3DSSCN: A Sentiment Analysis Model for Short Video Based on 3D-Dense Network

Michele Silva, Heitor Kruger, Muhammad Vieira, Carlos Stellato,

Abstract—In recent years, with the development of social media, people are more and more inclined to upload text, pictures and videos on the platform to express their personal emotions, thus the number of short videos is increasing and becoming the first choice for people to socialize. Unlike the traditional way, people can convey their personal emotions and opinions through media other than words, such as video images, etc. for external information. Therefore, the expression and analysis of emotions is not only through text, but also through the analysis of emotional needs in images and videos, and the research scholars have customized products for individual users. Compared with pure text content, video information can more intuitively express users' happiness, anger and sorrow, thus short video-related applications have gained more and more popularity among Internet users in recent years. However, not all short videos on social networking sites can accurately express users' emotions, and related text information can more accurately assist sentiment analysis and thus improve accuracy. However, short video sentiment analysis based on video frame images is inaccurate in some scenarios, such as when expressing tears of joy, the sentiment expressed by the user's facial expression and voice are different, which will cause errors in the analysis of sentiment. As a result, researchers began to consider multimodal sentiment analysis to reduce the impact of the above scenarios on short video sentiment analysis. This paper focuses on proposing a sentiment analysis method for short videos. We first propose a residual attention model to make full use of the information in audio to classify the emotions contained in them. Then the text information in the dataset is classified by feature extraction. The key to extract features from text information is not only to retain the semantic information of the text, but also to explore the potential emotional information in the text, so as to ensure the integrity of the text information features. The experiments show that the sentiment analysis model proposed in this paper is more superior than the baselines.

Index Terms—Short video, Sentiment Analysis, Feature, 3D Dense Net, 3D Residual Network.

1 Introduction

In recent years, with the development of social media, people are more and more inclined to upload text, pictures and videos on the platform to express their personal emotions, thus the number of short videos is increasing and becoming the first choice for people to socialize [1]. Unlike the traditional way, people can convey their personal emotions and opinions through media other than words, such as video images, etc. for external information. Therefore, the expression and analysis of emotions is not only through text, but also through the analysis of emotional needs in images and videos, and the research scholars have customized products for individual users [2]. Compared with pure text content, video information can more intuitively express users' happiness, anger and sorrow, thus short video-related applications have gained more and more popularity among Internet users in recent years. However, not all short videos on social networking sites can accurately express users' emotions, and related text information can more accurately assist sentiment analysis and thus improve accuracy [3].

Recently, video has gradually become an important resource in the web [4]. Sentiment analysis in video has received more and more extensive attention from researchers, and at present, a number of researchers have been engaged in the study of video sentiment analysis. Tran et al. [5] used 3D convolutional neural networks to extract the temporal features in video, and the deep network extracted the features better compared with the traditional video feature extraction methods. The research on emotion recognition of video, although some results have been achieved, is still some distance away from the desired human-computer interaction capability and semantic understanding of emotion due to the complexity and diversity of emotion and the human-computer interaction capable nature of video heterogeneity. The sentiment analysis discussed in the existing research mostly refers to the sentiment analysis in text (especially short text sentiment analysis), such as tweets and movie reviews [6]. Since text is abstract, individual text is independent of each other, and text with different tones carries different emotions of users, it is far from enough to complete sentiment analysis based on text alone. is far from enough. Considering that short videos have become a more mainstream sentiment carrier, researchers have started to think about the sentiment analysis can be done based on the frame information of short videos. However, short video sentiment analysis based on video frame images is inaccurate in some scenarios, such as when expressing tears of joy, the sentiment expressed by the user's facial expression and voice are different, which will cause errors in the analysis of sentiment. As a result, researchers began to consider multimodal sentiment analysis to reduce the impact of the above scenarios on short video sentiment analysis. Chen et al. [7], their research is mainly based on bimodal sentiment recognition of facial

*Michele Silva and Heitor Kruger are the corresponding authors.

