SENTIMENTAL ANALYSIS

PRANAY KUMAR and HARSHITA JAIN and ISHAN VARSHNEY
Bennett University, INDIA
E20CSE461@bennett.edu.in, E20CSE470@bennett.edu.in, E20CSE460@bennett.edu.in

Abstract

Audio Sentiment Analysis is popular research area contrary to the conventional text-based sentiment analysis to depend on the effectiveness of acoustic features extracted from speech. However, current progress in audio sentiment analysis focuses primarily on extracting homogeneous acoustic features or fails to effectively fuse heterogeneous features. In this paper, we propose an utterance-based deep neural network model with a parallel combination of CNN and LSTM-based networks to obtain Audio Sentiment features. ASV is a vector that can best reflect sentiment information in audio. Our model is specifically trained using utterance-level labels, and ASV can be extracted and creatively fused from two branches. Spectrum graphs generated by signals are fed as inputs to the CNN model branch, whereas spectral centroid, MFCC, and other recognised traditional acoustic features extracted from dependent utterances in an audio are fed as inputs to the LSTM model branch. BiLSTM with attention mechanism is also used for feature fusion. Our model can recognize audio sentiment precisely and quickly and hence, is better than traditional acoustic features or vectors extracted from other deep learning models. Furthermore, experimental results indicate that the proposed model outperforms the state-of-the-art approach by 9.33% on MOSI dataset.

1 Introduction

Sentiment Analysis is a well-studied research area in Natural Language Processing (NLP) (Pang B et at al. 2008), which is the computational study of peoples' opinions, sentiments, appraisals, and attitudes towards entities such as products, services, organizations and so on (Liu B et al. 2015). Traditional sentiment analysis methods are mostly based on text. Interestingly, a recent study shows that voice-only as modality seems best for humans empathic accuracy as compared to video-only or audiovisual communication (Kraus et al. 2017). In fact, audio sentiment analysis is a difficult task due to the complexity of audio signal. It aims to correctly analyse the sentiment of the speaker from speech signals, has drawn a great deal of attention of researchers.

In recent years, three main methods for audio sentiment analysis have emerged. First, automatic speech recognition (ASR) technology is used to convert speech into text, followed by traditional text-based sentiment detection systems. (S. Ezzat et at al. 2012). Secondly, adopts a generative model operating directly on the raw audio waveform (Van Den Oord A et at al. 2016). Third, it focuses on extracting signal features from raw audio files (Bertin et al. 2011), which accurately captures the tonal content of music and has been shown to be more effective than original audio spectrum descriptors such as Mel-frequency Cepstrum coefficients (MFCC).

However, for converting speech to text, recognise each word said by a person in an audio, convert them into word embedding, and use NLP techniques such as TF-IDF and bag of words model. The outcome is not always accurate, because the ability to reliably detect a very focused vocabulary in spoken comments is required (Kaushik L et al. 2015). Furthermore, when the voice is transferred to the text, some sentimentrelated signal characteristics are also lost, resulting in a decrease in the accuracy of the sentiment classification. Deep learning is popularly used in audio sentiment analysis in recent years.

We believe that information extracted from a single utterance must be context-dependent. A flash of loud expression, for example, may not indicate a person has a strong emotion because it could be caused by a cough, whereas a continuous loud one is far more likely to indicate the speaker has a strong emotion.

Based on a large number of experiments, we extract the features of each utterance in an audio using the Librosa toolkit and obtain the four most effective features in this paper. Using a BiLSTM with an attention mechanism, combine them to represent sentiment information. Furthermore, we develop a novel model for audio sentiment analysis called Audio Feature Fusion-Attention based CNN and RNN (AFF-ACRNN). We can obtain a new fusion of audio feature vectors before the softmax layer by feeding spectrum graphs and selected traditional acoustic features as input in two separate branches. is the class of sentiment.

Major contributions of the paper are that:

- We propose an effective AFF-ACRNN model for audio sentiment analysis, through combining multiple traditional acoustic features and spectrum graphs to

learn more comprehensive sentiment information in audio.

- Our model is language insensitive and pay more attention to acoustic features of the original audio rather than words recognized from the audio.
- Experimental results indicate that the proposed method outperforms the state-of-the-art methods (Poria et al. 2017) on Multimodal Corpus of Sentiment Intensity dataset(MOSI) and Multimodal Opinion Utterances Dataset(MOUD).

