each video frame for prediction. Because the datasets we used contains only one face at a time, we chose the Resnet for visual feature extraction. For more complex scenarios with multiple people per-frame, the face detector shall be implemented to capture features from certain face [14, 28].

Each input modality will generate feature vectors of the same length to avoid a dominant modality during fusion. The encoder structure finally outputs a visual, acoustic, and textual representation for each word.

3.3 Modality Attention and Fusion

To merge the word-level features for further processing, we introduce a novel modality fusion strategy. Modality fusion strategies are either early fusion or late fusion. Late fusion makes a final prediction from an ensemble of single-modality predictions (i.e. by voting); it is simple to implement but ignores associations between different modalities. Early fusion concatenates the features, but direct concatenation assumes that each modality contributes equally, which may not always be true (e.g. visual features are important for emotion recognition, but textual features contribute more to sentiment analysis). Attention is commonly used to select the most representative features.

Previous research focused on the attention over modality-specific features [17] or attention over shared feature vectors [5, 23]. We propose a fusion structure with two-level attention: (1) a cross-modality attention focusing on the important modality; (2) a modality-specific attention that highlights important feature dimensions within each modality. We first select a modality by attention, and then expand it to modality-specific attention using a one-dimensional convolution deconvolution network (Figure 3).

Our cross-modality attention is implemented with a set of convolution and pooling operations similar to visual attention used in image recognition [33]. Because convolution does not compromise spatial associations, the input vector for each modality will eventually be converted into a feature point with a high channel dimension ($3 \times \text{channel}$). This output denotes the importance of each input modality. This modality attention representation can be directly broadcasted to the input feature vector and used as the attention vector. However, such method overlooks the associations between modalities. We propose deconvolution and up-sampling (Figure 3) to create the attention vector of each input modality. Similarly, the attention vector of each modality is aligned to the modality's feature vector. Finally, we apply the attention to the feature vector input using soft attention [33].

3.4 Temporal Attention

We add attention for conversation-level feature fusion to select only the important word representations for final decision making. We adopted the temporal attention [2] similar to the attentive bi LSTM mechanism in textual feature extraction [35]. This enables the system to establish temporal association for both encoder and decoder between word-level representations (Figure 1, the region labeled by dash line).

3.5 The Decoder

The hierarchical decoder has two levels: instance-level and attribute level. The instance-level decoder decodes the encoder output across time instances to make continuous predictions (Figure 4(a)). The proposed decoder can be further modified based on two specific requirements. By setting the number of time instances to one (Figure 4(b)) the decoder performs single per-case multi-label predictions (similar to previous research [16]). By changing the number of attributes to one (Figure 4(c)), the system makes multiple binary or multi-class classification per-conversation.

The decoder can perform classification and regression regardless of the number of time instances or attributes. Classification is done by a single softmax activation in the output layer, subject to categorical cross-entropy loss. For regression, we scale all labels to (0, 1) and use one sigmoid neuron as the output, subject to mean-absolute error (MAE) loss.

3.6 Implementation

We synchronized all modalities to text using aeneas. Due to our limited hardware resources, the feature extractors for text, audio, and video are pre-trained and we set all bi-LSTMs with 256 hidden states. We selected the ReLU activation function except for the attention layers. We initialized the learning rate as 0.01 and used the Adam optimizer. We also applied batch normalization and dropout function to address the overfitting issue. We used the 80–20 training and testing split across all modalities. The implementation is based on Keras with Tensorflow (https://github.com/Arthurlxy/ACM_MM_18.git) and is trained on two GTX1080Ti GPUs.

4 EXPERIMENTAL RESULTS

4.1 Dataset and Experiments Configuration

We evaluated our model on five published datasets: two sentiment datasets (MOSI, MOUD), two emotion datasets (IEMOCAP, EmotiW), and one multi-label traits dataset (POM).

MOSI: A multimodal sentiment intensity and subjectivity dataset consisting of 93 English review videos with 2199 utterance segments [38]. We took the average score from five annotators as the ground truth label (as in previous research [37]). Considering the speaker independence, there are 1755 training and 444 testing utterances. We used video, audio, and text for classification (binary-category and 7-category) and regression tasks (Figure 4 (b, c)).

