Paper:

The Multi-modal Emotion Recognition Based on Text and Image

Wenlong Li[1,2],Kaoru Hirota[1,2], Xingwang Liu[1,2], Yaping Dai[1,2], Zhiyang Jia[1,2,*]

[1]School of Automation, Beijing Institute of Technology, Beijing 100081, China
E-mail: 3120180897@bit.edu.cn
[2]State Key Laboratory of Intelligent Control and Decision of Complex System, Beijing 100081, China
E-mail: 3120180897@bit.edu.cn

Abstract.The Multi-modal emotion recognition based on text and image (MMER) is proposed to solve the problem of inaccurate emotion recognition and poor model robustness of a single modality such as text, image or speech. The Multi-modal emotion recognition based on text and image compares the shallow features of text and image by cosine similarity, and inputs the obtained results to the decision-making layer, and participates in the final emotional decision-making together with the respective results of text and image. The experimental data set is made by ourselves, and each row includes an image, a sentence of text and the emotional label. Results of experiments on the dataset show that the Macro-F1 score for the multi-modal model based on text and image is 73.54, achieving 6.4% and 11.8% improvement compared with the text emotion recognition model various LSTM and the image emotion recognition model ResNet.

Keywords: Emotion recognition, Multi-modal, Text, Image, Deep learning

1. INTRODUCTION

Emotional interaction has received a lot of attention in the research of human-computer natural interaction, and emotion recognition is the key to human-computer emotional interaction. Its research purpose is to make machines perceive human emotions and improve the level of humanization of machines. For image-based emotion recognition, it can generally be divided into traditional methods and deep learning-based methods[1]. Traditional methods intuitively use artificially constructed feature tags to identify emotional states[2]. Some methods use changes in data structure or improve the effect of feature extractors to improve the accuracy of facial expression recognition[3]. The methods based on deep learning usually perform feature extraction in a data-driven manner. Li Yong et al[4] proposed an improved LeNet-5 convolutional neural network to extract low-level features and high-level features of facial expressions for emotion recognition. Yao Naiming et al[5] proposed a face image generation network based on the Wasserstein generation confrontation network to solve the problem of face occlusion in a large range. Tan Xiaohui et al[6]proposed a facial expression recognition method based on multi-scale detail enhancement, and proposed a local gradient feature calculation method using hierarchical structure.

In the field of text emotion recognition based on deep learning, Duyu Tang et al[7] proposed a deep memory network to capture the importance of context words for aspect-level sentiment analysis. Compared with recurrent neural network models (such as LSTM), this method is simple and fast. Experimental results on two data sets prove that the performance of this method is comparable to the SVM system based on the latest features, and is better than the LSTM architecture. Yequan Wang et al[8] proposed an attention-based LSTM for aspect-level emotion classification. The main idea of these suggestions is to learn the embedding of aspects and let each aspect participate in calculating the attention weight. When given different aspects, our proposed model can focus on different parts of the sentence, thus making them more competitive in aspect level classification.

In order to obtain accurate feature information from emotional multi-modal information and improve the accuracy of emotion recognition, many domestic and foreign researchers have done a lot of research on dual-modal fusion emotion recognition[9]. Xiao et al[10] proposed a multi-sensor data fusion method based on a new evidence confidence measure and confidence entropy. Yoshida et al[11] proposed a multi-modal emotion calculation model based on facial expressions, speech, text and other information to realize emotion measurement in the process of online learning.

The multi-modal model based on text and image is a two-branch neural network model that uses both image and text as input. It solves the problem than in a noisy or dim environment, once the voice or face cannot be clearly recognized, emotion recognition will not work. At the same time, the model add a data fusion layer after the output layer of the network, and input feature vectors of different levels into the data fusion layer to participate in the fusion decision of the final emotion classification, which increases the robustness of the model and further improves the emotion recognition Accuracy.

Experiments are conducted on Unbutu16.04 system by using deep learning frame called Pytorch programmed by Python3.6. The experimental dataset is made by ourselves, and it contains fifty thousand rows of data. Each row includes an image, a sentence of text and the emo-

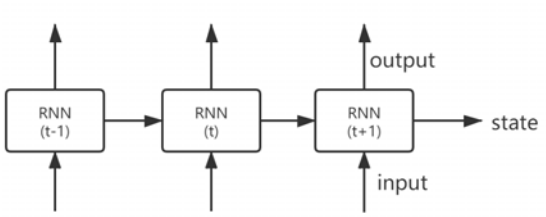

Fig. 1. A simple RNN model.

tional label. There are five types of labels, with Arabic numerals 0-5 representing anger, sadness, normal, happy and surprised.

