Attention based 3-D Convolution Neural Network for Speech Emotion Recognition

Meiying Chen

Goergen Institute for Data Science, University of Rochester

Abstract

Speech Emotion Recognition is important in many fields like human-machine interaction and health care. Yet SER remains a difficult task due to several obstacles, e.g. the sparsity of emotion-relevant frames in a long speech utterance, the difficulties of bettering extracted acoustic feature quality, and the long-dependencies problem in RNN models. In this paper:

- We apply a 3-D Mel-spectrograms
- We employ a convolution Neural Network with attention mechanism, which helps to generate higher quality features and alleviate the sparsity issue

Introduction

Recently, SER researchers utilize CNN to generate hierarchical features from spectrogram and employ LSTM-RNN to futher extract temporal information. This combination is called CRNN architecture. In some cases, CNN and RNN are also paralleled. But not all time-frequency units contribute equally to the whole utterance's emotional state. Attention mechanism avoids emphasis too much on the data point being close to one another by allowing the model to automatically search for parts of the source that are relevant to predicting the target, without fixing the length of the vectors used.

Another concern is that CNNs use personalized features as inputs, which are greatly influenced by the style of speaking, and not robust to the variation of speakers, speak contents and environments.We propose using log-Mel, deltas, delta-deltas as 3-D inputs. We believe that measured deltas and delta-deltas are capable of representing the changing emotions and maintaining effective emotional knowledge while decreasing the impact of non-relevant emotional factors such as speakers, contents, and environment.

Algorithm Description

- 3-D Log-Mels Features
  we use the log-Mels with deltas and delta-deltas as the CNN input, where the deltas and delta-deltas reflect the process of emotional change.

$$m_i = log(p_i) \tag{1}$$

$$m_i^d = \frac{\sum_{n=1}^{n} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{n} n^2} \tag{2}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{n} n(m_{i+n}^d - m_{i-n}^d)}{2\sum_{n=1}^{n} n^2} \tag{3}$$

- 3-D CNN Layers

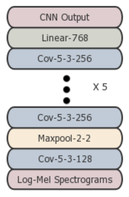

Figure 1:CNN blocks.

- Attention Layers
  The attention layer is worked to extract important elements in input sequences that are most emotion-relevant.

$$e_i = u^T tanh(Wa_i + b) \tag{4}$$

$$\alpha_i = \frac{exp(\lambda e_i)}{\sum_{k=1}^{L} exp(\lambda e_i)} \tag{5}$$

$$c = \sum_{i=1}^{L} \alpha_i a_i \tag{6}$$

$\{a_1, ..., a_L\}$ is the output of the linear layer. Feed it into MLP with tanh to get new representation of a.Then measure the importance weight e_i between this new vector and a learnable vector u. Normalized importance weight α_i. Scale factor $= 0.3$. Lastly, the utterance emotion vector c is the weighted sum of set a multiplied by important weights.

Experimental Set

- Dataset IEMOCAP is one of the most widely used databases on speech emotion recognition, which contains five sessions, each of them includes dialogues between two professional actors.We only focus on improvised speech data and use four emotional categories. The average duration of IEMOCAP speech recordings is 4.5 s.

Session	Neutral	Happy	Sad	Angry	Sum
1	223	132	104	62	521
2	217	191	100	22	530
3	198	149	190	90	627
4	174	195	81	84	534
5	287	280	133	31	731
Sum	1099	947	608	289	2943

Table 1:Instance distribution for the IEMOCAP Dataset.

- Baselines
  a) CNN in Soujanya, 2016. The method first extracts features from visual and textual modalities using CNN and then feeds such features to a multiple kernel learning classifier.
  b) CRNN model in Zhao, 2017. This is a convolutional LSTM-RNN architecture.
  c) CNN+LSTM+Attention Model with 2-D spectrogram features in Chen, 2018.

Results

As the dataset contains 10 speakers, we plan to employ 10-fold cross-validation to assess our model. 8 of 10 speakers are chosen as train set. For the rest 2 speakers, one is dev set and the other one is test set. We report the unweighted average recall(UAC) on the test set. Note only 1-fold test result is reported:

Model	UAC
DNN in [10]	51.24%
CNN+LSTM in [18]	59.40%
CNN+LSTM+Attention in [4]	62.47%
3-D CNN+Attention(our proposed)	63.93%

Table 2:UAC of Comparative Models.

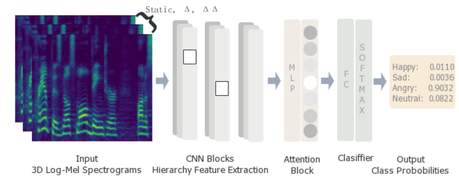

Figure 2:Overview of our proposed network architecture.

Conclusion

In this paper, we proposed an attention-based CNN model on 3-D Mel-spectrograms for speech emotion recognition. This process helps to generate higher quality features and alleviate the emotion sparsity issue of SER. The experimental results show our proposed model outperforms the baseline models, and the model size is smaller than LSTM-based ones.

Future Work

- Read more papers published 2019
- Combining the traditional model with DL
- Find a way to visualize and analyse why these models work or not work

Reference

[4] Chen et al. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition.IEEE Signal Processing Letters, 2018.

[10] Poria et al. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. ICDM,2016.

[18] Zhao et al. Recurrent convolutional neural network for speech processing. ICASSP, 2017.

air AUDIO INFORMATION RESEARCH

GOERGEN INSTITUTE FOR DATA SCIENCE