Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 3143748, 8 pages
https://doi.org/10.1155/2022/3143748

Research Article

Research on Animated GIFs Emotion Recognition Based on ResNet-ConvGRU

Qian Zhang[1], **Ren Qing-Dao-Er-Ji**[1], **and Na Li**[2]

[1]*School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China*
[2]*School of Science, Inner Mongolia University of Technology, Hohhot 010051, China*

Correspondence should be addressed to Ren Qing-Dao-Er-Ji; renqingln@sina.com

Received 23 April 2022; Revised 13 August 2022; Accepted 13 September 2022; Published 30 September 2022

Academic Editor: Francesco Lolli

Animated Graphics Interchange Format (GIF) images have become an important part of network information interaction, and are one of the main characteristics of analyzing social media emotions. At present, most of the research on GIF affection recognition fails to make full use of spatial-temporal characteristics of GIF images, which limits the performance of model recognition to a certain extent. A GIF emotion recognition algorithm based on ResNet-ConvGRU is proposed in this paper. First, GIF data is preprocessed, converting its image sequences to static image format for saving. Then, the spatial features of images and the temporal features of static image sequences are extracted with ResNet and ConvGRU networks, respectively. At last, the animated GIFs data features are synthesized and the seven emotional intensities of GIF data are calculated. The GIFGIF dataset is used to verify the experiment. From the experimental results, the proposed animated GIFs emotion recognition model based on ResNet-ConvGRU, compared with the classical emotion recognition algorithms such as VGGNet-ConvGRU, ResNet3D, CNN-LSTM, and C3D, has a stronger feature extraction ability, and sentiment classification performance. This method provides a finer-grained analysis for the study of public opinion trends and a new idea for affection recognition of GIF data in social media.

1. Introduction

Emotion recognition is one of the important research directions in affective computing, and its primary task is to identify people's emotional states through their facial expressions, behaviors, or physiological signals. In recent years, with the rapid development of social media, such as Weibo, forums, and TikTok, people are tending to express their personal feelings through various images and video data on social platforms. It is of great applicable value in box office earnings forecast [1], political elections [2], commodity recommendations [3], medical treatment [4], and education assistance [5] in the ways of mining valuable information from massive data and classifying the expressed affections and emotions.

Among them, animated Graphics Interchange Format (GIF) image as a significant data form of emotional expression, has been widely used on various social platforms. GIF is a shorter sequence of image data that is lighter than video and is sometimes referred to as a GIF short video. Compared to the emotion recognition of electroencephalogram (EEG), the GIF data has the advantages of convenient data collection and intuitive affection expression; compared to the emotion recognition of image, it can express more lively and convey much more information. The GIF data of mood changes is shown in Figure 1. Undeniably, GIF data has its unique advantages. Therefore, emotion recognition of GIF animated data is of great academic value and broad practical significance [6].

The early research on image emotion recognition mainly used the method based on manual features, which was composed of traditional image processing, pattern recognition, and classifiers, including the targeted determination such as histograms of oriented gradients (HOG), K-nearest neighbor algorithm (KNN), support vector machine (SVM), and other classification methods [7]. With the wide application of deep learning in affective computing, deep learning is applied to emotion recognition tasks by many experts and