The paper is organized as follows. In 2, The multi-modal model based on text and image and related work are presented. Description of the data set and setting of the experimental model are provided in 3. In 4, the experimental results are discussed.

2. MODELS

In this section, we describe the models used in this paper: RNN, LSTM, ResNet, The Multi-modal Model Emotion Recognition (MMER) based on text and image.

2.1. Long Short Term Memory networks (LSTM)

Recurrent neural networks (RNN) have been employed to produce promising results on a variety of tasks including language model[12] and speech recognition[13]. An RNN maintains a memory based on history information, which enables the model to predict the current output conditioned on long distance features.

Figure.1 shows the RNN structure which has an input layer, hidden layer and output layer. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input.

Unlike feedforward neural networks, RNNs can use their internal state to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

In this paper, we apply Long Short-Term Memory[14] to emotion recognition. Long Short-Term Memory networks are the same as RNNs, except that the hidden layer updates are replaced by purpose-built memory cells. LSTM solves the problems of RNN gradient disappearance and gradient explosion through the gate mechanism, so that the network can perform multiple back propagation. As a result, they may be better at finding and exploiting long range dependencies in the data. Figure.2 illustrates a single LSTM memory cell[15].

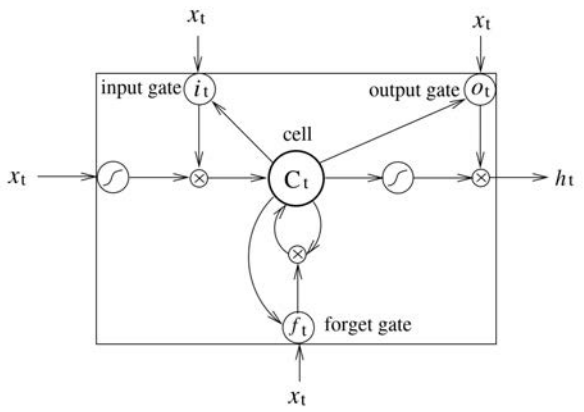

Fig. 2. A Long Short-Term Memory Cell.

LSTM memory cell is implemented as the following:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad . \; . \; . \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad . \; . \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad . \; . \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad . \; . \; . \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (5)$$

where σ is the logistic sigmoid function, i, f, o and c are the input gate, forget gate, output gate and cell vectors, all of which are the same size as the hidden vector h.The weight matrix subscripts have the meaning as the name suggests. For example, W_{ho} is the hidden-input gate matrix, W_{xo} is the input-output gate matrix etc. The weight matrices from the cell to gate vectors (e.g. W_{ci}) are diagonal,so element m in each gate vector only receives input from element m of the cell vector.

2.2. Residual Network (ResNet)

According to recent studies in the areas of image recognition, the depth of a network, i.e., the number of layers in CNN is crucial in its performance. However, when the depth is deeper than a certain level, degradation occurs such that the accuracy is saturated and then rapidly degrades. Experiments showed that such degradation is not caused by over-fitting, but a matter of optimization. ResNet resolves this degradation problem by introducing a residual learning framework[16].As shown in Figure.3, a block of plain CNN directly learns a target function $H(x)$.However, a ResNet block in Fig. 1 (b)has a different learning objective defined as $F(x) := H(x) - x$.It is called residual learning in the sense that the residual of x in $H(x)$ is learned. It implies that, it is easier to learn the residual than to learn whole $H(x)$, because ResNet learns complicated objective by detouring[17]. This concept also can be represented by a "shortcut connection", which performs identity mapping, and x is added to the output of the stacked layer.

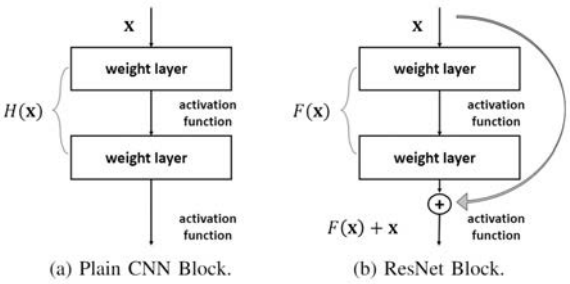

Fig. 3. Configuration of plain CNN block and ResNet block.

