 doi:10.20944/preprints202207.0308.v1

Multimodal Micro-video Classification Based on 3D Convolutional Neural Network

Yueyue Sun, Bin Chen, Fang Wei, Xinyue Chen, Qiang Gong, Peng Zhang

Abstract—Along with the popularity of the Internet, people are exposed to more and more ways of micro-videos, and a huge amount of micro-video data has emerged. micro-videos have gradually become the Internet content preferred by the public, and a large number of micro-video apps have also emerged, such as Tictok and Kwai. Intelligent classification and mining of micro-videos can greatly enhance user experience, improve business operation efficiency and enhance user experience. Through deep intelligent analysis and mining of micro-videos, important information in micro-videos can be extracted to provide an important basis for beautifying videos, content appreciation, video recommendation, content search, etc. In the past, content understanding for short videos often used human work annotation, but in recent years, with the great success of deep convolutional neural networks in image recognition, short video content understanding based on this method has gradually developed. Nowadays, most of the recognition algorithms extract the feature representation of each frame independently and then fuse them. However, while extracting the feature representation, some low-level semantic features are lost, which makes the algorithm unable to accurately distinguish the category of the video. At present, the algorithm of micro-video recognition based on deep learning has surpassed the iDT algorithm, making these traditional methods fade out of people's view. In this paper according to the micro-video classification task, a new network model is proposed to concat features of each modality into the overall features of various modalities through the network, and then fuse the various modal features with attention mechanism to obtain the whole micro-video features, which will be used for classification. In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

Index Terms—Micro-video Classification; 3D CNN; Multi-modal

1 Introduction

In recent years, micro-videos have gradually become the Internet content preferred by the public, and a large number of micro-video apps have also emerged, such as Tictok and Kwai. Intelligent classification and mining of micro-videos can greatly enhance user experience, improve business operation efficiency and enhance user experience. Through deep intelligent analysis and mining of micro-videos, important information in micro-videos can be extracted to provide an important basis for beautifying videos, content appreciation, video recommendation, content search, etc. Youtube statistics on micro-video traffic in 2017 found that the platform plays up to 400 hours of video per minute, and the click as well as the play of online video occupies a large amount of network channel bandwidth. The average daily active users and peak hours of Tictok micro-videos can reach 300 million, which fully illustrates that micro-videos have a huge user base and traffic value. Along with the popularity of the Internet, people are exposed to more and more ways of micro-videos, and a huge amount of micro-video data has emerged. In such a large amount of micro-video data, each person is interested in only a small part of it, and it will be a very difficult task to organize and manage these video data in an effective classification. micro-video classification algorithms, in general, are processed by obtaining the semantic related content of the video, such as analyzing and understanding human behavioral actions or other complex video clips, and finally classifying them into single or multiple categories automatically [1].

Before deep learning methods were widely used, most micro-video classification methods mainly relied on manually designed features [2] and machine learning methods for human action behavior recognition and specific event detection in videos. The core idea of these traditional micro-video classification studies is to extract action and appearance information from local spatio-temporal regions [3], obtain feature descriptors of video frames, and then use methods such as BagofWords model [4] to generate an encoding of the whole video, and finally train a machine learning classifier to achieve video classification. Among the many methods, the ODT algorithm [5] published in ICCV by the IEAR lab of INRIA is a very classical one. With the advent of convolutional neural networks, one no longer needs to design feature descriptors manually, but automatically learns video semantic features and understands image contents by convolutional neural networks, which eventually has achieved great success in image classification, detection and retrieval. At present, the algorithm of micro-video recognition [6] based on deep learning has surpassed the iDT algorithm, making these traditional methods fade out of people's view. FaceBookAIRearch [7] has proposed the use of 3D convolutional neural network to extract spatio-temporal information, which is an innovative approach for video classification. Karpathy et al. [8] proposed the concept of slow fusion of 3D convolutional networks to improve the temporal perception of convolutional networks. Simonyan et al. proposed the TwoStream method [9], which created a new direction in video research. The method based

*Yueyue Sun is the corresponding author.

