

DEEP MULTIMODAL LEARNING FOR EMOTION RECOGNITION IN SPOKEN LANGUAGE Yue Gu, Shuhong Chen, Ivan Marsic **Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, USA**

Abstract:

We present a novel deep multimodal framework to predict human emotions based on sentence-level spoken language. Our architecture has two distinctive characteristics. First, it extracts the high-level features from both text and audio via a hybrid deep multimodal structure, which considers the spatial information from text, temporal information from audio, and high-level associations from low-level handcrafted features. Second, we fuse all features by using a threelayer deep neural network to learn the correlations across modalities and train the feature extraction and fusion modules together, allowing optimal global fine-tuning of the entire structure. We evaluated the proposed framework on the IEMOCAP dataset. Our result shows promising performance, achieving 60.4% in weighted accuracy for five emotion categories.

Challenges:

- Lack of effective emotional modality-specific features and shared representations.
- ignoring the high-level associations across different modality and cannot guarantee global tuning of the parameters.

Contributions:

- A hybrid deep framework to predict the emotions from spoken language, which consists of ConvNets, CNN-LSTM, and DNN, to extract spatial and temporal associations from the raw text-audio data and low-level acoustic features.
- A four-layer deep neural network to fuse the features and classify the emotions, which allows global fine-tuning of the entire network.
- A detailed comparison with previous work and modality-specific models. 3.

Data Preprocessing:

- Used text as input and extracted the part-of-speech tags (POS) for each sentence using Natural Language Toolkit (NLTK) [1].
- Extracted the Mel-frequency spectral coefficients (MFSCs) from raw audio signals as audio input and extracted the low-level pitch and vocal related features using OpenSmile software [2].
- Evaluated on IEMOCAP including *anger, sad, neutral, frustration*, and *happy (happy+excited).*





Figure 2. Feature extraction structure for MFSC maps

. Text: word2vec + ConvNets with 2, 3, 4, and 5 as the widths.

2. POS: encoded the POS into a 10-dimensional vector and used the same ConvNets structure as the word branch to extract the POS features.

4. LLDs: a three-layer deep neural network of one input layer with two hidden layers to extract the high-level associations from the low-level features.

5) I	Fully- Connected Layer (4096)	Dense Layer (1024)	LSTM (1024)
	Layer5	Layer6	Layer7

- 3. MFSC: CNN-LSTM with seven layers to extract spatial-temporal associations.

Implementation:

- 80-20 training-testing split with speaker independence.
- 2. Rectified linear unit (ReLU) as the activation function.
- 3. Implemented the model with Keras and Tensorflow backend.
- value from categorical cross-entropy loss function.

Results:

- to better performance on *Hap*, *Sad*, *Neu*, and *Fru*.
- weighted accuracy.
- 4. Fine-tuning strategy increases weighted accuracy by 2.7%.
- achieves the best performance, improving accuracy by up to 8%.

Table 1. Accuracy comparis	on of different feature co
(p	ercentage)