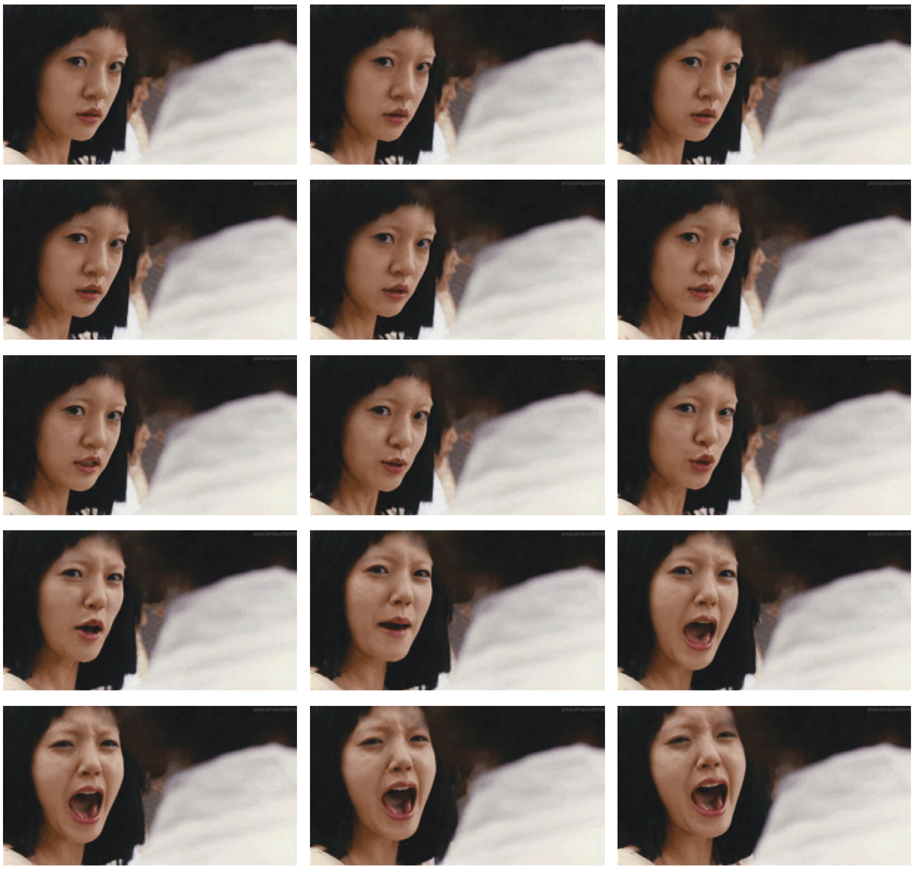

Figure 1: GIF graphs of emotional changes (example).

scholars. The emotion recognition method based on deep learning mainly extracts data features through convolution operation in the neural networks, including architectures such as GoogleNet, ResNet, VGGNet, and AlexNet [8]. This method has become the mainstream method of image emotion recognition because of its strong generalization ability and high accuracy of emotion recognition.

There are fewer related research studies at present because GIF data is widely used later than pictures, and the data features are more complex, the traditional convolution cannot make full use of the unique attributes of GIF to extract features. It has inspired us to build a model based on ResNet-ConvGRU to complete the emotion recognition task of animated GIFs referring to the algorithm idea of combining convolution feature extraction in pictures with spatial-temporal characteristics extraction in the video. The advantages of the ResNet-ConvGRU model compared with the conventional CNN-GRU network structure are that the residual network has stronger feature extraction capability and the ConvGRU network has stronger sequence prediction and generalization capability. The combined framework of the two is more conducive to learning animated GIF spatio-temporal features, which in turn improves the efficiency of emotion recognition. In this paper, we introduce the theory of animated GIFs emotion recognition and research progress from the aspects of GIF data emotion recognition classification system, feature extraction, network construction, attribute extraction, and classification recognition and finally summarize and prospect the existing problems and future challenges.

2. Related Work

2.1. Emotional Classification System. Affections and emotions are people's attitude experience and corresponding behavior reactions to objective things. Emotion classification is also called refined affection classification, which is because the early affection classification is usually divided into two categories of positive and negative, or three categories of positive, negative, and neutral, failing to meet people's needs for the simulation of complex emotion changes. Compared with affection classification, the emotion classification can reflect people's real psychological state in a certain period to a certain extent, hence, the study of emotion classification emerges from an opportunity. With the advancement of society, people's cognition of emotion is constantly updated. There is no unified standard for the classification of emotion in the academic community, but it can be divided into discrete emotion and dimensional emotion according to the different classification methods. The researchers of discrete emotions believed that people can have specific emotions that vary according to mood changes, but the basic emotions are rare and can be defined in a variety of ways. In 1992, American psychologist Ekman and his colleagues proposed a basic system of six emotions, including happiness, sadness, anger, fear, disgust, and surprise, based on the in-depth study of facial expressions [9]. Then, Robert Plutchik proposed a multidimensional emotional model in 2001, which contains eight basic two-way emotions, including happiness, trust, fear, surprise, sadness, disgust, anger, and expectation [10]. Izard [11] put forward the theory that human beings had ten basic emotions including anger, contempt, happiness, shame, disgust, interest, surprise, sadness, fear, and guilt employing factor analysis. Researchers of dimensional emotion theory tried to identify two emotions or define emotions in more dimensions. They believed that it was more reasonable to analyze emotional states in the emotional space. Russell and Barrett [12] adopted a two-dimensional valence-arousal model aiming at emotion recognition results. Wherein the valence corresponds to the pleasure value of the emotion felt by the individual, and the arousal corresponds to the emotional intensity.

We selected and sorted out seven emotion categories including happiness, like, surprise, sadness, anger, disgust, and fear based on the business requirements of most image emotion classification tasks and constructed the emotion corpus.