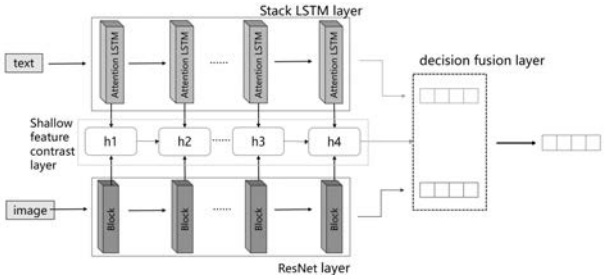

Fig. 4. The Multi-modal Emotion Recognition Based on Text and Image

2.3. The Multi-modal Emotion Recognition Based on Text and Image (MMER)

The text and image-based multimodal model is a two-branch neural network model that uses both image and text as input. The basic components of the model use ResNet model and stack LSTM model. In the data flow of the network model, the model extracts the shallow feature vectors of ResNet and LSTM in each unit block, and merges the two as the input of the shallow vector fusion layer of the next unit block. Through this form, the model can learn in advance the difference between images and text at the shallow feature level, and use this difference as a factor in the final emotional decision. Figure.4 illustrates the Multi-modal Emotion Recognition Based on Text and Image.

In addition, at the intersection of the two network branches, we set up a decision fusion layer. We input the output of the stack LSTM network, the output of the ResNet network and the output of the shallow feature fusion layer into the decision fusion layer for data fusion and final emotion recognition. Here we use three different fusion strategies, they are mean fusion, D-S evidence theory fusion and dynamic weighted fusion. In the mean fusion method, we add the positions corresponding to the three one-dimensional matrices obtained by the decision-making layer to calculate the average value.

D-S evidence theory[18] was first proposed by Dempster in 1967, and was further developed by his student Shafer in 1976. It was first applied to expert systems and has the ability to process uncertain information. As an uncertain reasoning method, the main characteristics of evidence theory is: it satisfies weaker conditions than Bayesian probability theory; it has the ability to directly express "uncertain" and "unknown". We first obtain the normalization constant K through the probability distribution matrix obtained by the decision-making layer, and then respectively obtain the mass functions of the 5 classification results, and finally concatenate these 5 values into a one-dimensional matrix as the output.

The dynamic weight fusion method is to add a hidden layer after the data fusion layer, and then use the neural network to train the weights w_1 and w_2 of the image text classification results

Because we want to perceive the semantic difference between the input image and text in advance, we use the cosine similarity method for the calculation of shallow feature vectors:

$$h_i = \frac{\sum_{j=1}^{n}(x_j \times y_j)}{\sqrt{\sum_{j=1}^{n}(x_j)^2} \times \sqrt{\sum_{j=1}^{n}(y_j)^2}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Where x, y are the shallow feature vectors of the picture and text respectively. We use cross entropy to calculate the loss function of the entire model, and dynamically balance the proportion of the shallow feature vector and the output of the two network branches in the final decision result through two hyperparameters λ_1 and λ_2:

$$F_{loss} = \frac{\lambda_1}{n}\sum_{i=1}^{n}(1-h_i) + \lambda_2 l, \lambda_1 < \lambda_2 \quad . \quad . \quad . \quad . \quad (7)$$

$$l = -\sum_{K=1}^{N} y^k \log \hat{y}^k + (1-y^k)\log(1-\hat{y}^k) \quad . \quad . \quad . \quad (8)$$

where λ_1 and λ_2 are hyperparameters, which respectively represent the proportion of the shallow feature vector and the output of the two-branch network in the decision result.

3. EXPERIMENTAL SETUP

In this section, we describe the dataset used in the experiment, the parameter setting of the model and the result evaluation standard.

3.1. Dataset

The Fer2013 facial expression dataset[19] is used for image emotion recognition. It consists of 35886 facial expression pictures. Each picture is composed of a grayscale image with a fixed size of 4848. There are 7 kinds of expressions, corresponding to the number labels 0-6, and the labels corresponding to the specific expression as follows: 0 anger; 1 disgust; 2 fear; 3 happy; 4 sad; 5 surprised; 6 normal, but we only select five of them, which are anger, sad, normal, happy and surprised.