(Percentuge)						(Percentage)					
Ang	Hap	Sad	Neu	Fru	Approach	Ang	Hap	Sad	Neu	Fru	
42.9	54.0	50.2	39.7	49.2	BoW+SVM	40.6	45.0	42.2	31.7	44.2	
10.3	33.2	30.3	12.9	39.5	CNN _{word} [16]	42.9	54.2	50.3	39.7	49.2	
51.5	50.6	52.3	43.2	49.2	$LHAF_{wo}+SVM[1]$	41.2	36.6	38.3	39.2	41.5	
54.3	44.1	40.4	39.8	41.7	$LHAF_w + SVM[1]$	40.2	37.1	40.2	40.1	41.8	
47.5	54.1	53.3	<i>A</i> 1 5	10.3	CNN _{mel} [7]	39.7	41.2	43.5	39.1	41.4	
47.5	54.1	55.5	41.5	49.5	$CNN_{word} + LHAF_{w} + MKL[2]$	50.3	52.5	53.2	49.2	52.2	
54.6	59.2	57.2	52.1	54.3	$CNN_{word} + CNN_{mfsc}$ [11]	50.1	52.3	56.3	51.2	50.4	
55.3	52.5	54.2	51.2	52.2	$CNN_{word} + CNN_{mfsc} + SVM$	51.2	50.8	55.3	51.7	51.4	
46.1	40.3	41.3	34.2	40.4	Our Method	57.2	65.8	60.2	56.3	61.6	
37.2	42.8	35.3	27.7	35.4	LHAF _{wo} : Low-level handcrafted acoustic features without feature						
53.7	51.3	51.1	41.3	49.5	selection. $LHAF_w$: Low-level handcraft acoustic feature			feature	s with		
55.7	61.3	57.4	52.6	57.5	feature selection. CNN _{mel} : Using ConvNet as feature extractor and						
55.9	60.2	54.1	50.3	54.3	mel-spectrogram as input data. CNN_{mfsc} : Using ConvNet as feature						
56.1	63.2	60.1	55.4	60.4 extractor and MFSC as input data. <i>MKL</i> : Using multiple kern					kernel		
47.2	42.3	40.1	36.2	40.5	learning as fusion strategy.						
55.3	61.4	57.2	52.3	58.1	Deference						
57.2	65.8	60.2	56.3	61.6						•	
	<i>Ang</i> 42.9 10.3 51.5 54.3 47.5 54.6 55.3 46.1 37.2 53.7 55.7 55.7 55.9 56.1 47.2 55.3 5 7.2	Ang Hap 42.9 54.0 10.3 33.2 51.5 50.6 54.3 44.1 47.5 54.1 54.6 59.2 55.3 52.5 46.1 40.3 37.2 42.8 53.7 51.3 55.7 61.3 55.9 60.2 56.1 63.2 47.2 42.3 55.3 61.4 57.2 65.8	Ang Hap Sad 42.9 54.0 50.2 10.3 33.2 30.3 51.5 50.6 52.3 54.3 44.1 40.4 47.5 54.1 53.3 54.6 59.2 57.2 55.3 52.5 54.2 46.1 40.3 41.3 37.2 42.8 35.3 53.7 51.3 51.1 55.7 61.3 57.4 55.9 60.2 54.1 56.1 63.2 60.1 47.2 42.3 40.1 55.3 61.4 57.2	AngHapSadNeu 42.9 54.0 50.2 39.7 10.3 33.2 30.3 12.9 51.5 50.6 52.3 43.2 54.3 44.1 40.4 39.8 47.5 54.1 53.3 41.5 54.6 59.2 57.2 52.1 55.3 52.5 54.2 51.2 46.1 40.3 41.3 34.2 37.2 42.8 35.3 27.7 53.7 51.3 51.1 41.3 55.7 61.3 57.4 52.6 55.9 60.2 54.1 50.3 56.1 63.2 60.1 55.4 47.2 42.3 40.1 36.2 55.3 61.4 57.2 52.3 57.2 65.8 60.2 56.3	AngHapSadNeuFru 42.9 54.0 50.2 39.7 49.2 10.3 33.2 30.3 12.9 39.5 51.5 50.6 52.3 43.2 49.2 54.3 44.1 40.4 39.8 41.7 47.5 54.1 53.3 41.5 49.3 54.6 59.2 57.2 52.1 54.3 55.3 52.5 54.2 51.2 52.2 46.1 40.3 41.3 34.2 40.4 37.2 42.8 35.3 27.7 35.4 53.7 51.3 51.1 41.3 49.5 55.7 61.3 57.4 52.6 57.5 55.9 60.2 54.1 50.3 54.3 56.1 63.2 60.1 55.4 60.4 47.2 42.3 40.1 36.2 40.5 55.3 61.4 57.2 52.3 58.1 57.2 65.8 60.2 56.3 61.6	AngHapSadNeuFruApproach 42.9 54.0 50.2 39.7 49.2 $BoW+SVM$ 10.3 33.2 30.3 12.9 39.5 $CNN_{word}[16]$ 51.5 50.6 52.3 43.2 49.2 $LHAF_w + SVM[1]$ 54.3 44.1 40.4 39.8 41.7 $LHAF_w + SVM[1]$ 47.5 54.1 53.3 41.5 49.3 54.6 59.2 57.2 52.1 54.3 55.3 52.5 54.2 51.2 52.2 46.1 