- *Yueyue Sun, Bin Chen, Fang Wei, Xinyue Chen, Qiang Gong, and Peng Zhang are with Yunnan University, China. (e-mail: sunyueyue.dlu@gmail.com).*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2022 doi:10.20944/preprints202207.0308.v1

on TwoStream network will split the network into two branches and the results of both networks will be fused to obtain the final class label.

Wu et al. [10] proposed different methods to combine timing feature information, which focus on how to combine the video timing information extracted by the trained CNN network. Firstly, a deep convolutional neural network is trained to extract the feature description vectors of different video frames, and then these temporal feature vectors are combined by designing temporal models, such as LSTM [11], NetVlad [12], etc., to obtain the category of video content. Ji proposed to use 3D convolution to extract the spatial and temporal information of video, while Tran trained a 3D convolutional neural network, this model is called C3D and is the original 3D convolutional neural network model, which is computationally efficient as compared to other types of methods that process multiple frames in C3D-session.Carreira and Zisserman proposed I3D [13], which combines two-s The 3D convolutional neural network model is a natural idea for extending the 2D convolutional neural network model that has been so successful in the field o f i mage c lassification to ap ply to video classification. by extending the 2D convolution to 3D convolution, the model is able to learn both spatial and temporal representations of the video and achieve end-to-end training. one of its drawbacks is the large number of parameters One of its drawbacks is the large number of parameters, which causes it to require a large amount of data to converge and its generalizability suffers [14].

In this paper, we proposed an efficient a nd accurate multi-modal fusion micro-video classification algorithm based on 3D convolutional neural network to further improve the classification a ccuracy. W e d esign and implement a convolutional neural network based video classification. T he n etwork c onsists o f a n i nput l ayer, a convolutional layer, a pooling layer, a BN layer, a fully connected layer and an output layer. The contriburions of this paper are as follows:

- According to the micro-video classification task, a new network model is proposed to concat features of each modality into the overall features of various modalities through the network, and then fuse the various modal features with attention mechanism to obtain the whole micro-video features, which will be used for classification.
- In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

2 Related Work

2.1 Video Classification

The temporal modeling of the last layer output of the 2D network using the LSTM series can capture the semantic information of the high layer but loses the information of the low layer. And LSTM is difficult to converge because it requires gradients to be back-propagated through the temporal layers for training. Another approach, which can be practically applied, was proposed by Simonyan and Zisserman, who modeled video frames by averaging a

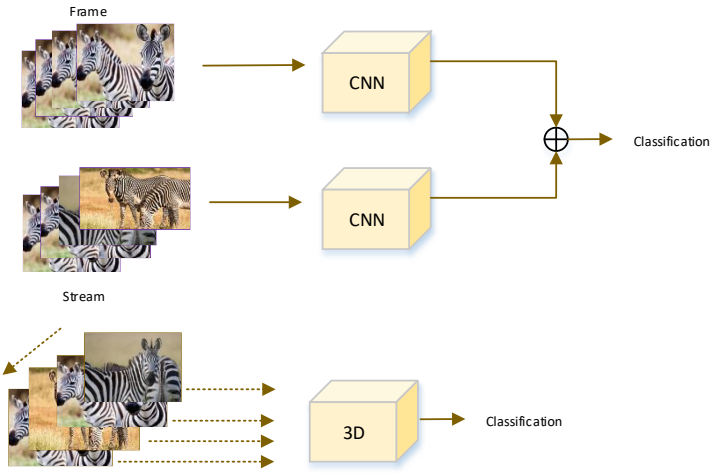

Fig. 1: Video classification model based on convolutional network.

single RGB frame and a pile of information consisting of 10 externally computed optical streams through two separate convolutional neural networks pre-trained by ImageNet. This approach achieves quite good results on top of some benchmark datasets. 3D convolutional neural networks are a natural choice for modeling video files because of the augmentation of the convolutional kernel dimensionality, which allows them to model both spatial and temporal information [15, 16].