2 Related Work

Current methods for audio sentiment analysis are mostly based on deep neural network. We briefly discuss the improvements made to the task of audio sentiment analysis using deep learning.

Convolutional Neural Network (CNN)

CNN (Y. Le Cun et at al. 1990) are well-known for extracting features from a image by using convolutional kernels and pooling layers to emulates the response of an individual to visual stimuli. Moreover, CNN have been successfully used not only for computer vision, but also for speech (T. N. Sainath et at al. 2015). For speech recognition, CNN is proved to be robust against noise compared to other DL models (D.Palaz et at al. 2015).

Audio Feature Representation and Extraction

Researchers have found pitch and energy related features playing a key role in affect recognition (Poria S et at al. 2017). Some researchers have also used formants, MFCC, root-mean-square energy, spectral centroid, and tonal centroid features for feature extraction. There are several utterances during speech production, and the audio signal can be divided into several segments for each utterance. Global features are calculated by calculating several statistics, such as the average, mean, and deviation of local features.. There are some drawbacks of calculating global features, as some of them are only useful to detect effect of high arousal, e.g., anger and disgust. For lower arousal, global features are not that effective, e.g., In addition, global features lack temporal information and dependencies between two segments in an utterance. In a recent study (Cummins N et al. 2017), deep spectrum features were derived from feeding spectrum graphs through a very deep image classification CNN and forming a feature vector from the results. The activation of the last fully connected layer. Librosa (McFee B et at al. 2015) is an open-source python package for music and audio analysis which can extract all the key features as elaborated above.

3 Methodology

In this section, we go over the proposed AFF-ACRNN model for audio sentiment analysis in detail. We begin by providing an overview of the entire neural network architecture. After that, two separate branch of AFF-ACRNN will be explained in details. We discuss our model's fusion mechanism in the final section.

Model—AFF-ACRNN

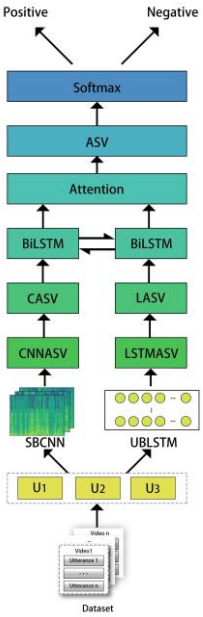

Figure 1: Overview of AFF-ACRNN Model

We concentrate on a model that has two parallel branches, the utterance based BiLSTM branch (UB-BiLSTM) and the spectrum-based CNN branch (SBCNN), whose core mechanisms are based on LSTM and CNN. One branch of proposed model uses the BiLSTM to extract temporal information between adjacent utterances, another branch uses the renowned CNN based network to extract features from spectrum graph that sequence model cannot achieve. Furthermore, audio feature vector of each piece of utterance is the input of the proposed neural network that based on Audio Feature Fusion (AFF), we can obtain a new fusion audio feature vector before the softmax layer, which we call the Audio Sentiment Vector (ASV). Finally, the output of the softmax layer produces our final sentiment classification results, as shown in Figure 1.

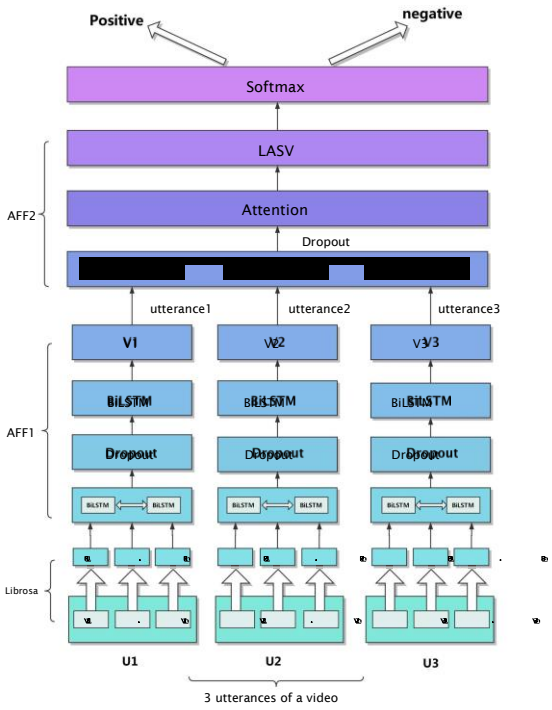

Figure 2: Overview of Our UB-BiLSTM Model

Audio Sentiment Vector (ASV) from Audio Feature Fusion (AFF)