- *Michele Silva, Heitor Kruger, Muhammad Vieira, and Carlos Stellato are with Federal University of Technology-Paran, Brazil. (e-mail: Michelesilva.utfpr@hotmail.com,HeitorKruger.css@utfpr.edu.br).*

expression and voice, by extracting the sentiment features of both modalities, then fusing the sentiment features of both modalities as the total sentiment features, and finally judging the sentiment category by classifier. PORIA et al. [8] used convolutional neural networks (CNN) to extract text, audio and visual features to connect their features, and multiple kernel learnin (MKL) for final sentiment classification. MA et al. [9] fused audio modalities and EEG signals for sentiment recognition. WU et al. [10] fused audio and text features at the decision level. The above studies show that multimodal systems have better performance than any single-modal system. Therefore, multimodal sentiment analysis [11, 12] has become a more important research direction for current short video sentiment analysis. Pang et al. [13] used support vector machines and Bayes as classifiers in their model to classify sentiment using a machine learning approach and achieved good classification results. The key to sentiment analysis, as with other supervised machine learning, is the selection of effective features. Unsupervised sentiment classification methods mainly use sentiment dictionaries to determine sentiment tendencies. Turney et al. [14] find matching words from the text, develop the pattern by hand, and use PMI algorithm to select a set of positive words and a set of negative words as benchmark words, and subtract the point-to-point mutual information of a word and a positive word from the point-to-point mutual information of the word and a negative word to get a difference value, and then the sentiment tendency of the word can be calculated based on this difference value to determine the sentiment tendency of the text [15]. Xu et al. [16] were able to judge the sentiment tendency of a text more accurately by introducing semantic information in word embedding to enhance the expression of text information. Hu et al. [17] used WordNet to calculate the sentiment tendency of each keyword in the comment text for the comment text dataset to determine the sentiment category of the text.

This paper focuses on proposing a sentiment analysis method for short videos. We first propose a residual attention model to make full use of the information in audio to classify the emotions contained in them. Then the text information in the dataset is classified by feature extraction. The key to extract features from text information is not only to retain the semantic information of the text, but also to explore the potential emotional information in the text, so as to ensure the integrity of the text information features. The experiments show that the sentiment analysis model proposed in this paper is more superior than the baselines.

2 Related Work

2.1 Video Sentiment Analysis

Traditional machine learning algorithms focus on the selection and extraction of features from the input data, and their models are relatively fixed, while deep learning methods focus more on modeling the extracted features, and directly use the image as input to learn the features of the image autonomously, avoiding the workload of manual design of data features in traditional machine learning algorithms, and focusing on modeling the features can better The sentiment in images can be better analyzed.

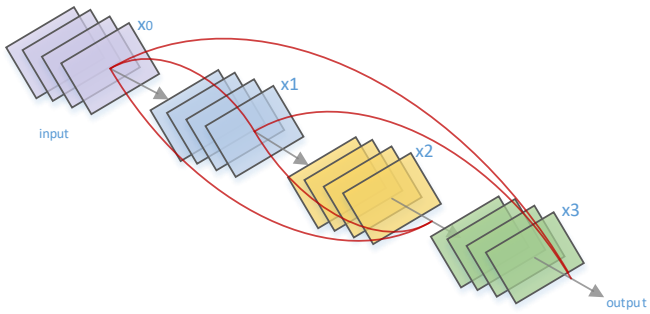

Fig. 1: Dense Block Structure Diagram.

Therefore, it is meaningful to use deep learning related algorithmic models in sentiment analysis of images and videos. The multimedia messages on social platforms are very diverse, from text and audio as the mainstay in the past, to images later, and now video streaming messages. A large number of researchers are promoting the development of cognitive intelligence by studying computer recognition of human sentiments. Along with the deepening research on sentiment recognition for images, sentiment recognition for short videos is gradually attracting people's attention. The video sentiment recognition task has two main challenges [18]. The first is the difficulty in understanding the content of the target video, i.e., when extracting the features of the video, it is necessary to extract not only a series of spatial dimensional features such as color, line, texture, etc. of the images in the video as image features, but also the video timing features between frames. The second challenge is the semantic divide problem, i.e., the human perception of the sentiment reflected by the video.