40.3 41.3 34.2 40.4 37.2 42.8 35.3 27.7 35.4 46.1 40.3 41.3 49.5 55.7 61.3 57.4 52.6 57.7 61.3 57.4 52.6 55.9 60.2 54.1 50.3 54.1 55.4 60.4 47.2 42.3 40.1 36.2 60.1 55.4 60.4 47.2 42.3 40.1 36.2 40.5 55.3 61.4 57.2 55.3 61.4 57.2 52.3 58.1 60.2 56.3 61.6 57.2 65.8 60.2 56.3 61.6 60.2 56.3 61.6	AngHapSadNeuFruApproachAng 42.9 54.0 50.2 39.7 49.2 $BoW+SVM$ 40.6 10.3 33.2 30.3 12.9 39.5 $BoW+SVM$ 40.6 51.5 50.6 52.3 43.2 49.2 $LHAF_{wo}+SVM[1]$ 41.2 54.3 44.1 40.4 39.8 41.7 $LHAF_w+SVM[1]$ 40.2 47.5 54.1 53.3 41.5 49.3 $CNN_{mord}+LHAF_w+MKL[2]$ 50.3 54.6 59.2 57.2 52.1 54.3 $CNN_{word}+CNN_{mfsc}$ 50.1 55.3 52.5 54.2 51.2 52.2 $CNN_{word}+CNN_{mfsc}+SVM$ 51.2 46.1 40.3 41.3 34.2 40.4 $Our Method$ 57.2 46.1 40.3 51.1 41.3 49.5 51.2 $CNN_{word}+CNN_{mfsc}+SVM$ 51.2 55.7 61.3 57.4 52.6 57.5 57.2 $Cur Method$ 57.2 55.9 60.2 54.1 50.3 54.3 54.3 60.4 60.4 47.2 42.3 40.1 36.2 40.5 60.4 $earning$ as fusion strategy. 55.3 61.4 57.2 52.3 58.1 60.2 56.3 61.6 57.2 65.8 60.2 56.3 61.6 60.4 60.2	AngHapSadNeuFru 42.9 54.0 50.2 39.7 49.2 10.3 33.2 30.3 12.9 39.5 51.5 50.6 52.3 43.2 49.2 51.5 50.6 52.3 43.2 49.2 54.3 44.1 40.4 39.8 41.7 47.5 54.1 53.3 41.5 49.3 54.6 59.2 57.2 52.1 54.3 54.6 59.2 57.2 52.1 54.3 54.6 59.2 57.2 52.1 54.3 54.6 59.2 57.2 52.1 54.3 57.3 52.5 54.2 51.2 52.2 46.1 40.3 41.3 34.2 40.4 37.2 42.8 35.3 27.7 35.7 51.3 51.1 41.3 49.5 55.7 61.3 57.4 52.6 57.5 55.9 60.2 54.1 50.3 54.3 56.1 63.2 60.1 55.4 60.4 47.2 42.3 40.1 36.2 40.5 55.3 61.4 57.2 52.3 58.1 57.2 65.8 60.2 56.3 61.6	AngHapSadNeuFruApproachAngHapSad 42.9 54.0 50.2 39.7 49.2 $BoW+SVM$ 40.6 45.0 42.2 10.3 33.2 30.3 12.9 39.5 $CNN_{word}[16]$ 42.9 54.2 50.3 51.5 50.6 52.3 43.2 49.2 $CNN_{word}[16]$ 41.2 36.6 38.3 54.3 44.1 40.4 39.8 41.7 $LHAF_w + SVM[1]$ 40.2 37.1 40.2 47.5 54.1 53.3 41.5 49.3 $CNN_{word} + CNN_{mfsc}[11]$ 50.1 52.3 56.3 54.6 59.2 57.2 52.1 54.3 $CNN_{word} + CNN_{mfsc}[11]$ 50.1 52.3 56.3 55.3 52.5 54.2 51.2 52.2 $CNN_{word} + CNN_{mfsc}[11]$ 50.1 52.3 56.3 55.7 61.3 57.4 52.6 57.5 57.2 65.8 60.2 54.3 55.9 60.2 54.1 50.3 54.3 54.3 55.4 60.4 47.2 42.3 40.1 36.2 40.5 55.3 61.4 57.2 52.3 58.1 57.2 65.8 60.2 56.3 61.6 61.6 61.6 61.6 61.6 61.6	AngHapSadNeuFruApproachAngHapSadNeu 42.9 54.0 50.2 39.7 49.2 $BoW+SVM$ 40.6 45.0 42.2 31.7 10.3 33.2 30.3 12.9 39.5 $CNN_{word}[16]$ 42.9 54.2 50.3 39.7 51.5 50.6 52.3 43.2 49.2 $LHAF_{wo}+SVM[1]$ 41.2 36.6 38.3 39.2 54.3 44.1 40.4 39.8 41.7 $LHAF_w+SVM[1]$ 40.2 37.1 40.2 40.1 47.5 54.1 53.3 41.5 49.3 $CNN_{word}+LHAF_w+SVM[1]$ 40.2 37.1 40.2 40.1 55.3 52.5 54.2 51.2 52.2 $CNN_{word}+CNN_{mfsc}[11]$ 50.1 52.3 56.3 51.2 55.3 51.3 51.1 41.3 49.5 57.2 65.8 60.2 56.3 51.2 55.7 61.3 57.4 52.6 57.5 57.5 56.3 51.4 60.4 57.2 65.8 60.2 56.3 55.9 60.2 54.1 50.3 54.3 54.3 $CNN_{mel}:$ $CNN_{mel}:$ Cus -level handcrafted acoustic features without 55.7 61.3 57.4 50.3 54.3 57.5 55.3 60.2 56.3 61.6 57.2 65.8 60.2 56.3 61.6 60.4 67.2 65.8 60.2 56.3	