The amazon reviews dataset[20] is used for text emotion recognition. It contains 600,000 book shopping reviews, each of which contains between 5 and 60 words. Each sentence also corresponds to a score label, and the

score ranges from 0 to 5, indicating that the shopping satisfaction is from low to high.

We selected 50,000 data sets from the above two data sets, with 10,000 data sets for each label. Then the image labels angry, sad, normal, happy, surprised are corresponding to the text labels 0,1,2,3,4,5, so that each row of the final data set consists of three columns, namely image, text and corresponding tags, there are 5 types of tags, 0 to 5 indicate the change of emotion from negative to positive.

3.2. Model Training and Parameters

For the word embedding method in the input module and aspect embedding module, we pretrained a GloVe[21] model. The training corpora was Leipzig Corpora Collection[22] and the dimension of embedding vector was set to 300. To avoid out-of-vocabulary (OOV) problems, we sampled from a uniform distribution $U(0.1, 0.1)$ for the words that were not recorded in the dictionary. The Chinese dataset was preprocessed by word segment with Stanford Word Segmenter.

For the CNN in the attention module, 100 filters were used to reduce the possibility of missing important signals. The CNN network contained one convolution layer, in which the filter was of $R^{3\times 300}$ with no padding and the stride was 1.

We trained the model through mini-batch gradient descend, and the mini-batch was 32. L2 regularization factor was added to the loss function and the weight for L2 factor was 3. The dropout rate was set to 0.5 and the learning rate was set to 0.001.

3.3. Evaluation Criterion

This paper uses precision rate, recall rate and Macro-F1[23] value to evaluate the accuracy of emotion recognition.

We know that in the binary classification task, the accuracy rate P represents the proportion of correct words in all the extracted words, the recall rate R represents the proportion of correct words in all ground truth, and the $F1$ value is the harmonic average of the precision rate and the recall rate.

The confusion matrix[24] is shown in Table 1. We can calculate the value of precision rate, recall rate and F1 value according to the confusion matrix.

Table 1. Confusion matrix for binary classification task

	Predicated Positives	**Predicated Negatives**
Actual Positives	True Postives (TP)	False Negatives (FN)
Actual Negatives	False Positives (FP)	True Negatives (TN)

$$P = \frac{TP}{TP+FP} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$R = \frac{TP}{TP+FN} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

$$F1 = \frac{2\times P\times R}{P+R} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

Macro-F1 first calculates the total number of TP, FP and FN, and then calculates PRF. That is, the positions corresponding to TP, FP, TN, and FN of multiple confusion matrices are averaged first, and then calculated according to the PRF value formula and inverse. The formula is as follows:

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

$$Macro\_F = \frac{1}{n}\sum_{i=1}^{n} F_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

Macro-F1 can treat each category equally, and its value will be affected by the rare category.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We performed three sets of experiments. The first set of experiments was the comparison of the effects of the MMER model with the four models of BI-LSTM, BI-LSTM+Attention, BI-LSTM+Pooling, Transformer, and calculated the Macro-F1 scores of five emotional labels. The second set of experiments is the comparison of the effects of the MMER model and the four models of VGG16, ResNet, and AlexNet. The Macro-F1 scores of five emotional labels are calculated respectively. In the third set of experiments, we set up three data fusion methods at the fusion decision-making layer, namely weighted fusion, D-S evidence theory fusion and dynamic weight fusion. We compared the accuracy of sentiment label classification under the conditions of three different data fusion methods.

4.1. Comparison With Various LSTM Networks

We compare the MMER model with four LSTM network models such as LSTM, LSTM+Attention[25], LSTM+Pooling, Transformer[26]. LSTM network has seven LSTM cells and 280 hidden nodes. LSTM+Pooling network model adds a max pooling layer to the original LSTM. The maximum pooling layer retains the main features while reducing the amount of parameters and calculations, preventing over-fitting and improving the generalization ability of the model.LSTM+Attention adds attention mechanism to the original LSTM. The attention mechanism can select information that is more critical to the current task goal from a large number of information. The structure of Transformer is more complex, using multi-head attention and position-wise feed-forward

Table 2. The average Macro-F1 of different network models for text emotion recognition