The role of sentiment classifiers is to make predictions about unknown data. The choice of classifier also plays a critical role in the final classification effect key role. Common classifiers include the plain Bayesian algorithm, support vector machines [19], etc. The plain Bayesian algorithm is one of the classical machine learning algorithm, which originates from classical mathematical theory, is a classification algorithm based on Bayes' theorem and the assumption of conditional independence. The classification efficiency is relatively high and the algorithm is simple and easy to understand, but the plain Bayesian algorithm also has some limitations, it needs to know the prior probability It is sensitive to the presentation of the input data. Support vector machine is a binary classification model that It can solve the classification problem of small samples by finding a hyperplane to segment the samples. Support vector machines have excellent generalization ability, and its computational complexity depends on the number of support vector machines rather than the spatial dimensionality of the samples, thus avoiding the The problem of dimensional disaster is avoided, and the problem of neural network structure selection and local

minima is avoided, however, in practical applications However, in practical applications, support vector machines have limitations for solving multi-classification problems.

In many scenarios, there are many duplicate information in each frame of a video. Due to the existence of a large amount of redundant information, the workload is huge if each frame is processed, so it is not possible to process every frame of images in a video [20]. Video key frame extraction is to sample the repetitive content in the video, remove the redundant frames, and select the key frames that can describe the video content to reduce the computational effort [21]. Image preprocessing can eliminate irrelevant information in the image by segmenting the extracted keyframes, and can also enhance the identifiability of the relevant information by operations such as enhancement. Image preprocessing operations can simplify the data to the maximum extent and select the useful information features in the image to improve the value of the video frame. Feature extraction is generally performed using a convolutional neural network model for deep learning operations. The design of the classifier is a cross-cutting discipline with many research points in several fields, and its selection directly affects the final result of the classification task and is where many researchers are currently improving and innovating on the classification task [22, 23].

2.2 Convolutional Neural Networks

Convolutional neural network (CNN) is a widely used researcher in recent years CNN is an efficient recognition method based on neurocognitive machine model for deep learning, which is highly valued in several fields of pattern recognition and machine learning because it avoids the complicated pre-processing of input images and CNNs are highly valued in several fields of pattern recognition and machine learning because they avoid the complicated preprocessing of input images and the large workload of feature selection. As early as in 1980 A new recognition machine proposed by K. Fukushima in 1980 was the first network to implement a convolutional neural network. The following research of AlexNet [24], VGG-16 [25], etc. were developed based on convolutional neural networks. By simulating two important perceptual units in human visual cortex, the complex unit and the simple unit, researchers have designed the convolutional and pooling layers in the CNN model.

The overall structure of a convolutional neural network is a multilayer feedforward network that contains an input layer that directly receives two-dimensional visual images that do not require much manual processing, a convolutional layer and a pooling layer that extract image features and feature mapping, and finally a fully connected layer. Each convolutional layer consists of multiple convolutional neurons, each of which is connected to a perceptual region at the corresponding location in the previous layer, a feature known as local perception in CNNs, and the image features are extracted from the corresponding local region according to the different weights of the convolutional neurons connected to the local perceptual region in the previous layer. Also, in order to restrict the convolutional layer to extract only the same features at different locations in the previous layer of the network when extracting image features, a weight sharing strategy for CNNs is applied, i.e., restricting different neurons in the same layer of the convolutional layer to have equal weights when they are connected to the corresponding local sensory domain [26]. Many current models based on convolutional neural networks extract richer features by increasing the number of convolutional layers [27]. Each pooling layer consists of multiple sampling neurons, similar to convolutional neurons, which are connected to the perceptual regions at the corresponding locations in the previous layer, except that the connection weight of each sampling neuron to the corresponding region is a fixed value, so the pooling layer does not generate new parameters during the model training [28]. At the same time, the down-sampling of features from the latter pooling layer to the previous layer reduces the memory consumption of the entire network model and improves the robustness of the network model to potential deformations of the input pattern. In common neural network models, maximum pooling or average pooling is generally used. The main role of the fully connected layer of a CNN is to stretch the last layer of the hidden layer (convolutional or pooling layer). The main function of FC (Fully Connect) is to stretch the two-dimensional feature map obtained from the last layer of the hidden layer (convolutional or pooling layer) to obtain a one-dimensional vector for the final classification. In common neural network models, the multi-layer fully connected layer is generally designed to better map the target features extracted from the hidden layer to the output category labels [29].