1: procedure CNN BRANCH
2: for i:[0,n] do
3: $x_i \leftarrow$ *get SpectrogramImage*(u_i)
4: $c_i \leftarrow$ *CNNModel*(x_i)
5: $l_i \leftarrow$ *BiLSTM*(c_i)
6: end for
7: end procedure
8: procedure FIND CORRESPONDING LABEL
9: for i:[0:2199] do
10: *rename*(u_i) // for better order in sorting
11: $NameAndLabel = createIndex(u_i)$
12: // A dictionary [utterance Name: Label]
13: end for
14: $Label_x = NameAndLabel(u_x)$
15: end procedure

CNN Layers Similar to the UB-BiLSTM model proposed above, we extract the spectrum graph of each utterance through the Librosa toolkit and use it as the input of our CNN branch. The convolutional layer performs 2-dimensional convolution between the spectrum graph and the predefined linear filters. A number of filters with different functions are used to enable the network to extract complementary features and learn the characteristics of the input spectrum graph. Deep convolutional neural networks are used to obtain a more refined audio feature vector, which is then fed into the BiLSTM layer to learn related sentiment information between adjacent utterances. Finally, before the softmax layer, we get another effective vector CASV extracted by our CNN framework, as shown in Figure 3. The procedure is explained in Algorithm 1's CNN branch procedure.

We extract the Audio Sentiment Vector (ASV) that has the greatest impact on the sentiment classification in the three utterances through the action of the attention mechanism while successfully learning the pertinent sentiment information of adjacent utterance. Finally, Softmax determines the final sentiment classification outcome.

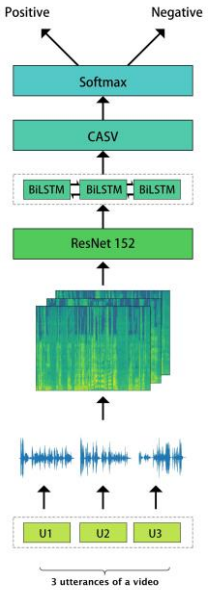

Figure 3: Overview of Our ResNet152 CNN Model

layer. In (J. Donahue et at al. 2015), long-term recurrent convolution network (LRCN) model was proposed for visual recognition. LRCN is a consecutive structure of CNN and LSTM. LRCN processes the variable-length input with a CNN, whose outputs are fed into LSTM network, which finally predicts the class of the input. In (T. N. Sainath et at al. 2015), a cascade structure was used for voice search. Compared to the method mentioned above, the proposed network forms a parallel structure in which LSTM and CNN accept different inputs separately. Therefore, the Audio Sentiment Vector (ASV) can be extracted more comprehensively, and a better classification result can be got.

4 Experiments

In this section, we exhibit our experimental results and the analysis of our proposed model. More specifically, our model is trained and evaluated on utterance-level audio of categories positive neutral negative.

Experiment Setting

Evaluation Metrics We evaluate our performance by weighted accuracy on both 2-class, 5-class and 7-class classification.

$$weighted\ accuracy = \frac{correct\ utterances}{utterances}$$

Additionally, F-Score is used to evaluate 2-class classification.

$$\mathrm{F}^{\beta} = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Where β represents the weight between precision and recall. During our evaluation process, we set β = 1 since we regard precision and recall has the same weight thus F_1-score is adopted.

However, in 5-class and 7-class classification, we use Macro F_1-Score to evaluate the result.

$$Macro\ F_1 = \frac{\sum_{1}^{n} F_{1n}}{n}$$

where n represents the number of classification and F_{1n} is the F_1 score on nth category.

We took the average of three annotations as the sentiment polarity and considered three conditions where consists of two classes (positive and negative), three classes (positive, neutral, negative).

Our dataset's train and test splits are totally unrelated in terms of speakers. In order to better compare with the previous work, similar to (Poria et at al. 2017), we divide the data set by 7:3 approximately.