Model	Average	Angry	Sad	Normal	Happy	Surprised
LSTM	69.1	87.2	72.2	60.2	85.1	40.8
LSTM+Pooling	70.66	87.5	73.4	64.4	87	41
LSTM+Attention	72.36	90.2	75.5	65.7	87.3	43.1
Transformer	69.76	85	73	63.1	85.4	42.3
MMER	73.54	90.7	73.9	66.9	89.1	47.1

Table 3. The average Macro-F1 of different network models for image emotion recognition

Model	Average	Angry	Sad	Normal	Happy	Surprised
VGG16	58.06	72.8	58	54.2	77.4	27.9
ResNet	66.7	85.2	67.2	60.1	86.9	34.1
AlexNet	65.78	83	66.1	62.7	84	33.1
MMER	73.54	90.7	73.9	66.9	89.1	47.1

networks. Table 2 shows the experimental results comparing MMER with various LSTM models.From the table, we can see that compared with other LSTM models, the sentiment classification results of MMER have been significantly improved, especially under the normal and surprised labels.

4.2. Comparison With VGG16, ResNet, AlexNet Networks

We compare the proposed method MMER with VGG16, ResNet, AlexNet network models. Table 3 shows that compared with the other three network models, MMER has a significant improvement in the effect of image emotion recognition. Especially for surprised and sad images, the improvement in recognition accuracy is the most obvious. The Macro-F1 values of sad and surprised increased from 58 and 27.9 to 73.9 and 47.1 respectively.

4.3. Comparison of Three Data Fusion Methods

We compared three fusion algorithms, mean fusion, D-S evidence theory fusion, dynamic weighted fusion, and their data fusion effects at the emotional decision-making level.In the mean fusion method, we add the positions corresponding to the three one-dimensional matrices obtained by the decision-making layer to calculate the average value. In the D-S evidence theory, We first obtain the normalization constant K through the probability distribution matrix obtained by the decision-making layer, and then respectively obtain the mass functions of the 5 classification results, and finally concatenate these 5 values into a one-dimensional matrix as the output. The dynamic weight fusion method is to add a hidden layer after the data fusion layer, and then use the neural network to train the weights w_1 and w_2 of the image text classification results. Table 4 shows the accuracy of emotion recognition under the three data fusion algorithms.

Table 4. The Accuracy of three data fusion methods for emotional decision

Decision fusion method	Accuracy
Mean fusion	78.4%
D-S evidence theory fusion	80.0%
Dynamic weighted fusion	85.8%

5. CONCLUSION

The Multi-modal emotion recognition based on text and image (MMER) is proposed for emotion recognition. The basic idea is to fusion of multi-modal data and solve the problem of inaccurate emotion recognition and poor model robustness of a single modality. Through experiments, we demonstrated that Macro-F1 score for the multi-modal model based on text and image is 73.54, achieving 6.4% and 11.8% improvement compared with the text emotion recognition model LSTM and the image emotion recognition model ResNet.

Three sets of experiments, MMER and text emotion recognition model comparison, MMER and image emotion recognition model comparison and fusion algorithm comparison, prove that not only the accuracy of MMER is higher than the single-modal emotion recognition model of image or text, but also MMER can effectively deal with scenes where some data of text or image is lost.In the future, we consider the use of capsule models and fusion functions to enable the model to further fusion features of different levels of data and increase the flexibility of the model.

Acknowledgements

This work is supported by the State Recruitment Program of Global Experts under Grant WQ20141100198.

References:

[1] Wang S H, Phillips P, Dong Z C, et al. Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm[J]. Neurocomputing, 2018, 272: 668-676.

[2] Zeng N, Zhang H, Song B, et al. Facial expression recognition via learning deep sparse autoencoders[J]. Neurocomputing, 2018, 273: 643-649.

[3] Yan Y, Zhang Z, Chen S, et al. Low-resolution facial expression recognition: A filter learning perspective[J]. Signal Processing, 2020, 169: 107370.

[4] Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition based on cross-connection LeNet-5 network[J]. Acta Automatica Sinica.

[5] Yao Naiming, Guo Qingpei, Qiao Fengchun, et al. Robust facial expression recognition based on generative confrontation network[J]. Acta Automatica Sinica, 2018, 44(5): 865-877.

[6] Tan Xiaohui, Li Zhaowei, Fan Yachun. Facial expression recognition method based on multi-scale detail enhancement[J]. Journal of Electronics & Information Technology, 2019, 41: 11.

