Journal of Southeast University (English Edition) Vol. 33, No. 4, pp. 444 - 447 Dec. 2017 ISSN 1003—7985

Multimodal emotion recognition based on deep neural network

Ye Jiayin Zheng Wenming Li Yang Cai Youyi Cui Zhen

(School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210029, China)

Abstract: In order to increase the accuracy rate of emotion recognition in voice and video, the mixed convolutional neural network (CNN) and recurrent neural network (RNN) are used to encode and integrate the two information sources. For the audio signals, several frequency bands as well as some energy functions are extracted as low-level features by using a sophisticated audio technique, and then they are encoded with a one-dimensional (1D) convolutional neural network to abstract high-level features. Finally, these are fed into a recurrent neural network for the sake of capturing dynamic tone changes in a temporal dimensionality. As a contrast, a two-dimensional (2D) convolutional neural network and a similar RNN are used to capture dynamic facial appearance changes of temporal sequences. The method was used in the Chinese Natural Audio-Visual Emotion Database in the Chinese Conference on Pattern Recognition (CCPR) in 2016. Experimental results demonstrate that the classification average precision of the proposed method is 41.15%, which is increased by 16.62% compared with the baseline algorithm offered by the CCPR in 2016. It is proved that the proposed method has higher accuracy in the identification of emotional information.

Key words: emotion recognition; convolutional neural network (CNN); recurrent neural networks (RNN)

DOI: 10.3969/j.issn.1003-7985.2017.04.009

In recent years, many researchers have devoted themselves to multimodal emotion recognition research[1-5]. For audio emotion recognition, a common practice is fusing the most typical audio emotion feature to form high-dimensional emotion feature sets, thus improving the performance of the recognition system. For video, there are many common feature extraction methods of dynamic image sequence, such as the optical flow method, LBP-TOP[6] features and so on. Compared with a single modality, multimodal emotion recognition is more complete and more robust. Petridis et al.[7] studied the video-audio emotion recognition based on model layer fusion and decision layer fusion and obtained the best accuracy of 96% and 98%. Wang et al.[8] used kernel cross modal factor analysis (KCFA) to reduce the feature dimension and fuse the feature. Metallinou et al.[9] proposed multimodal emotion recognition based on the policy maker fusion and Gaussian mixture model. Although multimodal emotion recognition has achieved initial results, it still requires further research.

In this paper, the convolutional neural network (CNN) was adopted to extract the high-level features of the spatial domain and the RNN was applied to build a dynamic model of the time domain to deal with the audio-visual data on the Chinese natural audio-visual emotion database (CHEAVD)[10].

1 Model of Multimodal Emotion Recognition

1.1 Segmentations and MFCC features extraction

Each audio sample is divided into several speech frames with fixed frame length and the adjacent two frames are shared fixed sampling points to keep sufficient global and local information. For each speech frame, we use open-SMILE[11] (speech and music interpretation by large-space extraction) software employed by the INTERSPEECH 2010 paralinguistic challenge[12] to extract the MFCC features which encode the target audio sequence in several frequency bands as well as some energy functions.

1.2 Spatial-temporal features extraction

A feature map is obtained by the repeated application of a function across the entire audio segments. If we denote the k-th feature map at a given layer as $\boldsymbol{h}^k$, their filters are determined by the weights $\boldsymbol{W}^k$ and the bias $\boldsymbol{b}_k$, then the feature map $\boldsymbol{h}^k$ is obtained as follows:

$$\boldsymbol{h}_i^k = f(\boldsymbol{W}^k \boldsymbol{x}_i + \boldsymbol{b}_k) \tag{1}$$

where $\boldsymbol{x}_i$ is the input pixel of the corresponding layer at coordinates i.

To effectively learn contextual dependencies among sequential data, the RNN is chosen owing to its feedback connections. Rectified linear units $f(\cdot)$ is added to have a higher level of nonlinearity. The output nodes $\boldsymbol{y}_i$ and the hidden nodes $\boldsymbol{h}_i$ can be calculated as

$$\boldsymbol{y}_i = f(\boldsymbol{U}\boldsymbol{h}_i) \tag{2}$$

$$\boldsymbol{h}_i = f(\boldsymbol{V}\boldsymbol{x}_i + \boldsymbol{W}\boldsymbol{h}_i) \tag{3}$$

where $\boldsymbol{W}$ and $\boldsymbol{V}$ are the weight matrices of $\boldsymbol{x}$ and $\boldsymbol{h}$; $f(\cdot)$ is the max value of input.