2.2. GIF Data Emotion Recognition Method. Jou et al. [13] firstly predicted the emotion in GIF and compared the influence of different types of features in GIF, including color histogram, aesthetics, intermediate semantic features inspired by affection modeling, and facial features on emotion recognition. The experimental results showed that the facial expression features had the highest accuracy in predicting 17 emotions compared with other features. Inspired by video analysis, Chen and Picard [14] employed a three-dimensional convolution neural network (CNN) to extract spatial-temporal characteristics from GIFs for emotion recognition, achieving good results. Yang et al. [15] proposed a key point attend visual attention network (KAVAN), and improved the traditional LSTM layer into an (HSLSTM) module to further improve the affection recognition performance of GIF. Liu et al. [16] adopted the network model of a combination with C3D and ConvLSTM for affection analysis of GIF short videos, achieving good experimental results. Lin [17] proposed an affection analysis method (RFCM) based on a 3D residual network (Res Net3D) and improved convolution long and short time memory network (FConvLSTM). As a result, it could effectively improve the effect of affection classification. Ma et al. [18] proposed a new method for animated GIFs content classification based on keyframe attention and entropy loss, reducing the interference of redundant frames to a certain extent, thus improving the classification efficiency of the model. The above-given algorithms had achieved good results in some aspects, but the GIF features were not fully utilized to extract its spatial features affecting the final GIF recognition efficiency. Although the transfer learning of the 3D convolution method in video recognition can be simultaneous to the extraction of the spatial-temporal characteristics of the GIF data, the 3D convolution can not obtain a longer time dependence. In addition, there are some problems in GIF data such as the large difference in image sequence length distribution and the inconsistent image size, bringing challenges to GIF emotion recognition. An animated GIFs emotion recognition algorithm based on ResNet-ConvGRU is proposed in this paper to solve the above-given problems. The efficiency of emotion recognition of GIF data is further improved by learning the spatial and temporal characteristics of GIF data, respectively.

3. Model Introduction

An animated GIFs emotion recognition method based on ResNet-ConvGRU is proposed in this paper. The model structure is shown in Figure 2 below. It mainly includes four modules: the input layer, the ResNet layer, the ConvGRU layer, and the emotion classification layer.

(1) The input layer preprocesses the source data including converting the GIF data to a static discrete image, i.e., the steps of gif converting to .png and adjusting the size of the image.

(2) The ResNet layer is used to extract the spatial features of static images.

(3) The ConvGRU layer is used to extract the temporal features of the animated GIFs image sequence. It takes the image spatial features extracted from the ResNet network layer and stitches them by column and inputs them to the ConvGRU layer to further extract the image sequence features.

(4) The emotion classification layer is used for classifying the total characteristics of the data with the Softmax function and outputting the result.

3.1. Inputs Layer. The input layer mainly preprocesses the input animated GIFs data, making it more standard and convenient for model training. First, the GIF short video is converted into an image format for saving, i.e., converting the suffix name of the data .gif to .png. The naming of static images is arranged according to time series convenient for the model to extract its temporal features. Second, all images are converted to RGB format. It is because the data in the data set is the gray image and the data decoding mode is gray image mode, the RGB image format needs to be further converted into three channels. Third, the image size is adjusted to be 224 × 224. Finally, canonical image data are input into the model.

3.2. Spatial Feature Learning. ResNet [19] is usually used to explore the local features in an image. The spatial feature of the image is extracted from the animated GIF with ResNet in this paper. The spatial feature learning of GIF images is adopted by the ResNet18 network because the deep network model may lead to insufficient training and degradation due to the small data set in this paper. Figure 3 shows the ResNet

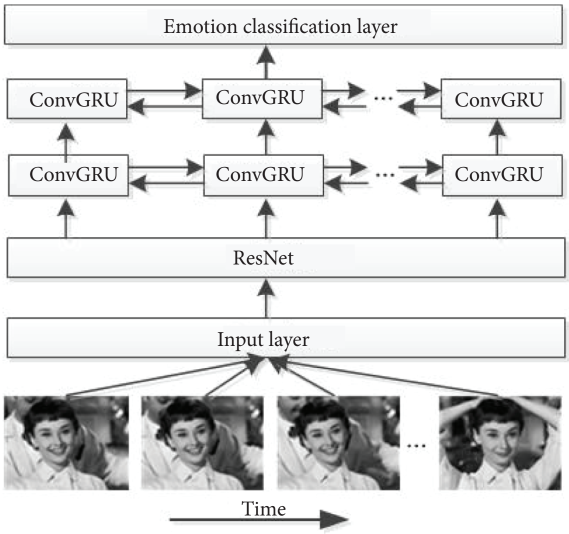

Figure 2: Structure diagram of ResNet-ConvGRU model.