For proposed CNN framework, the input images are warped into a fixed size of 512 $*$ 512. If the training samples' bounding boxes are provided, we crop the images first before warping them to the fixed size. We use the fine-tuning training strategy to train the feature encoder.

In every experiment, Adam or SGD optimizer trains our networks. In the CNN branch, we initiate the learning rate to be 0.001, and there are 20 epochs in training Resnet-152 with batch size equals to 20 in each epoch.

epoch.

Performance Comparison

Comparison of different feature combinations. Firstly, we have considered three types of acoustic features that can best represent an audio, which mainly includes MFCC, rootmean-square energy, spectral and tonal features. A lot of experiments have be done in order to get the best feature combinations with different model on three types of classification. on the acoustic information previous. It can be seen that the best number of feature combination is four and those four features are MFCC, spectral centroid, spectral contrast and chroma stft. That means the other three features, which are root-mean-square energy, spectral contrast and tonal centroid may introduce some noise or misleading in our sentiment analysis since all seven types of features do not have the best result.

Discussion

The model's effectiveness is assessed using metrics like weighted accuracy and F1-Score and Macro F1-Score. Numerous studies in the UB-Bilstm branch have demonstrated that four types of heterogeneous traditional features trained by BiLSTM will produce the best results, with a weighted accuracy of 68.72% on MOSI. We conducted seven experiments in the SBCNN branch to demonstrate that the ResNet152 utilised in SBCNN will have

Table 5: Comparison with traditional methods on MOUD

Model	MOUD	
	ACC(%)	F1
SVM	57.23	54.83
Naive Bayes	55.72	52.14
GMM	54.66	52.89
HMM	56.63	55.84
DTW	53.92	53.06
AFF-ACRNN	68.74	66.37

Methods	2-class		5-class		7-class	
	Acc(%)	F1	Acc(%)	Macro F1	Acc(%)	Macro F1
LeNet	56.75	55.62	23.67	21.87	15.63	15.12
AlexNet	58.71	57.88	26.43	23.19	19.21	18.79
VGG16	57.88	55.97	27.37	25.78	17.34	16.25
ZFNet	55.37	53.12	21.90	21.38	12.82	11.80
ResNet18	58.94	56.79	25.26	24.63	18.35	17.89
ResNet50	62.52	61.21	28.13	27.04	20.21	20.01
ResNet152	65.42	64.86	28.78	28.08	21.56	20.57

Table 3: Comparison of SBCNN with different structure

Methods	2-class		5-class		7-class	
	Acc(%)	F1	Acc(%)	Micro F1	Acc(%)	Micro F1
UB-LSTM+Res18	67.19	66.37	33.83	31.97	26.78	25.83
UB-LSTM+Res50	67.83	66.69	34.21	33.78	27.75	26.41
UB-LSTM+Res152	68.64	67.94	35.87	34.11	28.15	27.03
UB-BiLSTM+Res18	68.26	66.25	35.43	33.52	27.63	26.09
UB-BiLSTM+Res50	69.18	68.22	36.93	34.67	28.11	27.54
UB-BiLSTM+Res152	69.64	68.51	37.71	35.12	29.26	28.45

Table 4: Comparison of different combinations between SBCNN and UB-BiLSTM

Model	ACC(%)		
	MOSI	→	MOUD
State-of-the-art	60.31		59.99
AFF-ACRNN	69.64		57.74

Table 6: Comparison with state-of-art result (Poria et al.2017) . The right arrow means the model is trained and validated on the MOSI and tested on the MOUD

the best result, for instance, with the weighted accuracy of 65.42% on the dataset, due to its extreme depth and the helpful residual units used to prevent degradation. We selected six best combinations of SBCNN and UB-BiLSTM and find that the best is ResNet152 used in SBCNN with UB-Bilstm, whose weighted accuracy is 69.42% on MOSI and outperforms not only the traditional classifier like SVM, but also the state-of-the-art approach by 9.33% on MOSI dataset. Attention mechanism is used in both branch to subtly combine the heterogeneous acoustic features and choose the feature vectors that have the greatest impact on the sentiment classification. Furthermore, in the experiment of using MOSI as training set and verification set and MOUD as test set, it also shows that our proposed model has strong generalization ability.