As a result, an end-to-end network combined convolu-

Received 2017-05-16, **Revised** 2017-08-25.

Biographies: Ye Jiayin (1993—), female, graduate; Zheng Wenming (corresponding author), male, doctor, professor, wenming_zheng@seu.edu.cn.

Citation: Ye Jiayin, Zheng Wenming, Li Yang, et al. Multimodal emotion recognition based on deep neural network[J]. Journal of Southeast University (English Edition), 2017, 33(4): 444 - 447. DOI: 10.3969/j.issn.1003-7985.2017.04.009.

tional layers over 1D space and RNN layers are designed to obtain high-level spatial-temporal features.

1.3 Spatial-temporal features extraction

To further characterize the globalism of hidden states, alignment is adopted to assign weights to each hidden state from temporal features and a nonlinear transform is introduced to explore a deep relationship among them. The concrete process can be written as

$$\boldsymbol{h} = \sum_t \boldsymbol{a}_t \boldsymbol{s}_t \tag{4}$$

$$\boldsymbol{a}_t = \exp(\boldsymbol{W}\boldsymbol{s}_t + \boldsymbol{b}) \tag{5}$$

where $\boldsymbol{s}_t$ is the state of hidden unit; $\boldsymbol{a}_t$ is the weight vector; $\boldsymbol{W}$ is the transform matrix; and $\boldsymbol{b}$ is the bias.

1.4 Face region cropping

The state-of-art detector faster-RCNN method[13] is adopted to crop the face region and resize it. To confirm reliability, we check the cropped face images one by one and manually fix the error results.

1.5 Fine-tuning on a pre-trained model

For the sake of simplicity, we choose the existing VGG _Face16 as our pre-trained model, which consists of five stacks of Conv-Net, three fully-connected layers and one soft-max layer, for its robust performance. As for merely extracting the spatial facial features, we only use the output of the last fully-connected layer. The parameters of the first several convolutional layers are fixed while the final three fully-connected layers' parameters are set randomly and they change after several iterations.

1.6 Features rearrangement

Since the lengths of each video are variable, we cannot simply combine the feature of every frame together. To solve this problem, we calculate the discrimination of features between the adjacent two frames in time order. If the difference is very small, the changes of emotions in these two frames are so insignificant that they can be discarded. In contrast, the second frame will be saved due to the high intensity variation of facial information. Owing to this strategy, we can obtain a certain number of frames with the largest changes. After selection, we rearrange the selected frames in time order to keep the context of the frames. If the selected frames are smaller than the fixed length of the feature sequence, the first frame is copied several times to make the length of the whole sequence up to the fixed length and the copies are inserted into the beginning of the whole sequence to maintain the temporal relationship of the sequence features.

1.7 Multimodal emotion recognition

As for audio and video from the same clips having the supplementary information in expressing emotions, we fuse the two results to make multimodal emotion recognition. After obtaining the soft-max layer scores of the audio network and video network, we compute the score according to the following criterion to determine the emotion label of the testing sample.

$$C = \arg\max(\alpha s_{\text{audio}}(i) + \beta s_{\text{video}}(i)) \tag{6}$$

where $s(i)$ is the score of the i-th class; α and β are the weight of scores in two modalities; and C denotes the prediction of the classification category.

2 Experiments

2.1 Database and data augment

In this section, CHEAVD[10] was used to evaluate the performance of the proposed method. It contained 2 582 video clips including happy, sad, angry, worried, anxious, surprise, disgust and neutral emotions and consisted of audio, video and multimodal emotion recognition tasks.

We abandoned the first or final 4 000 to 8 000 sampling points of the produced noisy speech signals so that we could obtain a large number of augmented speech samples. 800 samples of each class from the augmented dataset were randomly selected to supply samples of each emotion category, which enables the number of samples in each emotion category to be approximately equal. Since the first or final several frames of a video sequence may have neutral expressions, we selected part of the middle frames manually from the video sequence for the next step. Meanwhile, for balancing the number of samples in each emotion, some samples from the multi-PIE database was added. When the number of emotions in multi-PIE was 5 rather than 8, we copied several pictures according to the serial number from the corresponding classes of the remaining three classes in CHEAVD to ensure that the order of magnitudes of eight classes were the same. Meanwhile, we adopted two tactics, rotation and mirroring part of the selected sequences, to guarantee that the numbers of sequences in all classes were balanced.

2.2 Configures and parameters.

The structure and parameters of the audio sub-challenge network are shown in Fig. 1 and Tab. 1.