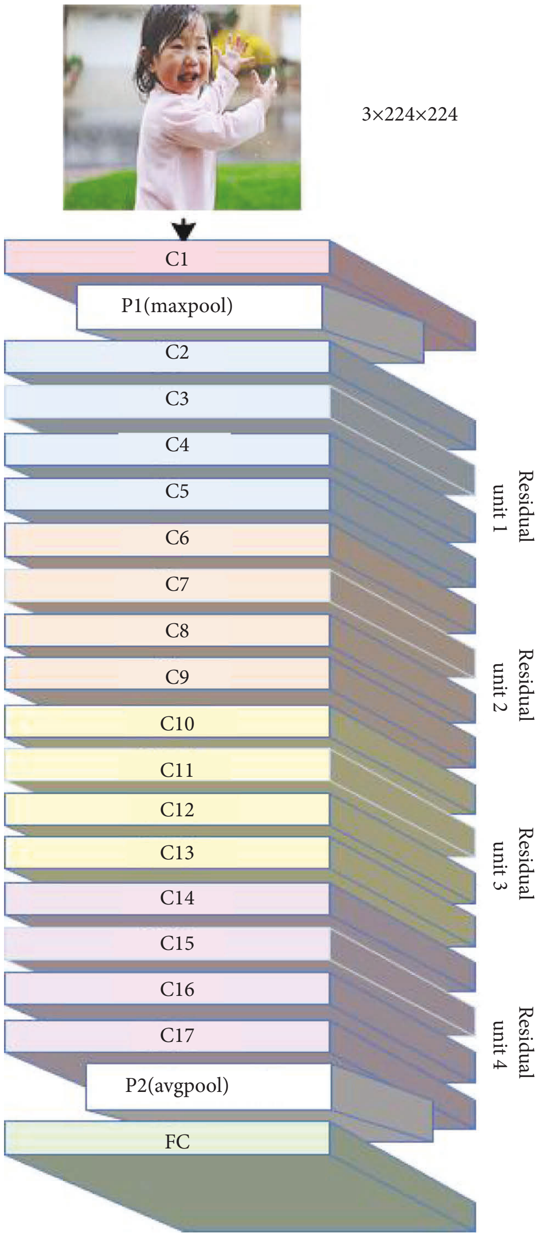

Figure 3: Structure diagram of ResNet network.

image spatial features extraction network structure diagram, including 17 convolution layers, 1 maximum pooling layer, 1 average pooling layer, and 1 full connection layer. The ResNet network layer in this paper is convolved by a small convolution kernel of 3×3 to extract the deep spatial features of the image.

Each residual unit in Figure 3 is composed of two residual blocks. Where the residual block contains identity mapping and ReLU connection. The function of identity mapping is to establish a direct correlation channel between input and output. A function expression of ReLU connection:

$$F(x) = W_1 \sigma(W_2 x), \tag{1}$$

where σ is the nonlinear activation function ReLu, W_1 and W_2 represent the model parameters. When the numbers of channels of the input characteristic pattern and output characteristic pattern of the residual block are equal, the output of the residual block is

$$y = F(x, W_i) + x. \tag{2}$$

3.3. Temporal Feature Learning. The traditional recurrent neural network (RNN) cannot process long sequence information well because of the problems of gradient disappearance and gradient explosion in training that is easier to occur. The Gated recurrent unit (GRU) is obtained during the improvement of the traditional recurrent neural network. As a variant of the Long Short-Term Memory(LSTM) networks, GRU only retains the reset gate and the update gate with the advantages of fewer parameters, a simple model, and fast convergence speed. Its unit structure is shown in Figure 4.

The ConvGRU [20, 21] network is used to extract the temporal features of the animated GIFs image sequence. This is because GIF images, unlike static images, have time-series features that change and can dynamically represent changes in human emotion. That is also the reason that makes it very difficult to extract the emotion features of GIF images. In this paper, the image features extracted by the ResNet network are stitched together and fed into the ConvGRU network, so as to learn the changes in the emotion features of GIF images in the temporal dimension. ConvGRU is an improvement based on retaining the advantages of the GRU structure. The activate function σ and full connection operation in tanh in its structure are replaced by a convolution operation. By doing so, the RNN's ability to explore the inherent dependencies of sequence data can be inherited, and the underlying temporal characteristics of the data can be captured. The basic calculation formula of ConvGRU is as follows:

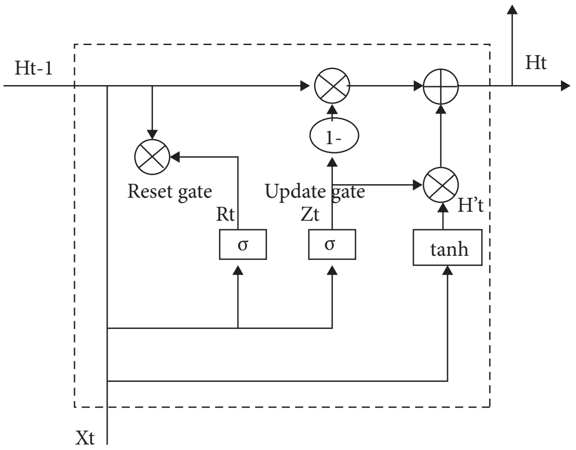

Figure 4: GRU structure.

$$\begin{aligned} Z_t &= \sigma\left(W_{xz} * X_t + W_{hz} * H_{t-1}\right), \\ R_t &= \sigma\left(W_{xr} * X_t + W_{hr} * H_{t-1}\right), \\ H_t' &= f\left(W_{xh} * X_t + R_t \circ \left(W_{hh} * H_{t-1}\right)\right), \\ H_t &= \left(1 - Z_t\right) \circ H_t' + Z_t \circ H_{t-1}, \end{aligned} \tag{3}$$

where Z_t, $R_{t,}$ and H_t stand for update gate, reset gate, and memory state, respectively. The update gate defines the amount of information that the emotional features preserved by the memory state are finally preserved to the ConvGRU network. The reset gate determines how the image space features output by the ResNet network are combined with the sentiment features stored in the memory state of the ConvGRU network. The memory state is used to preserve the image features learned by the ResNet network. H_t' is the new image features obtained by aggregating the image features outputted by the ResNet network with the output of the previous hidden layer. $*$ represents convolution operation, $\circ$ stands for Hadamard product [22], f stands for activation function. W_{xz}, W_{hz}, W_{xr}, W_{hr}, W_{xh}, and W_{hh} represent the training parameter matrix. In this paper, the output result of the ResNet network is used as the input of the ConvGRU layer through the superposition of the multilayer ConvGRU, and the time sequence relationship is established. The convolution kernel size of each layer of ConvGRU is 3×3.

3.4. Emotional Classification System. The emotion classification layer is mainly to classify the final emotion characteristics and output the results. The loss function is cross-entropy in this paper, and the Softmax function is used as the classification function in the full connection layer to output the prediction results.

$$\begin{aligned} L &= \frac{1}{N} \sum_i L_i \\ &= -\frac{1}{N} \sum_i \sum_{c=1}^{M} y_{ic} \log\left(p_{ic}\right), \end{aligned} \tag{4}$$

where M is the number of categories, here in this paper it presents a seven classified emotion task, thus $M = 7$. y_{ic} is the symbolic function, the value is 0 or 1. When the real class of the sample i is equal to c, it is 1, otherwise, it is 0. p_{ic} is the prediction probability that the observed sample i belongs to the class c.

$$y_i = \operatorname{softmax}\left(W_{\text{soft}}\left(\tanh\left(W_l F_i^* + b_l\right)\right) + b_{\text{soft}}\right), \tag{5}$$

where W_l, b_l stand for the weight and offset of the full connection layer, W_{soft} and b_{soft} stand for the weight and offset of the classification layer softmax, y_i represents the emotion classification result of the final data.

4. Experiment

The experimental codes are implemented on GPU-Nvidia Tesla P100, training with Python3.7 and Torch1.10.0.

4.1. Experimental Data. The GIFGIF data set was used in the experiment in this paper. This data of 17 emotion classifications were provided by MIT multimedia labs, including pleasure, disgust, happiness, excitement, embarrassment, surprise, fear, satisfaction, pride, anger, and guilt, a total of 6170 animated GIFs. The animated GIFs content covered many fields such as movies, TV programs, cartoons, and sports. The data label was pair-compared and voted from users' emotions on more than 2.7 million animated GIFs content.

We selected 7 out of 17 emotion categories including happiness, like (satisfaction), surprise, sadness, anger, disgust, and fear for a better experiment. The first 300 animated GIFs in each category were used, and a total of 2100 GIF data were used in the experiment. In the experiment, 80% of each category of data in the data set was selected as the training set, and the remaining 20% was selected as the test set.