3 Methodology

3.1 Short Video Nets

Sentiment analysis of visual content can help to correctly identify the sentiment contained in short videos. Compared with traditional feature extraction methods such as SIFT and HOG, deep learning methods are widely used because feature extraction becomes more and more complete as the depth of the network increases [30]. First, since 2D CNN does not consider the temporal coherence of features in video frame sequences [31], it can reduce the accuracy of short video sentiment classification results. Second, anomalies in short videos (e.g., distorted target positions) can also degrade the accuracy of sentiment results. Finally, for CNN-based methods, as the depth increases, it will fall into local optimal solutions, which will also reduce the accuracy of the results and the training process becomes extremely slow. In this paper, we use a 3D convolutional form of ResNet. The time dimension is the third 3D convolution is performed by stacking multiple consecutive frames to form a cubic The 3D convolution is achieved by stacking several consecutive frames to form a cube, and then applying a 3D convolution kernel. In this structure, each feature map in the In this structure, each feature map in the convolution layer is connected to multiple neighboring frames in the previous layer, thus capturing temporal information.

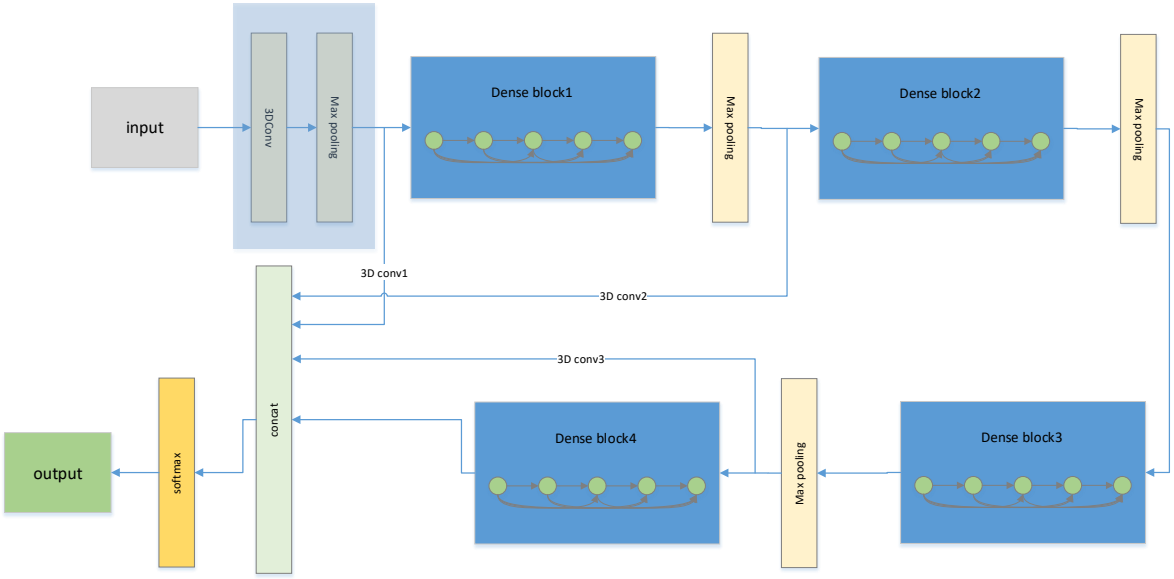

Fig. 2: Framework of our model.