The batch size, momentum, total epoch, learning rate, learning decay and weight regression of our network were fixed at 40, 0.9, 600, 0.01, 0.1 and 0.01, respectively. For the video part, after data augmentation, there were approximately 23 000 training images in each class. In the stage of fine-tuning, we set the batch-size and learning rate to be 32 and 0.005. Moreover, the momentum term weight and weight decay factor were fixed at the default value, which are 0.9 and 0.001 5. Finally, a dropout with a rate of 0.5 was applied to the fully connected layers.

When preparing the input of the BRNN, 40, 50, 60, 70 and 80 are tried to be the fixed length of the sequence and 50 is set to be the final threshold due to its better performance. That is to say, each image will be rotated in sequence by $\{\pm 7, \pm 15\}^\circ$ or mirroring the images in some sequences, and hence the numbers of eight emotion sequences are 798, 801, 585, 614, 723, 768, 737 and 744, respectively.

For the BRNN configurations, we set the batch-size and learning rate to be 256 and 0.005. The momentum term weight and weight decay factor were fixed at the values of 0.9 and 0.000 5.

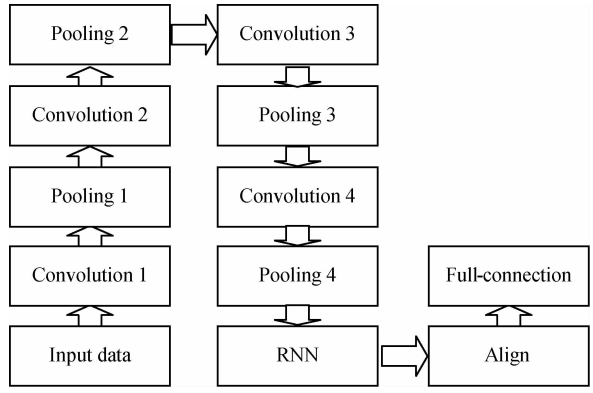

Fig. 1 The network structure for audio sub-challenge

Tab 1 Parameters of network for audio sub-challenge

Layer name	Conv1	Pool 1	Conv 2	Pool 2	Conv 3	Pool 3	Conv 4	Pool 4	RNN	Align	FC
Kernel size of layer	3	2	3	2	3	2	3	2			
Kernel number	32		32		64		64				
Stride of kernel	1	2	1	2	1	2	1	2			
Dimensionality of layer	1 582	791	791	395	395	197	197	98	512	512	512

3 Results

The challenge organizers provided two evaluation indicators, including average precision and classification accuracy. Tab. 2 and Tab. 3 depict the best results of audio, video, and multimodal emotion recognition among our submitted ones as well as the baselines provided by the challenge organizers, SYSU[14], which proposed a residual network architecture within the convolutional neural networks framework. Another system i.e., LBPTOP, DCNN and LRCN[15] combined traditional features and deep learning features to address the challenge from the MEC 2016 in terms of average precision and classification accuracy.

Tab. 2 Classification accuracy of test sets in emotion recognition %

Method	Audio	Video	Multimodal
Baseline	24.36	19.59	24.52
SYSU[14]	26.11	27.38	29.93
LBPTOP + DCNN + LRCN[15]	27.87	24.68	19.90
Proposed	31.53	31.85	34.55

Tab. 3 Classification average precision of test sets in emotion recognition %

Method	Audio	Video	Multimodal
Baseline	24.02	34.28	24.53
SYSU[14]	25.98	36.56	36.42
LBPTOP + DCNN + LRCN[15]	26.33	28.35	25.35
Proposed	37.06	37.63	41.15

From Tab. 2 and Tab. 3, it is clear to see that we achieve the classification accuracy of 31.53% and the average precision of 37.06% in audio emotion recognition, which has an improvement of more than 5% over the baseline and other methods. Furthermore, our method has the classification accuracy of 31.85% and the average precision of 37.63% in video emotion recognition tasks. In this case, our results are also higher than the baseline and other methods in terms of both evaluation indicators, which increases by 1% or more. Finally, in the multimodal emotion recognition sub-challenge, we have the accuracy of 34.55% and the average precision of 41.15%. Compared with the baseline, it can be clearly seen that our average precision achieved 16.62% improvement and it is 4.73% higher than that of SYSU[12], and 15.8% higher than that of LBPTOP + DCNN + LRCN. In addition, from the three sub-challenge results, it is interesting to see that in terms of the recognition rates including accuracy and average precision, the results of the multimodal case are clearly better than the single modal ones, regardless of audio or video. It is believed that integrating multimodalities benefits emotion recognition performance and one certain modality may contain the discriminative ability for distinguishing some emotions that the other lacks.

