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

RUTGERS School of Engineering

Supervised by: Ivan Marsic
Multimedia and Image Processing Lab

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 three-layer 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:

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

Contributions:

1. 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.
2. A four-layer deep neural network to fuse the features and classify the emotions, which allows global fine-tuning of the entire network.
3. A detailed comparison with previous work and modality-specific models.

Data Preprocessing:

1. Used text as input and extracted the part-of-speech tags (POS) for each sentence using Natural Language Toolkit (NLTK) [1].
2. 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].
3. Evaluated on IEMOCAP including *anger, sad, neutral, frustration*, and *happy* (*happy+excited*).

System Structure:

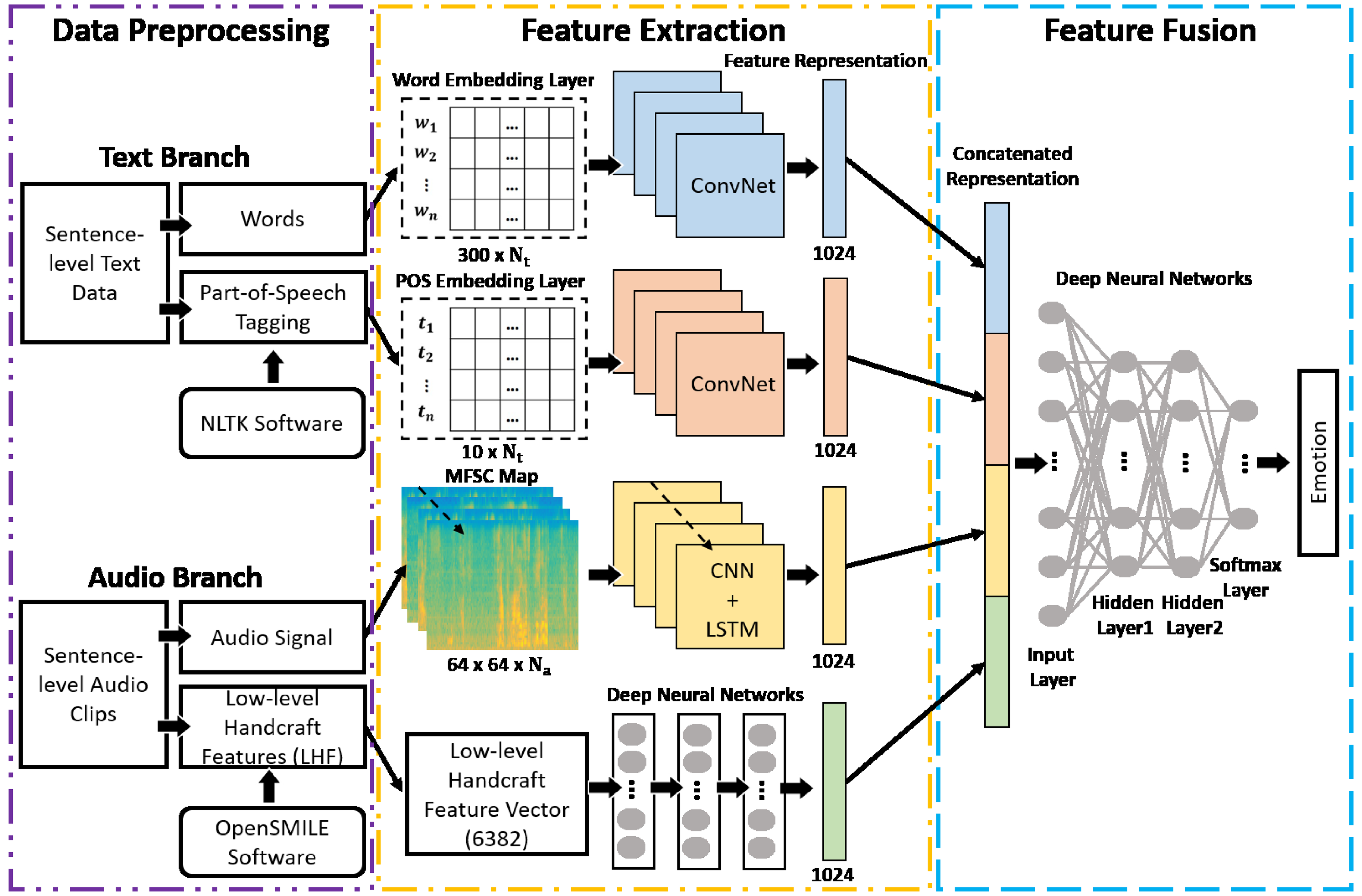

Figure 1. Overall structure of the proposed deep multimodal framework

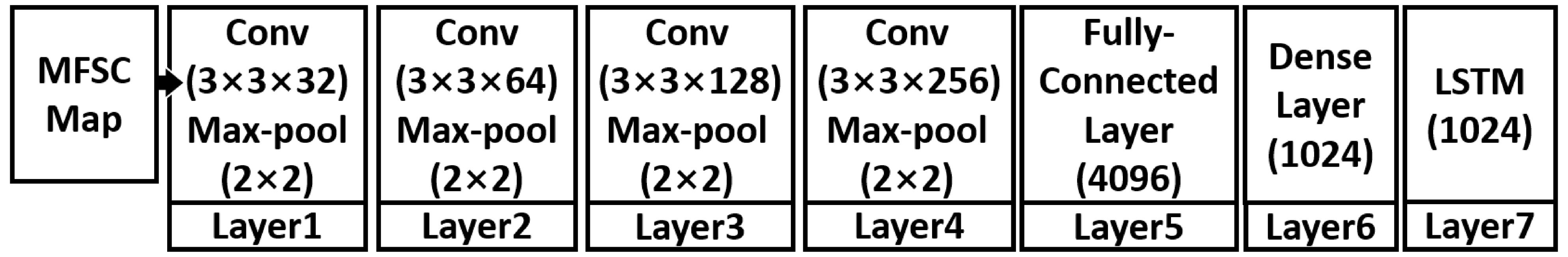

Figure 2. Feature extraction structure for MFSC maps

1. 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.
3. MFSC: CNN-LSTM with seven layers to extract spatial-temporal associations.
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.

Implementation:

1. 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.
4. Initialized the learning rate at 0.01 and use Adam optimizer to minimize the value from categorical cross-entropy loss function.

Results:

1. The spatial-temporal high-level acoustic features extracted from the CNN-LSTM lead to better performance on *Hap*, *Sad*, *Neu*, and *Fru*.
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% weighted accuracy.
4. Fine-tuning strategy increases weighted accuracy by 2.7%.
5. Compared with previous approaches, the proposed hybrid deep multimodal structure achieves the best performance, improving accuracy by up to 8%.