3.2 Pooling Layer

A global maximum pooling operation, which is equivalent to the global average pooling operation in the original compression operation, is performed. The global maximum pooling operation is used because it can effectively preserve the background information and maximize the extraction of useful texture features, which is important for performing sentiment classification.

$$\alpha_i = argminL(\alpha, D) \tag{1}$$

$$h_{m,j} = max_{i \in N_m} \alpha_{i,j}, j = 1, \ldots, K \tag{2}$$

where α_i is the minimum distance between each feature vector and the reference after the convolution operation, and the maximum value of which can be obtained by the max function.

Since the output value after the global maximum pooling operation sometimes fluctuates in magnitude fluctuates greatly. It is necessary to add a normalization operation.

$$z_c = Norm(h_{m,j}) \tag{3}$$

Full capture of channel dependence by excitation. and employs a simple gating mechanism using the Sigmoid activation function as follows.

$$S = F_{ex}(z, W) = \delta(g(z, W)) = \delta(W_2 \theta(W_1 z)) \tag{4}$$

where δ is the Sigmoid function, θ is the linear activation function operation. W_1 and W_2 prevent the model from becoming complex and to take into account generalization factors, 2 layers of fully connected layers need to be set up around the nonlinearity, which acts as a bottleneck to parameterize the gate mechanism. Finally, the Sigmoid function is passed again to obtain S.

After obtaining S, the output of the original residual network can be manipulated.

$$\tilde{x} = F_{scale}(U_c, S_c) \tag{5}$$

where U_c is the output of the residual network, a two-dimensional matrix with subscript c denoting the channel. S_c is a value in the output S vector from the previous step and is also the weight, so it is equivalent to multiplying each value in the U_c matrix by S_c.

3.3 3D Residual Network

The original convolutional neural networks focus on processing image data cannot capture the features in time dimension well, while video feature extraction includes not only spatial information but also information in time dimension, so 3D convolutional neural networks have great application value in the fields of human pose reconstruction and video sentiment classification, and the residual dense networks also have great advantages in building deep networks. The residual dense network is reconstructed to make it suitable for handling aspects of video sentiment analysis. In this paper 5 3D convolutions are used for each 3D residual dense block. The first three convolution layers are each followed by an activation layer and a batch normalization layer. Finally, the features extracted from the network are tensor stitched and another 3D convolution layer is performed to obtain the final output tensor.

Similar to the 2D residual dense network, the 3D residual dense network is also divided into three parts, the underlying feature extraction module, the residual dense module, and the global feature aggregation module. The underlying feature extraction module includes two layers of 3D Conv. The residual density module includes multiple residual density blocks, pooling layers, and convolutional layers such as 3D Conv1 and 3D Conv2 for convolutional downsampling. The global feature aggregation module consists of a concatenate operation for features and a convolutional layer for $1 \times 1 \times 1$ feature aggregation and channel adjustment. The first two layers of the 3D residual dense network are used to extract the underlying features, which can be represented as follows.

$$P = F_{sh}(P_{in}) \tag{6}$$

where F_{sh} denotes the composite function of the first two convolutional and downsampling operations and P_{in} is the input to this model. P is the extracted feature map, which is used as the input of the first residual dense block.

$$P_n = F_{3D-RDB,n}(F_{3D-RDB,n-1}(\dots(F_{3D-RDB,0}, P_0))\dots) \tag{7}$$

where $F_{3D-RDB,n}$ denotes the nth residual dense block (3D-RDB) and the computational operation of downsampling (Maxpooling), and $F_{3D-RDB,n}$ is a composite operator function, which contains multi-layer convolution and rectified linear units. In the model designed in this paper, four residuals are set dense blocks.

The input features P_n from different levels are convolutionally sampled to generate $1 \times 7 \times 7$ feature maps X_n , and then normalized using the l_2 parametric process. Then the local features X_n from different levels are stitched together. Finally, the $1 \times 1 \times 1$ convolution is used for feature aggregation and channel adjustment to obtain the global feature map.

$$P_G = F_G(X_0, X_1, \dots, X_n) \tag{8}$$

In which, P_G is the feature map output by global feature aggregation and F_G is the composite function of $1 \times 1 \times 1$ convolution. $[X_0, X_1, \dots, X_n]$ is the concatenate of 3D residual dense blocks and the feature map after convolution sampling. After the local features are extracted using multiple 3D residual dense blocks, the global features are obtained by the global feature aggregation operation. Finally, a nonlinear mapping function is used to map the global features to the specified sentiment category space.