MOUD: A multimodal Spanish sentiment dataset including 79 videos with a *positive/negative* label for each of 498 utterances [28]. Instead of translating the sentences into

English as previous research [22], we randomly initialized the word vectors. Considering its small size, we only used it for generalization experiments. We used 59 videos for training and the remaining 20 for testing. Following previous research [22], we removed the *neutral* label and kept only *positive/negative* labels.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture database is an acted, multimodal, and multi-speaker dataset [4] containing ~12 hours of video, speech, and text. For each sentence in the dataset, we took the voted results from different annotators as labels. We performed experiments on the 10-category configuration and tested the classification (Figure 4 (a, b)) with this dataset.

EmotiW²: A multimodal audio-visual emotion recognition dataset. We used IBM Watson speech-to-text software³ to transcribe the text data. We used the official training and evaluation set (we did not use the test set due to the lack of labels). We only used it for generalization experiments on the classification task (7-category) due to its small size.

POM: The Persuasion Opinion Multimodal dataset contains 904 movie review videos [20]. Each video contains one speaker and is annotated with several traits. Following previous research [37], we used *confidence*, *passion*, *dominance*, *credibility*, *entertaining*, *reserved*, *trusting*, *relaxed*, *nervous*, *humorous*, and *persuasive* (11 multi-labels). To compare, we followed the same 700–204 training-testing split. We performed multiclass classification, regression, and multi-label classification (Figure 4 (a, b, c)) on this dataset.

4.2 Experimental Results and Comparison

We first compared our system with previous research using the same setup on three relatively large datasets [37]. The proposed method outperformed the previous state-of-the-art on all three datasets on different tasks (Table 1). Observe that: 1). Our system outperforms previous work with the same network configuration on different applications (emotion recognition, semantic analysis, and trait analysis). 2). The major differences between our system and previous state-of-the-art are the feature extraction and fusion strategies; our hierarchical encoder and modality attention help the system select representative features and result in higher performance (Table 2). 3). Our proposed structure with attribute-level decoder achieves 37.9% accuracy on POM dataset which is the second-best result compared with the multiclass classification baselines (Table 1). The state-of-the-art system used 11 separate models for multiclass classification [1] while we used single model. Our model is more scalable and easier to be implemented compared with multi-binary model based solution.

4.3 Impact of Encoder Modalities and Attentions

We evaluated the importance of each modality during IEMOCAP and MOSI tests. We analyze how each of the following affects performance: having multimodal data, the proposed word-level encoder (WLE) with modality attention (MA), and the conversation

²<https://cs.anu.edu.au/few/ChallengeDetails.html>

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

level encoder (CLE) with temporal attention (TA). By removing components of our model, we were able to study the impact of different modalities and attentions (Table 2). We found that: 1). The multimodal structure with tri-modality achieves the highest accuracy and F1 score on IEMOCAP and MOSI, indicating that the different modalities indeed complement each other. 2). Text alone outperforms both video and audio on MOSI. However, text and audio have similar performance on IEMOCAP, indicating that vocal delivery is more important for emotion recognition than for sentiment analysis. 3). Removing modality attention (and directly using concatenated features for conversation-level encoding) causes a significant accuracy decrease on both datasets; -4.5% accuracy on IEMOCAP and -3.3% accuracy on MOSI (Table 2). This shows that the word-level encoder with modality attention has a positive influence. 4). Conversation-level encoding with temporal attention brings 4.2% and 4.7% accuracy improvement on IEMOCAP and MOSI (Table 2). This demonstrates that the temporal attention and hierarchical encoder also improves performance.

4.4 Impact of Recurrent Unit

We also made the quantity evaluation of different sequential models on the encoder structure. We did the baseline experiments that using GRUs and LSTMs as the encoder, respectively. The result in Table 3 indicates that the bi-LSTM as we used in the proposed architecture has the best performance on MOSI, IEMOCAP, and POM dataset.