*CNN*_{word}: Using ConvNet as feature extractor and text as input; CNN_{pos}: Using ConvNet as feature extractor and part-of-speech tags as input data; CNN_LSTM_{mfsc} : Using CNN-LSTM as feature extractor and MFSC energy maps as input data; DNN_{lhaf}: Using DNN as feature extractor and low-level handcraft features as input data; *Both_text*: Including both *CNN_{word}* and *CNN_{pos}*; *Both_audio*: Including both *CNN_LSTM_{mfsc}* and *DNN_{lhaf}*.



Supervised by: Ivan Marsic Multimedia and Image Processing Lab

4. Initialized the learning rate at 0.01 and use Adam optimizer to minimize the

1. The spatial-temporal high-level acoustic features extracted from the CNN-LSTM lead

2. The DNN_{lhaf} achieves the best result on Ang category in all unimodal structures.

3. Combining all the features from four branches achieves the best result, with 60.4%

5. Compared with previous approaches, the proposed hybrid deep multimodal structure

 Table 2. Comparision of previous emotion recognition structures
(nercentage)

||| Bird, Steven. "NLTK: the natural language toolkit." In Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69-72. Association for Computational Linguistics,

[2] Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor." In Proceedings of the 18th ACM international conference on Multimedia, pp. 1459-1462. ACM, 2010.