$$d_c = tanh(W_c P_G + b_c) \tag{9}$$

where W_c is the parameter matrix of global features and b_c is the bias vector. Finally, the mapped short video features are classified using a softmax classifier to obtain the sentiment category of short videos.

$$p_c^i = \frac{exp(d_c^i)}{\sum_{i=1}^{c} exp(d_c^i)} \tag{10}$$

where c is the number of categories for sentiment classification, d_c^i is the representation of short video features mapped to the ith sentiment category, and p_c^i is the predicted probability of a video with sentiment category i.

3.4 Loss Function

The cross-entropy loss function is used to optimize the trained model, and the loss function is calculated as follows.

$$J = -\sum_{i=1}^{c} y_i log(p_c^i) \tag{11}$$

where c is the number of categories for sentiment classification, and y_i denotes the ground truth of the sample as the i-th category. The stochastic gradient descent method is used to optimize the loss function and learn to obtain the minimum J .

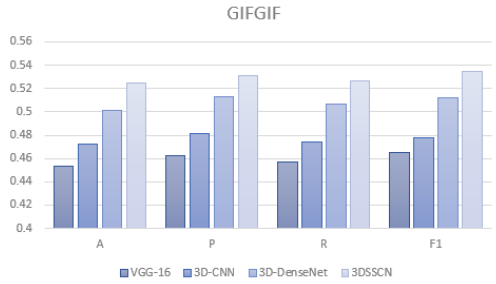

Fig. 3: Comparison of different models on GIFGIF dataset.

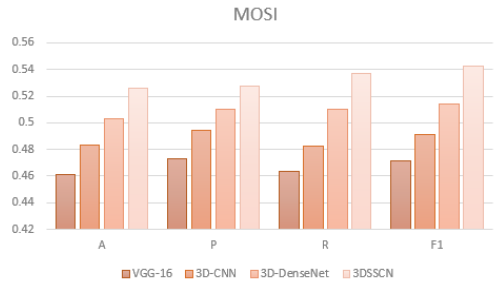

Fig. 4: Comparison of different models on MOSI dataset.

4 EXPERIMENTS

4.1 Datasets

The GIFGIF dataset is a short video site provided by the MIT Multimedia Lab.The dataset has 6,119 short videos divided into 17 categories: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilty, happiness, pleasure, pride, relief, sadness, satisfaction, shame, surprise. The data is obtained from 3,272,523 users who voted on the emotion categories. A total of 5,100 short videos, 300 from each category, are used as the dataset for this paper. The MOSI dataset is a three-modality dataset containing text, speech, and video [40], and each video contains a piece of sentiment-labeled text data. A total of 2199 samples were collected in this dataset, which were obtained from 89 web users. The MOSI dataset is a publicly available dataset and one of the most frequently used datasets in multimodal sentiment computing. The sentiment label in this dataset corresponds to a sentiment score, and in the MOSI dataset, the sentiment label corresponds to a sentiment score of -3 to 3. In the MOSI dataset, this sentiment score ranges from -3 to +3. +3 indicates strongly positive, +2 indicates positive, +l indicates more positive, 0 indicates neutral, -1 indicates more negative, and -2 indicates more negative. -1 means more negative, -2 means negative, and -3 means strongly negative.

4.2 Metrics

In our task, Precision, Accuracy,Recall, and F-Measure are the most commonly used evaluation metrics. P is the probability of actual positive samples among all predicted positive samples. A is the number of correctly classified categories as a percentage of the total sample. R is the

probability that the actual positive sample is predicted to be positive. F1 is the average of the combined P and R. The formulae for each of the above evaluation indicators are as follows.

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

where TP indicates the number of correctly discriminated positive categories. FP indicates the number of incorrectly discriminated negative categories. TN indicates the number of correctly discriminated negative categories. FN indicates the number of incorrectly identified positive categories.