4.5 Decoder Analysis

We evaluated the proposed decoder by comparing the different decision-making model performances on POM, IEMOCAP, and MOSI. Specifically, we tried to answer the following questions: 1). Is the proposed hierarchical decoder sufficiently flexible for different tasks? 2). How does multi-label classification and regression compare with baseline multiple-classifier solutions [37]? We found that: 1). The proposed architecture achieved state-of-the-art classification and regression on MOSI and POM (Table 1 and Table 4). The only difference between these two models is the activation function; softmax for classification and hard sigmoid for regression. This shows that our decoder can handle both classification and regression very well. 2). Compared with MOSI and POM, IEMOCAP has emotional changes over the conversation. Nevertheless, our model achieved state-of-the-art (Table 1 and Table 2) on IEMOCAP simply by changing the number of decoding instances from 1 prediction per case to 100 predictions per case (one at each percentile). This shows our model can handle predictions over time, answering questions 1. 3). We further compared the performance of multiclass and multi-label [15] tasks on both regression and classification (Table 4). Using multiple classifiers (one for each attribute) still achieves 1.7% better performance but requires ~ 7 times more parameters and ~ 6.5 times more training time than our proposed single multi-label classifier. The proposed multi-label architecture still achieved performance comparable to many previous baselines (Table 1). To our best knowledge, we are the first to use multi-label classification and regression for conversation understanding, and our model shows the potential of multi-label decoders for conversation understanding tasks.

4.6 Generalization Test

Our system should be transferable enough to achieve good performance on one dataset despite being trained on another. We performed two sets of experiments to test the system's generalizability.

We first trained a model on the IEMOCAP emotion dataset and tested on both IEMOCAP and EmotiW (Table 5). To match the labels between the IEMOCAP and EmotiW, we changed *surprise* to *exciting* due to the lack of *surprise* in IEMOCAP. Then, we trained our model on the MOSI sentiment dataset and tested on MOUD (Table 5). Similar to previous experiments [37], we processed the MOUD labels to positive/negative to perform transfer learning from the MOSI-trained model.

The results show that training on IEMOCAP achieves 26.1% accuracy with 0.244 F1 score on EmotiW. We further compared the generalization ability of our method with previous approaches [24]. We re-implemented their method with the same training and testing split. The proposed system outperforms the previous approach by 2.7% accuracy. The MOSI-trained model achieves 59.5% accuracy with 0.592 F1 score on MOUD. Compared to the 77.5% accuracy with 0.774 F1 score on the MOSI test set, the lower generalization accuracy might be caused by the language difference (MOSI is English, but MOUD is Spanish). However, compared with the approach in [22], our system still outperforms the previous method by 6.8% accuracy.

5 DISCUSSION

5.1 Visualization of Attentions

Unlike previous approaches using uninterpretable attention at the hidden-state-level [37] or feature-level [23], our system provides direct visualizations of modality and temporal attention. The word level design allows intuitive understanding of modality importance of each word and its importance in the text sequence.

We noticed that the model assigns the text modality a high attention score on the words that carry emotion information, such as “dumb” (Figure 5). The system also focuses on audio or video when tone or facial expression changes. For example, the system focuses on video when the kid has knitted brows and then on audio when he has a higher acoustic energy distribution (Figure 5 (a)). The visualization demonstrates that the modality attention selects useful information from different sources.

Unlike modality attention, temporal attention assigns the importance of each word representation. We see that the temporal attention vector peaks on “great”, indicating this representation contains the most informative information (Figure 5 (b)). Because temporal attention is computed over a multimodal representation, it selects the most important words considering all modalities.

5.2 Possible Extensions

Our hierarchical decoder can be used for text and conversation understanding. It is common that a conversation or text has multiple attributes: a person can be both *angry* and *sad*,

excited and *happy*. Unlike previous strategies [37], our hierarchical encoder-decoder framework flexibly decodes one or multiple attributes simultaneously over time with a single model. In addition, our temporal and modality attentions allow visual interpretation of the model's inner workings (Figure 5).

We demonstrated that our hierarchical decoder is able to make both multiclass and multi-label predictions with performance comparable to most of previous systems (Table 1). Yet there are still some limitations: 1). the attribute-decoder framework assumes associations are between conversational attributes: e.g. if a conversation is convincing, it is usually pervasive. However, such complicated associations cannot be fully captured through limited data size (only 904 samples) with an imbalanced amount of data across attributes (e.g. there is a lack of labels with <3 *confidence* score). We will continue to test our system on different datasets with concurrent labels. 2). The model simply decodes the entire conversation into time instances with equal length, which cannot make sentence level predictions as in previous research. This may lead to some performance mismatches between our system and the previous approaches. We plan to decode features for sentence-level instance predictions in the future.