Table 1.Accuracy comparison of different feature combinations (percentage)

Approach	*Ang*	*Hap*	*Sad*	*Neu*	*Fru*
CNN_{word}	42.9	54.0	50.2	39.7	49.2
CNN_{pos}	10.3	33.2	30.3	12.9	39.5
CNN_LSTM_{mfsc}	51.5	50.6	52.3	43.2	49.2
DNN_{lhaf}	54.3	44.1	40.4	39.8	41.7
$CNN_{word} + CNN_{pos}$	47.5	54.1	53.3	41.5	49.3
$CNN_{word}+CNN\_LSTM_{mfsc}$	54.6	59.2	57.2	52.1	54.3
$CNN_{word}+DNN_{lhaf}$	55.3	52.5	54.2	51.2	52.2
$CNN_{pos}+CNN\_LSTM_{mfsc}$	46.1	40.3	41.3	34.2	40.4
$CNN_{pos}+DNN_{lhaf}$	37.2	42.8	35.3	27.7	35.4
$CNN\_LSTM_{mfsc}+DNN_{lhaf}$	53.7	51.3	51.1	41.3	49.5
Both_text+CNN_LSTM_{mfsc}	55.7	61.3	57.4	52.6	57.5
Both_text+DNN_{lhaf}	55.9	60.2	54.1	50.3	54.3
CNN_{word}+*Both_audio*	56.1	63.2	60.1	55.4	60.4
CNN_{pos}+*Both_audio*	47.2	42.3	40.1	36.2	40.5
Our Method_Separate	55.3	61.4	57.2	52.3	58.1
Our Method _Together	**57.2**	**65.8**	**60.2**	**56.3**	**61.6**

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}.

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

Approach	*Ang*	*Hap*	*Sad*	*Neu*	*Fru*
BoW+SVM	40.6	45.0	42.2	31.7	44.2
CNN_{word}[16]	42.9	54.2	50.3	39.7	49.2
$LHAF_{wo}+SVM$[1]	41.2	36.6	38.3	39.2	41.5
$LHAF_{w}+SVM$[1]	40.2	37.1	40.2	40.1	41.8
CNN_{mel} [7]	39.7	41.2	43.5	39.1	41.4
$CNN_{word}+LHAF_{w}+MKL$[2]	50.3	52.5	53.2	49.2	52.2
$CNN_{word}+ CNN_{mfsc}$ [11]	50.1	52.3	56.3	51.2	50.4
$CNN_{word}+CNN_{mfsc}+SVM$	51.2	50.8	55.3	51.7	51.4
Our Method	**57.2**	**65.8**	**60.2**	**56.3**	**61.6**

$LHAF_{wo}$: Low-level handcrafted acoustic features without feature selection. $LHAF_{w}$: Low-level handcraft acoustic features with feature selection. CNN_{mel}: Using ConvNet as feature extractor and mel-spectrogram as input data. CNN_{mfsc}: Using ConvNet as feature extractor and MFSC as input data. *MKL*: Using multiple kernel learning as fusion strategy.

Reference:

[1] Bird, Steven. "NLTK: the natural language toolkit." In Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69-72. Association for Computational Linguistics, 2006.

[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.