[7] Tang D, Qin B, Liu T. Aspect Level Sentiment Classification with Deep Memory Network[J]. 2016.

[8] Wang Y, Huang M, Zhu X, and Zhao L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), 2016

[9] Noroozi F, Marjanovic M, Njegus A, et al. Fusion of classifier predictions for audio-visual emotion recognition[C]. 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 61-66.

[10] Xiao F. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy[J]. Information Fusion, 2019, 46: 23-32.

[11] Yoshida S, Hirota K. Concepts of D, T, SR Fuzzy Flip Flop and Their Circuit Design Using FPGA[J]. International Journal of Japan Society for Fuzzy Theory and Systems, 2000, 12(1): 160-168

[12] Li S , Li W , Cook C , et al. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN[J]. 2018.

[13] Battenberg E , Chen J , Child R , et al. Exploring neural transducers for end-to-end speech recognition[C]// 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2018.

[14] Ghaeini R , Hasan S A , Datla V , et al. DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference[J]. 2018.

[15] Gers F A . Learning to forget: continual prediction with LSTM[C]// 9th International Conference on Artificial Neural Networks: ICANN '99. IET, 1999.

[16] Caihua X , Yanli G , Hua L , et al. An improved Faster R-CNN hand gesture recognition algorithm based on ResNet-50[J]. Computer Era, 2019.

[17] Chen C , Qi F . Single Image Super-Resolution Using Deep CNN with Dense Skip Connections and Inception-ResNet[C]// 2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE Computer Society, 2018.

[18] Wang H , Guo L , Dou Z , et al. A New Method of Cognitive Signal Recognition Based on Hybrid Information Entropy and D-S Evidence Theory[J]. Mobile Networks and Applications, 2018.

[19] Barsoum , Emad and Zhang , Cha and Canton Ferrer , Cristian and Zhang , Zhengyou. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. ACM International Conference on Multimodal Interaction (ICMI), 2016.

[20] R. He, J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016.

[21] Pennington J , Socher R , Manning C . Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014.

[22] Rosanne, M, Leipzig. Drugs and Falls in Older People: A Systematic Review and Meta-analysis: II. Cardiac and Analgesic Drugs[J]. Journal of the American Geriatrics Society, 2015.

[23] Wang D , Zhang H , Liu R , et al. t-Test feature selection approach based on term frequency for text categorization[J]. Pattern Recognition Letters, 2014, 45(aug.1):1-10.

[24] Xu J , Zhang Y , Miao D . Three-way Confusion Matrix for Classification: A Measure Driven View[J]. Information ences, 2019, 507.

[25] Yangsen Z , Jia Z , Yuru J , et al. A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model[J]. Chinese Journal of Electronics, 2019, 28(001):120-126.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al.Attention is all you need. Advances in neural information processing systems, 2017.

[27] Liu Z T, Wu M, Hirota K, et al. Concept of Fuzzy Atmosfield for Representing Communication Atmosphere and Its Application to Humans-robots Interaction[J]. International Journal of Enterprise Information Systems, 2013, 17(1): 3-17.

[28] Liu Z T, Wu M, Hirota K, et al. Communication Atmosphere in Humans and Robots Interaction Based on The Concept of Fuzzy Atmosfield Generated by Emotional States of Humans and Robots[J]. Journal of Automation Mobile Robotics & Intelligent Systems, 2013, 7.

[29] Poria S, Cambria E, Howard N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content[J]. Neurocomputing, 2016, 174: 50-59.

[30] Gosztolya G. Posterior-thresholding feature extraction for paralinguistic speech classification[J]. Knowledge-Based Systems, 2019, 186: 104943.

[31] Zhang L, Mistry K, Neoh S C, et al. Intelligent facial emotion recognition using moth-firefly optimization[J]. Knowledge-Based Systems, 2016, 111: 248-267.

[32] Cui Z , Maojie Z . A text emotion classification method based on CNN and bidirectional LSTM fusion[J]. Computer Era, 2019.

[33] Gen-Sheng W , Xue-Jian H , Lu M . GRU Neural Network Text Emotion Classification Model Based on Multi-feature Fusion[J]. Journal of Chinese Computer Systems, 2019.

[34] Atkinson A P , Smithson H E . The impact on emotion classification performance and gaze behavior of foveal versus extrafoveal processing of facial features[J]. Journal of Experimental Psychology Human Perception & Performance, 2020, 46(3):292-312.