4.3 Result Analysis

The model designed in this paper (3DSSCN) improves the video sentiment classification on P, R, F1, and A. The 3D-CNN network improves 1.12%, 1.56%, 1.47%, and 1.98% on P, R, F1, and A, respectively, compared with the VGG-16 network. 3D-CNN can extract video feature information better than VGG network. The 3D-DenseNet network improves 2.13%, 0.31%, 0.51%, 1.35% on P, R, F1, A, respectively, compared to 3D-ResNet network. The 3DSSCN network improves 1.37%, 2.75%, 2.33%, and 1.17% over the 3D-DenseNet network for P, R, F1, and A, respectively. It can be clearly seen that the classification performance of the 3D residual dense network model is higher than that of the VGG-16 network, 3D-CNN network, and 3D-DenseNet network because the 3D residual dense network extracts multi-level spatio-temporal features through residual learning and dense connections, thus reducing the risk of losing the original video information during the detection process. The sentiment recognition effect of the model used in this paper is improved by the previous models. This indicates the effectiveness of the video sentiment classification model proposed in this paper.

From Table 2, we can see that: the video sentiment classification of the model designed in this paper (3DSSCN) improves on P, R, F1, and A. The 3D-CNN network improves on P, R, F1, and A by 2.07%, 2.43%, 1.21%, and 2.15%, respectively, compared to the VGG-16 network, which is due to the problem of video analysis for which the 2D network cannot capture the information on timing well, and the 3D- CNN is better at extracting video feature information compared to VGG network. The 3DSSCN network improves 1.42%, 2.75%, 2.46%, and 1.48% in P, R, F1, and A, respectively, over the 3D-DenseNet network. It can be clearly seen that the classification performance of 3D residual dense network model is more effective than that of VGG-16 network, 3D-CNN network and 3D-DenseNet network due to the fact that 3D residual dense network extracts multi-level features through residual learning and dense connections, thus reducing the risk of losing the original video information in the detection process.

5 Conclusion and Future Work

Recently, video has gradually become an important resource in the web. Sentiment analysis in video has received more and more extensive attention from researchers, and at present, a number of researchers have been engaged in the study of video sentiment analysis. Tran et al. used 3D convolutional neural networks to extract the temporal features in video, and the deep network extracted the features better compared with the traditional video feature extraction methods. The research on emotion recognition of video, although some results have been achieved, is still some distance away from the desired human-computer interaction capability and semantic understanding of emotion due to the complexity and diversity of emotion and the human-computer interaction capable nature of video heterogeneity. The sentiment analysis discussed in the existing research mostly refers to the sentiment analysis in text (especially short text sentiment analysis), such as tweets and movie reviews. Since text is abstract, individual text is independent of each other, and text with different tones carries different emotions of users, it is far from enough to complete sentiment analysis based on text alone. is far from enough. Considering that short videos have become a more mainstream sentiment carrier, researchers have started to think about the sentiment analysis can be done based on the frame information of short videos. However, short video sentiment analysis based on video frame images is inaccurate in some scenarios, such as when expressing tears of joy, the sentiment expressed by the user's facial expression and voice are different, which will cause errors in the analysis of sentiment. As a result, scholars began to consider multimodal sentiment analysis to reduce the impact of the above scenarios on short video sentiment analysis. Chen et al., their research is mainly based on bimodal sentiment recognition of facial expression and voice, by extracting the sentiment features of both modalities, then fusing the sentiment features of both modalities as the total sentiment features, and finally judging the sentiment category by classifier. PORIA et al. used convolutional neural networks (CNN) to extract text, audio and visual features to connect their features, and multiple kernel learnin (MKL) for final sentiment classification. MA et al. fused audio modalities and EEG signals for sentiment recognition. WU et al. fused audio and text features at the decision level. The above studies show that multimodal systems have better performance than any single-modal system. Therefore, multimodal sentiment analysis has become a more important research direction for current short video sentiment analysis. This paper focuses on proposing a sentiment analysis method for short videos. We first propose a residual attention model to make full use of the information in audio to classify the emotions contained in them. Then the text information in the dataset is classified by feature extraction. The key to extract features from text information is not only to retain the semantic information of the text, but also to explore the potential emotional information in the text, so as to ensure the integrity of the text information features. The experiments show that the sentiment analysis model proposed in this paper is more superior than the baselines.