4.2. Experimental Parameters and Evaluation Indicators. The experimental parameters are set as shown in Table 1.

The evaluation indexes of the model were accuracy (Acc), precision (P), Recall (R), and F1 Score (F1) to measure the experimental effect of the model. Accuracy represented the proportion of the number determined to be correct to the total number:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

where TP was the number of true positive classes, TN was the number of true negative classes, FP was the number of misjudging the negative classes as positive classes, and FN was the number of misjudging the positive classes as negative classes. The following formula was the same. Precision was the proportion of the true positive class to the predicted positive class.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

Table 1: Experimental parameters.

Parameters	Value	Parameters	Value
kernel_size	(3,3)	Loss function	Cross entropy
Learning rate	0.0001	Epoch	100
weight_decay	1e-4	Batch size	30
The function optimized	Adam	Number of categories	7

The recall rate refers to the proportion of positive class judged to the true positive class.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{8}$$

The F1 score was the harmonic average of accuracy and recall.

$$F1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{9}$$

As shown above, the higher the accuracy, precision, recall, and *F*1 score, the better the performance of the emotion recognition model.

4.3. Experimental Analysis and Results. Animated GIF is a kind of image sequence data. Its image sequence number is unevenly distributed due to the different sources of the animated GIFs in the data set. As shown by the relevant experiments, the problems of uneven distribution of the number of image sequences and too long sequences all have different degrees of influence on the emotional recognition of animated GIFs. We made the following statistics on the number frequency distribution of all animated GIF image sequences in the experimental data to explore the distribution characteristics of animated GIF image sequences in the data set:

Figure 5 showed that most of the animated GIF sequences number used in the experiment were within 100, mostly, between 0 and 50. The shortest animated GIF image sequence was 2, and the maximum could be up to 303 in this paper according to the experimental data statistics. It could be seen that the number of animated GIF image sequences in this data set was relatively concentrated despite the large differences in the number of image sequences.

To alleviate the problem above that the number of animated GIF image sequences had a greater difference which influenced the emotion recognition efficiency, in this paper, after the animated GIF images were converted to static images for saving every 4 images were input into the model as an image sequence according to the time sequence and their spatial-temporal characteristics were analyzed, respectively. Data of image sequences that were less than 4 images were supplemented randomly, while longer image sequences were split, and input in multiple batches. Various animated GIF emotion recognition methods were compared to verify the proposed effectiveness of the animated GIF emotion recognition method based on ResNet-ConvGRU.

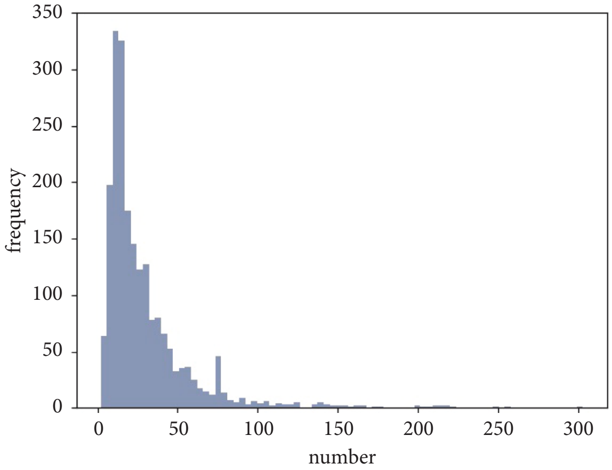

Figure 5: Number frequency distribution diagram of the animated GIF image sequence.

Table 2 showed the comparison of the evaluation indexes of GIF data of each emotion category in the emotion classification model based on ResNet-ConvGRU. It could be seen that the category of anger has a higher score on precision, indicating that the category had higher prediction accuracy in the positive sample results. It might be because the facial expressions and body movements of the animated GIF images of the anger emotion category were more obvious and easier to be distinguished than other emotions. While like had a higher score on recall, indicating that the probability of the actual emotion category being predicted was higher. It might be because the GIF image of like had a higher correlation between features, such as a tendency to contain obvious love and easier to be recognized. There were still gaps in the F1 score among the seven emotional categories, which indicated that there was a certain difference in the difficulty of recognizing various emotional features. Of course, it might be related to the animated GIF images themselves in the model training.

In order to verify the advantages of the proposed model backbone network ResNet feature extraction, the proposed model is experimentally compared with a similar deep network VGGNet in this paper. It is found that the ResNet network used in this paper contains 17 convolutional layers and 1 fully connected layer, while the VGG19 network contains 16 convolutional layers and 3 fully connected layers Table 3.