6 CONCLUSIONS

We introduced a novel human conversation analysis system using a hierarchical encoder-decoder framework. To better combine features extracted from different modalities, we proposed word-level fusion with modality attention. Our system achieved state-of-the-art performance on three published datasets and outperformed others at generalization testing. We hope to deliver the following contributions to the community:

1. A hierarchical encoder-decoder framework that can recognize emotion, sentiment, and speaker traits.
2. A word-level feature extraction strategy that can be widely used for emotion recognition, sentiment analysis, and associated applications.
3. An attentive modality fusion strategy that can be used for any multimodal application.
4. A detailed comparison with previous work for future reference, including generalization and classification vs. regression tests.
5. Datasets synchronized to the word-level and our source code for future comparative research.

ACKNOWLEDGMENTS

We would like to thank the four reviewers' valuable feedback. This research was supported in part by the National Institutes of Health under Award Number R01LM011834.

REFERENCES

- [1]. Ossama Abdel-Hamid Abdel-rahman Mohamed, Jiang Hui, Deng Li, Penn Gerald, and Yu Dong. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing* 22, 10 (2014), 1533–1545.
- [2]. Bahdanau Dzmitry, Cho Kyunghyun, and Bengio Yoshua. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [3]. Breiman Leo. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [4]. Busso Carlos, Bulut Murtaza, Narayanan Shrikanth, Gratch J, and Marsella S. 2013 Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction*, Gratch J and Marsella S, Eds (2013), 110–127.
- [5]. Chen Minghai, Wang Sen, Paul Pu Liang Tadas Baltrušaitis, Zadeh Amir, and Morency Louis-Philippe. 2017 Multimodal sentiment analysis with word-level fusion and reinforcement learning In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 163–171.
- [6]. Cortes Corinna and Vapnik Vladimir. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
- [7]. Degottex Gilles, Kane John, Drugman Thomas, Raitio Tuomo, and Scherer Stefan. 2014 COVAREP-A collaborative voice analysis repository for speech technologies In *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on. IEEE, 960–964.
- [8]. Eyben Florian, Martin Wöllmer Alex Graves, Schuller Björn, Cowie Ellen Douglas, and Cowie Roddy. 2010. 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 (2010), 7–19.
- [9]. Eyben Florian, Wöllmer Martin, and Schuller Björn 2010 Opensmile: the munich versatile and fast open-source audio feature extractor In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.
- [10]. Gu Yue, Chen Shuhong, and Marsic Ivan. 2018. Deep Multimodal Learning for Emotion Recognition in Spoken Language. *arXiv preprint arXiv:1802.08332* (2018).
- [11]. Gu Yue, Yang Kangning, Fu Shiyu, Chen Shuhong, Li Xinyu, and Marsic Ivan. 2018. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. *arXiv preprint arXiv:1805.08660* (2018).
- [12]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2016 Identity mappings in deep residual networks In *European conference on computer vision*. Springer, 630–645.
- [13]. Krizhevsky Alex, Sutskever Ilya, and Hinton Geoffrey E. 2012 Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
- [14]. Li Xinyu, Zhang Yanyi, Zhang Jianyu, Chen Yueyang, Li Huangcan, Marsic Ivan, and Burd Randall S. 2017 Region-based Activity Recognition Using Conditional GAN In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1059–1067.
- [15]. Liu Shuhua Monica and Chen Jiun-Hung. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications* 42, 3 (2015), 1083–1093.
- [16]. Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg S, and Dean Jeff. 2013 Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [17]. Mirsamadi Seyedmahdad, Barsoum Emad, and Zhang Cha. 2017 Automatic speech emotion recognition using recurrent neural networks with local attention In *Acoustics, Speech and Signal Processing (ICASSP)*, 2017 IEEE International Conference on. IEEE, 2227–2231.
- [18]. Morency Louis-Philippe, Mihalcea Rada, and Doshi Payal. 2011 Towards multimodal sentiment analysis: Harvesting opinions from the web In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 169–176.