4 Conclusion

In this paper, we propose a multimodal emotion recognition method based on a deep neural network, which combines the structure of 1D CNN, alignment and B_RNN. The voice feature is extracted with the openSMILE and the image feature is extracted with adjustive VGG16. Then, the RNN layer is used to find the association between the front and rear frames of the voice and images. All the features are combined to build a new feature to complete emotion recognition. The results from the Chinese Conference on Pattern Recognition in 2016 show that the proposed method is suitable for recognizing emotion.

References

[1] Zeng Z, Pantic M, Huang T S. Emotion recognition based on multimodal information [M]//A*ffective Information Processing*. Springer, 2009: 241 - 265. DOI: 10.1007/978-1-84800-306-4_14.

[2] Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(1): 39 - 58. DOI: 10.1109/TPAMI.2008.52.

[3] 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. DOI: 10.1016/j.patcog.2010.09.020.

[4] Chen L, Mao X, Xue Y, et al. Speech emotion recognition: Features and classification models [J]. *Digital Signal Processing*, 2012, **22**(6): 1154 - 1160. DOI: 10.1016/j.dsp.2012.05.007.

[5] Yan J, Wang X, Gu W, et al. Speech emotion recognition based on sparse representation [J]. *Archives of Acoustics*, 2013, **38**(4): 465 - 470. DOI: 10.2478/aoa-2013-0055.

[6] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 2007, **29**(6): 915 - 928. DOI: 10.1109/TPAMI.2007.1110.

[7] Petridis S, Gunes H, Kaltwang S, et al. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities [C]//*Proceedings of the* 2009 *International Conference on Multimodal Interfaces*. Cambridge, MA, USA, 2009: 23 - 30. DOI: 10.1145/1647314.1647321.

[8] Wang Y, Guan L, Venetsanopoulos A N. Audiovisual emotion recognition via cross-modal association in kernel space[C]// 2011 *IEEE International Conference on Multimedia and Expo* (*ICME*). Barcelona, Spain, 2011: 6011949-1 - 6011949-6. DOI: 10.1109/icme.2011.6011949.

[9] Metallinou A, Lee S, Narayanan S. Audio-visual emotion recognition using gaussian mixture models for face and voice[C]//*Tenth IEEE International Symposium on Multimedia*. Berkeley, CA, USA, 2008: 250 - 257. DOI: 10.1109/ism.2008.40.

[10] Li Y, Tao J, Schuller B, et al. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016 [M]//*Pattern Recognition*. Springer, 2016: 667 - 678.

[11] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor [C]//*Proceedings of the* 18*th ACM International Conference on Multimedia*. Firenze, Italy, 2010: 1459 - 1462. DOI: 10.1145/1873951.1874246.

[12] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2010 paralinguistic challenge [C]//11*th Annual Conference of the International Speech Communication Association*. Makuhari, Chiba, Japan, 2010: 2795 - 2798.

[13] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **39**(6): 1137 - 1149. DOI: 10.1109/TPAMI.2016.2577031.

[14] He G, Chen J, Liu X, et al. The SYSU system for CCPR 2016 multimodal emotion recognition challenge [M]// *Pattern Recognition*. Springer, 2016: 707 - 720.

[15] Sun B, Xu Q, He J, et al. Audio-video based multimodal emotion recognition using SVMs and deep learning[M]// *Communications in Computer and Information Science*, 2016: 621 - 631. DOI: 10.1007/978-981-10-3005-5_51.

基于深度神经网络的多模态情感识别

叶佳音　郑文明　李　阳　蔡友谊　崔　振

(东南大学生物科学与医学工程学院, 南京 210029)

摘要: 为了提升音频和视频载体中的情感识别准确率,采用混合卷积神经网络和递归神经网络编码和集成视频与音频信息来源.通过智能的音频技术,从音频信号提取底层特征,然后用一维卷积神经网络抽象出高级特征,最后送入递归神经网络捕捉时间维度上的语调变化.作为对比,使用二维卷积神经网络和一个类似的卷积神经网络捕捉动态面部外观变化.该方法在2016年度中国模式识别会议提供的中国视觉与听觉情感数据库上达到了41.15%的平均精确度,相比会议基准算法的准确率提升了16.62%.证明所采用方法在情感信息识别中有更高的准确性.

关键词: 情感识别;卷积神经网络;递归神经网络

中图分类号: TP751