As can be seen from the above experimental results, it is clear that the precision, recall, and F1-score of the ResNet-ConvGRU model proposed in this paper are higher than those of the VGGNet-ConvGRU model for animated GIF emotion recognition by 0.8%, 4.09%, and 3.57%, respectively. The results show that the network structure with residual connections is more capable of feature extraction. This is because ordinary convolutional networks experience gradient disappearance as the number of layers deepens, which results in a loss of information about the learned image features. In contrast, the ResNet network uses jump connections internally, which to some extent avoids the

Table 2: Animated GIF emotion recognition confusion matrix.

Label	Precision	Recall	$F1$ score
Anger	0.7976	0.6331	0.7059
Disgust	0.5959	0.6438	0.6189
Fear	0.6322	0.6577	0.6447
Happiness	0.6384	0.6215	0.6299
Like	0.6912	0.8079	0.7450
Sadness	0.7121	0.6403	0.6743
Surprise	0.6592	0.6940	0.6762

Table 3: Comparison of ResNet-ConvGRU and VGGNet-ConvGRU.

Model	Precision	Recall	$F1$-score
ResNet-ConvGRU	0.6752	0.6712	0.6707
VGGNet-ConvGRU	0.6672	0.6303	0.6350

phenomenon of information loss as the network deepens. Therefore, the use of the backbone network ResNet to extract animated GIF image features has obvious advantages in terms of structure.

In addition, it was explored that the length of the input image sequence has an impact on the accuracy of the model. Specifically, the models were trained with 2, 4, 6, and 8 images as an input sequence. The experiments showed that the two-image sequence was the least effective, with a model accuracy of 61%, the 4- and 6-image sequences were more effective, with an accuracy of 67% and 66.56%, respectively, while the 8-image sequence had a reduced model training effect, with an accuracy of 65.21%. This may be related to the dataset on the one hand, and the higher correlation of consecutive 4–6 images on the other hand, which is more conducive to the model to extract the sentiment features of short image sequences.

The proposed model and the classical video analysis model were compared on the emotion recognition task to further verify the effectiveness of the proposed model.

There were 4 classic video analysis models:

(1) VGGNet [23]. It is a deep network structure consisting of 3×3 small convolutional kernels, commonly used in VGG16 and VGG19. It can extract image features and is commonly used in areas related to image recognition.

(2) ResNet3D [24, 25]. 3D convolution kernel was used to extract the spatial-temporal characteristics, and the network structure was deeper (18 layers and 34 layers) than the C3D model. This model could be used in video analysis and motion recognition.

(3) CNN-LSTM model [26]. CNN model was used to learn visual features, the output of CNN was used as the input of the LSTM module to learn, data sequence features, and the parameter weights of the two were shared across time. This model could be used in the image sequence and video recognition.

(4) C3D model [27, 28]. The network constructed with 3D convolution and 3D pooling could be used to extract the spatial-temporal characteristics of video data for video recognition and other fields.

Table 4: Comparison of emotion recognition accuracy of different methods.

Model	Accuracy
ResNet-ConvGRU	0.6700
VGGNet-ConvGRU	0.6349
ResNet3D	0.6450
CNN-LSTM	0.6427
C3D	0.6248

It could be seen from Table 4 that the classification accuracy of the ResNet-ConvGRU model was improved by 2.5 %–4.52% compared to the classic emotion recognition model in the animated GIFs emotion recognition task. The experimental results show that the feature extraction capability of the residual network is better than that of the ordinary convolutional network, and the effects of different residual network models are also different, the accuracy of the ResNet-ConvGRU model is improved by 2.5% compared with the ResNet3D model. In general, the proposed animated GIFs affection recognition model based on ResNet-ConvGRU could improve the performance of feeling classification.

5. Conclusions

An animated GIFs emotion recognition method based on ResNet-ConvGRU is proposed in this paper. The model can improve further the training efficiency by converting GIF short videos into static image sequences and extracting the spatial and temporal features of image sequences using ResNet and ConvGRU networks, respectively. In the experiment, we first showed the frequency distribution of the number of animated GIF images in the data set. Then, the recognition efficiency of various emotion categories of the proposed method was verified. Finally, different emotion recognition models were compared by experiments to show that the proposed model has higher accuracy than the experimental results. The model was validated on only one data set in this paper. The effective performance of emotion recognition on different data sets and the robustness and universality of the model still need further research.

Data Availability

The dataset was used to support the findings of this study Stored on GIPHY https://giphy.com/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant nos. 61966027, 61966028, and 62141603), Inner Mongolia Natural Science Foundation (Grant nos. 2020MS07006 and 2022MS06013), Inner Mongolia Science and Technology Program Project (Grant no. 2021GG0140), Inner Mongolia Natural Science

Foundation General Project (Grant no. 2022MS06013), and universities directly under the autonomous region Funded by the Fundamental Research Fund Project (Grant no. JY20220122).