- [19]. Nojavanasghari Behnaz, Gopinath Deepak, Koushik Jayanth, Baltrušaitis Tadas, and Morency Louis-Philippe. 2016 Deep multimodal fusion for persuasiveness prediction In Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 284–288.
- [20]. Park Sunghyun, Han Suk Shim Moitrey Chatterjee, Sagae Kenji, and Morency Louis Philippe. 2014 Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach In Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 50–57.
- [21]. Poria Soujanya, Cambria Erik, and Gelbukh Alexander. 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. 2539–2544.
- [22]. Poria Soujanya, Cambria Erik, Hazarika Devamanyu, Majumder Navonil, Zadeh Amir, and Morency Louis-Philippe. 2017 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), Vol. 1 873–883.
- [23]. Poria Soujanya, Cambria Erik, Hazarika Devamanyu, Mazumder Navonil, Zadeh Amir, and Morency Louis-Philippe. 2017 Multi-level multiple attentions for contextual multimodal sentiment analysis In Data Mining (ICDM), 2017 IEEE International Conference on. IEEE, 1033–1038.
- [24]. Poria Soujanya, Chaturvedi Iti, Cambria Erik, and Hussain Amir. 2016 Convolutional MKL based multimodal emotion recognition and sentiment analysis In Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 439–448.
- [25]. Quattoni Ariadna, Wang Sybor, Morency Louis-Philippe, Collins Morency, and Darrell Trevor. 2007. Hidden conditional random fields. IEEE transactions on pattern analysis and machine intelligence 29, 10 (2007).
- [26]. Shyam Sundar Rajagopalan Louis-Philippe Morency, Baltrušaitis Tadas, and Goecke Roland. 2016 Extending long short-term memory for multi-view structured learning In European Conference on Computer Vision. Springer, 338–353.
- [27]. Ranjan Rajeev, Patel Vishal M, and Chellappa Rama. 2017 Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
- [28]. Verónica Pérez Rosas Rada Mihalcea, and Morency Louis-Philippe. 2013. Multimodal sentiment analysis of spanish online videos. IEEE Intelligent Systems 28, 3 (2013), 38–45.
- [29]. Rozgic Viktor, Ananthakrishnan Sankaranarayanan, Saleem Shirin, Kumar Rohit, and Prasad Rohit. 2012 Ensemble of svm trees for multimodal emotion recognition In Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 1–4.
- [30]. Sakoe Hiroaki and Chiba Seibi. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing 26, 1 (1978), 43–49.
- [31]. Savran Arman, Cao Houwei, Shah Miraj, Nenkova Ani, and Verma Ragini. 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. ACM, 485–492.
- [32]. Song Yale, Morency Louis-Philippe, and Davis Randall. 2012 Multi-view latent variable discriminative models for action recognition In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2120–2127.
- [33]. Wang Fei, Jiang Mengqing, Qian Chen, Yang Shuo, Li Cheng, Zhang Honggang, Wang Xiaogang, and Tang Xiaoou. 2017. Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017).
- [34]. Martin Wöllmer Felix Weninger, Knap Tobias, Schuller Björn, Sun Congkai, Sagae Kenji, and Morency Louis-Philippe. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28, 3 (2013), 46–53.

- [35]. Yang Zichao, Yang Diyi, Dyer Chris, He Xiaodong, Smola Alex, and Hovy Eduard. 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 1480–1489.
- [36]. Zadeh Amir, Chen Minghai, Poria Soujanya, Cambria Erik, and Morency Louis-Philippe. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017).
- [37]. Zadeh Amir, Paul Pu Liang Soujanya Poria, Vij Prateek, Cambria Erik, and Morency Louis-Philippe. 2018. Multi-attention recurrent network for human communication comprehension. arXiv preprint arXiv:1802.00923 (2018).
- [38]. Zadeh Amir, Zellers Rowan, Pincus Eli, and Morency Louis-Philippe. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016).
- [39]. Zhang Shiqing, Zhang Shiliang, Huang Tiejun, Gao Wen, and Tian Qi. 2017. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology (2017).

CCS CONCEPTS

- Computing methodologies → Natural language processing; Computer vision;

Figure 1:
The structure of our proposed hierarchical encoder-decoder for conversation understanding.

Figure 2:
Word-level data synchronization and feature extraction using attention mechanisms.