TABLE 1: Model Performance Comparison on GIFGIF.

Modle	A	P	R	F1
VGG-16	0.4532	0.4628	0.4571	0.4653
3D-CNN	0.4726	0.4815	0.4738	0.4782
3D-DenseNet	0.5013	0.5132	0.5067	0.5124
3DSSCN	0.5247	0.5315	0.5263	0.5345

TABLE 2: Model Performance Comparison on MOSI.

Modle	A	P	R	F1
VGG-16	0.4614	0.4732	0.4635	0.4717
3D-CNN	0.4834	0.4941	0.4822	0.4912
3D-DenseNet	0.5034	0.5099	0.5105	0.5141
3DSSCN	0.5256	0.5274	0.5374	0.5428

6 Conflict of Interest Statement

All authors have no conflict and declare that: (i) no support, financial or otherwise, has been received from any organization that may have an interest in the submitted work ; and (ii) there are no other relationships or activities that could appear to have influenced the submitted work.

References

[1] S.-M. Choi, S.-K. Ko, and Y.-S. Han, "A movie recommendation algorithm based on genre correlations," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8079–8085, 2012.

[2] A. M. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 278–288.

[3] H. Zarzour, Z. Al-Sharif, M. Al-Ayyoub, and Y. Jararweh, "A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques," in *2018 9th international conference on information and communication systems (ICICS)*. IEEE, 2018, pp. 102–106.

[4] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.

[5] 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.

[6] X. Yu, T. Gan, Y. Wei, Z. Cheng, and L. Nie, "Personalized item recommendation for second-hand trading platform," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3478–3486.

[7] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 366–371.

[8] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[9] J. Ma, Y. Sun, and X. Zhang, "Multimodal emotion recognition for the fusion of speech and eeg signals," *Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University*, vol. 46, no. 1, pp. 143–150, 2019.

[10] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2010.

[11] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[12] I. O. Hussien and Y. H. Jazyah, "Multimodal sentiment analysis: a comparison study," *Journal of Computer Science*, vol. 14, no. 6, pp. 804–818, 2018.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[14] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *acm Transactions on Information Systems (tois)*, vol. 21, no. 4, pp. 315–346, 2003.

[15] H. Jiang, W. Wang, Y. Wei, Z. Gao, Y. Wang, and L. Nie, "What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3487–3495.

[16] H. Xu, M. Dong, D. Zhu, A. Kotov, A. I. Carcone, and S. Naar-King, "Text classification with topic-based word embedding and convolutional neural networks," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016, pp. 88–97.

[17] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[18] D. Lin, D. Cao, Y. Lv, and Z. Cai, "Gif video sentiment detection using semantic sequence," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[19] H. Xia, M. Tao, and Y. Wang, "Sentiment text classification of customers reviews on the web based on svm," in *2010 Sixth International Conference on Natural Computation*, vol. 7. IEEE, 2010, pp. 3633–3637.

[20] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, "Personalized hashtag recommendation for micro-videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1446–1454.

[21] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 483–487.

[22] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," in *Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 189–200.

[23] ——, "A feature-based algorithm for detecting and classifying production effects," *Multimedia systems*, vol. 7, no. 2, pp. 119–128, 1999.

[24] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[25] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.

[26] Y. Cao, R. Xu, and T. Chen, "Combining convolutional neural network and support vector machine for sentiment classification," in *Chinese national conference on social media processing*. Springer, 2015, pp. 144–155.

[27] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 1–14, 2019.

[28] S. Ruder, P. Ghaffari, and J. G. Breslin, "Insight-1 at semeval-2016 task 4: Convolutional neural networks for sentiment classification and quantification," *arXiv preprint arXiv:1609.02746*, 2016.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[30] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-

 doi:10.20944/preprints202212.0132.v1

generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[31] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.