References

[1] S. F. Lehrer and T. Xie, "The bigger picture: combining econometrics with analytics improves forecasts of movie success," *Management Science*, vol. 68, no. 1, pp. 189–210, 2022.

[2] A. Badawy, K. Lerman, and E. Ferrara, "Who falls for online political manipulation?" in *Proceedings of the 2019 World Wide Web Conference*, pp. 162–168, San Francisco, CA, USA, May 2019.

[3] D. Cao, L. Miao, H. Rong, Z. Qin, and L. Nie, "Hashtag our stories: hashtag recommendation for micro-videos via harnessing multiple modalities," *Knowledge-Based Systems*, vol. 203, pp. 106114–114, 2020.

[4] M. S. Kraus, T. M. Walker, D. Perkins, and S. E. K. Richard, "Basic auditory processing and emotion recognition in individuals at clinical high risk for psychosis. Schizophrenia research," *Cognition*, vol. 27, p. 100225, 2021.

[5] H. Zhang, *Application Research of Facial Emotion Recognition Based on Deep Learning in Intelligence Education*, North University of China, Taiyuan, China, 2021.

[6] I. Rúa-Hidalgo, M. Galmes-Cerezo, C. Cristofol-Rodríguez, and I. Aliagas, "Understanding the emotional impact of GIFs on instagram through consumer neuroscience," *Behavioral Sciences*, vol. 11, no. 8, p. 108, 2021.

[7] F. Z. Canal, T. R. Müller, J. C. Matias et al., "A survey on facial emotion recognition techniques: a state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022.

[8] H. Arabian, V. Wagner-Hartl, and K. Moeller, "Traditional versus neural network classification methods for facial emotion recognition," *Current Directions in Biomedical Engineering*, vol. 7, no. 2, pp. 203–206, 2021.

[9] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[10] R. Plutchik, "The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[11] C. E. Izard, *Basic Emotions, Relations Among Emotions, and Emotion-Cognition Relations*, American Psychological Association, Washington, DC, USA, 1992.

[12] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.

[13] B. Jou, S. Bhattacharya, and S. F. Chang, "Predicting Viewer Perceived Emotions in Animated GIFs," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 213–216, Mountain View, CA, USA, June 2014.

[14] W. Chen and R. W. Picard, "Predicting Perceived Emotions in Animated GIFs with 3D Convolutional Neural Networks," in *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)*, pp. 367-368, San Jose, CA, USA, December 2016.

[15] Z. Yang, Y. Zhang, and J. Luo, "Human-centered Emotion Recognition in Animated Gifs," in *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1090–1095, Shanghai, China, July 2019.

[16] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, and J. Luo, "Sentiment recognition for short annotated GIFs using visual-textual fusion," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1098–1110, 2020.

[17] M. Lin, *Sentiment Analysis of Short Annotated Videos Integrating Text Information*, Nanjing University of Posts and Telecommunications, Nan Jing Shi, China, 2020.

[18] Y. Ma, Y. Wang, P. Zhu, J. Pan, and H. Shi, "Get to the point: content classification of animated Graphics Interchange formats with key-frame attention," in *Proceedings of the 2021 IEEE international conference on image processing (ICIP)*, pp. 409–413, Anchorage, AK, USA, August 2021.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[20] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[21] T. Ma, L. Zhang, X. Diao, and O. Ma, "ConvGRU in fine-grained pitching action recognition for action outcome prediction," 2020, https://arxiv.org/abs/2008.07819.

[22] X. Shi, Z. Gao, L. Lausen et al., "Deep learning for precipitation nowcasting: a benchmark and a new model," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5622–5632, Long Beach CA, USA, December 2017.

[23] P. Mahalakshmi and N. S. Fatima, "Ensembling of text and images using deep convolutional neural networks for intelligent information retrieval," *Wireless Personal Communications*, pp. 1–19, 2021.

[24] K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3154–3160, Venice, Italy, January 2017.

[25] Y. Wang, *Research on action recognition based on resnet3D convolutional network and action knowledge base*, Xidian University, 2021.

[26] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., "Long-term Recurrent Convolutional Networks for Visual Recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, Boston, MA, USA, June 2015.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision. 2015*, pp. 4489–4497, Santiago, Chile, December 2015.

[28] S. Liu, Y. Ren, L. Li, X. Sun, Y. Song, and C. C. Hung, "Micro-expression recognition based on SqueezeNet and C3D," *Multimedia Systems*, pp. 1–10, 2022.