5 Conclusion

In this research paper, we propose the AFF-ACRNN, a novel utterance-based deep neural network model that simultaneously combines CNN and LSTM-based networks to produce representative features ASV that can best capture the emotional content of an audio utterance.We extract several traditional heterogeneous acoustic features by Librosa toolkit and choose the four most representative features through a large number of experiments, and regard them as the input of the neural network. The CNN branch and the LSTM branch can both produce CASV and LASV, which we can then combine to produce the final ASV for each utterance's sentiment classification. Feature fusion also uses BiLSTM with an attention mechanism. The experimental findings demonstrate that our model can accurately and quickly recognise audio sentiment, and they also show that our heterogeneous ASV outperforms conventional acoustic features or vectors extracted from other deep learning models. Additionally, experimental results show that the suggested model outperforms the cutting-edge method by 9.33% on MOSI dataset. In order to demonstrate that our model won't be greatly influenced by language types, we also tested it on MOUD. To further discuss the fusion dimension of audio features and take into consideration the fusion of various dimensions of various categories of features, we will combine feature engineering technologies in the future, and even apply them to multimodal sentiment analysis.

References

Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(12): 1-135.

Liu B, ”Sentiment analysis: mining opinions, sentiments, and emotions”, The Cambridge University Press, 2015.

S. Poria, E. Cambria, R. Bajpai, and A. Hussain, ”A review of affective computing: From unimodal analysis to multimodal fusion”, Information Fusion, vol. 37, pp. 98125, 2017.

Kraus, ”M.W. Voice-only communication enhances empathic accuracy”, American Psychologist 72, 7 (2017), 644. S. Zhang, T. Huang and W. Gao, ”Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching,” in IEEE Transactions on Multimedia, vol. PP. 99 (2017):1-1.

S. Ezzat, N. Gayar and M.M. Ghanem, Sentiment Analysis of Call Centre Audio Conversations using Text Classification, in International Journal of Computer Information Systems and Industrial Management Applications, vol. 4, pp. 619-627, 2012.

Van Den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[C]//SSW. 2016: 125.

Bertin-Mahieux, T., and Ellis, D. P. 2011. Large-scale cover song recognition using hashed chroma landmarks. In 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 117120. IEEE.

Kaushik L, Sangwan A, Hansen J H L. Automatic audio sentiment extraction using keyword spotting[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

Mariel W C F, Mariyah S, Pramana S. Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naŁve Bayes for Indonesian text[C]//Journal of Physics: Conference Series. IOP Publishing, 2018, 971(1): 012049.

G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, ADIEU Features? End-to-end Speech Emotion Recognition using A Deep Convolutional Recurrent Network, in IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 52005204.

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

Neumann, Michael, and Ngoc Thang Vu. ”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 (2017). Wang, Zhong-Qiu, and Ivan Tashev. ”Learning utterancelevel representations for speech emotion and age/gender recognition using deep neural networks.” Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

Chen M, He X, Yang J, et al. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition[J]. IEEE Signal Processing Letters, 2018. Poria, Soujanya, et al. ”Context-dependent sentiment analysis in user-generated videos.” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2017.

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

Y. Le Cun, B. Boser et al., Handwritten digit recognition with a back-propagation network, in Advances in neural information processing systems, 1990.

T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, Convolutional, long short-term memory, fully connected deep neural networks, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 45804584.

D.Palaz,R.Collobertetal.,Analysisofcnn-based speechrecognition system using raw speech as input, in Proceedings of Interspeech, 2015.

El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3): 572-587.

Cummins N, Amiriparian S, Hagerer G, et al. An Imagebased deep spectrum feature representation for the recognition of emotional speech[C]//Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017: 478-484. McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference. 2015: 18-25.

Bae S H, Choi I, Kim N S. Acoustic scene classification using parallel combination of LSTM and CNN[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). 2016: 11-15.

Prez-Rosas V, Mihalcea R, Morency L P. Utterance-level multimodal sentiment analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, 1: 973-982.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(2): 295307.
Balamurugan B, Maghilnan S, Kumar M R. Source camera identification using SPN with PRNU estimation and enhancement[C]//Intelligent Computing and Control (I2C2), 2017 International Conference on. IEEE, 2017: 1-6.

www.google.com
www.kaggle.com
https://ceur-ws.org/Vol-2328/3_2_paper_17.pdf