Figure 3:

Our fusion strategy based on 1D fully convolutional network and soft attention.

Figure 4:

(a) The hierarchical decoder. (b) Modification 1: The hierarchical decoder with single time instance. (c) Modification 2: The hierarchical decoder with single attribute.

Figure 5:
Visualization of modality attention (MA) and temporal attention (TA) on MOSI. (a) Negative example. (b) Positive example.

Table 1:

Experimental results and comparison on MOSI, IEMOCAP, and POM. (BC) for binary classification, (MCC) for multiclass classification, (MCR) for multiclass regression, and (MLC) for multi-label classification (accuracy in percentage).

	MOSI(BC)		MOSI(MCC)		MOSI(MCR)		IEMOCAP(MCC)		POM(MCC)		POM(MLC)	
	Acc.	F1-Score	Acc.		MAE		Acc.	F1-Score	Acc.		Acc.	
Majority	50.2	0.501	17.5		1.864		21.2	0.074	24.0		✓	
RF[3]	56.4	0.563	21.3		✓		24.1	0.180	32.6		✓	
SVM[6]	71.6	0.723	26.5		1.100		27.3	0.253	34.4		✓	
THMM[18]	50.7	0.454	17.8		✓		23.5	0.108	23.8		✓	
C-MKL[21]	72.3	0.720	30.2		✓		34.0	0.311	✓		✓	
EF-HCRF[25]	65.3	0.654	24.6		✓		32.0	0.205	✓		✓	
MV-HCRF[32]	65.6	0.657	24.6		✓		32.0	0.205	✓		✓	
DF[19]	72.3	0.721	26.8		1.143		26.1	0.200	34.3		✓	
EF-LSTM[37]	73.3	0.732	32.4		1.023		34.1	0.323	36.4		✓	
MV-LSTM[26]	73.9	0.740	33.2		1.019		31.3	0.267	36.1		✓	
BC-LSTM[22]	73.9	0.739	28.7		1.079		35.0	0.341	34.1		✓	
TFN[36]	74.6	0.745	28.7		1.040		36.0	0.345	31.9		✓	
MARN[37]	77.1	0.770	34.7		0.968		37.0	0.359	39.4		✓	
Ours	77.5	0.774	38.5		0.932		39.4	0.383	39.6		37.9	

Table 2:

Experimental results and comparison of modality importance (accuracy in percentage)

Method	IEMOCAP(MCC)		MOSI(MCC)	
	Acc.	F1	Acc.	F1
T	31.8	0.307	32.7	0.271
A	31.5	0.310	27.5	0.225
V	28.4	0.268	23.3	0.191
T+A	33.4	0.325	32.8	0.278
T+V	32.7	0.321	32.1	0.270
A+V	31.8	0.301	23.8	0.187
T+A+V (no MA & WLE)	34.9	0.341	35.2	0.301
T+A+V (no TA & CLE)	35.2	0.352	33.8	0.298
T+A+V (with all)	39.4	0.383	38.5	0.331

Table 3:

Encoder quantity analysis (accuracy in percentage)

	MOSI(MCC)	IEMOCAP(MCC)	POM(MCC)
GRU	37.4	38.6	38.9
LSTM	37.9	39.1	39.3
bi-LSTM	38.5	39.4	39.6

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4:

Comparison of multiclass classification (MCC), multiclass regression (MCR), and multi-label classification (MLC), multi-label regression (MLR) on POM dataset (accuracy in percentage).

	Acc. or MAE	Trainable Parameters	Training Time (s)
MCC	39.6	3.45×10^8	3.96×10^5
MLC	37.9	5.14×10^7	6.15×10^4
MCR	0.102	4.45×10^8	1.22×10^6
MLR	0.158	3.18×10^7	8.74×10^4

Table 5:

Experimental results on generalization (accuracy in percentage)

IEMOCAP (MCC)→EmotiW (MCC)		
Testing Set	Acc.	F1
IEMOCAP	39.4	0.383
EmotiW with[22]	23.4	0.231
EmotiW with our method	26.1	0.244
MOSI (BC)→MOUD (BC)		
Testing Set	Acc.	F1
MOSI	77.5	0.774
MOUD with[2]	52.7	\
MOUD with our method	59.5	0.592