The important feature of these methods is that they directly divide the spatial and temporal dimensions, allowing the model to be modeled in both spatial and temporal dimensions. The disadvantage of these methods is the huge computational effort, because the convolution kernel adds an extra dimension, which also leads to the additional problem that the model is difficult to train and tends to overfit on small data sets. To solve this problem, Qiu et al. [15] proposed a pre-training method, in which the model is first pre-trained on ImageNet and Kinetics datasets and then migrated to the target dataset. In this paper, in order to balance efficiency and accuracy, a temporal partial channel fusion algorithm is proposed, which is able to model temporal features on top of 2D networks. The proposed algorithm can be added to different positions of the network in the form of modules to extract both spatial and temporal semantic features at low and high levels, achieving the accuracy of a 3D network with the speed of a 2D network. This initiative has injected new dynamics into the field of video classification research. This method adds a dimension to the 2D convolutional kernel, i.e., the planar convolutional kernel becomes a three-dimensional convolutional kernel. The two-stream CNN approach was proposed by Simonyan et al. [9] where the SpatialStream network accepts the original still video frames and the TemporalStream network accepts the optical flow field as input. In the dual-stream network, the SpatialStream network uses a 2D convolutional kernel to classify the sparsely sampled picture frames, and the TemporalStream

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2022 doi:10.20944/preprints202207.0308.v1

network extracts the optical flow field information of frames around the sampling point, and finally fuses the results of the two networks to achieve classification.

2.2 Multi-modal Representation

In video classification task multi-modal representation is one of the most important problem [17, 18]. Existing multi-modal representations can be grouped into two categories. On is combine the various single-modal information into a single representation and project it into the same representation space, i.e. concatenation of single-modal features. Recently, neural networks are increasingly used in the multi-modal domain [19, 20, 21, 22], especially on multimodal representations [23, 24, 25, 26, 27]. The probabilistic graphical models [28, 29] are another way to construct a joint representation for multi-modal information using the latent random variable [30]. The other category of multi-modal representation is s learn separate representations for each modality but coordinate them with constraints. Frome et al. [31] proposed a deep visual-semantic embedding model which projects the visual information and semantic information into a common space constrained by distance between the visual embedding and the corresponding word embedding. Similarly, Wang et al. [32] constructed a coordinated space which enforces images with similar meanings to be closer to each other.

3 Methodology

3.1 Feature Extraction

Feature extraction is divided into scene feature extraction and behavior feature extraction. First, scene features are extracted formicro-videos. Let the micro-video sequence be $V = \{v_1, v_2, \ldots, v_n\}$, where, n is the number of micro-videos. The scene features are extracted using a deep fusion network based on VGGNet. The global features in the scene are learned and extracted using the VGGNet16 network, and the local detailed features in the scene are learned and extracted using VGGNet19 , and the learned features are fused respectively. The reason for using VGGNet is that the network chooses a 3×3 convolutional kernel, which makes the number of parameters smaller, and the superposition of small convolutional layers enables multiple nonlinear computations and better learning ability of features. Assuming that the number of scene categories is sceneN , for the i-th micro-video, the purpose of scene recognition is to find the maximum value of the scene prediction probability as follows:

$$p^s_{v_i} = max\{p^{s_j}_{v_i}\}(1 \le j \le N_s, 1 \le i \le n), \tag{1}$$

where, v_i is the i-th micro-video and $p^{s_j}_{v_i}$ is the probability value of the i-th micro-video corresponding to the j-th scene. But here, it is necessary to keep the probability values of v_i in all scenes to retain as much useful information in the video as possible.

$$f^s_{v_i} = \{p^{s_j}_{v_i}\}(1 \le j \le N_s, 1 \le i \le n), \tag{2}$$

Assuming that the number of behavior categories is N_a , for the i-th micro-video, the result for behavior recognition can be defined as:

$$p^A_{v_i} = max\{p^{a_k}_{v_i}\}(1 \le k \le N_a, 1 \le i \le n), \tag{3}$$

where, v_i is the i-th micro-video and $p^{a_k}_{v_i}$ is the probability value of the i-th micro-video corresponding to the k-th action. Furthermore, that can be defined as:

$$p^{a_k}_{v_i} = p^{a_k^{RGB}}_{v_i} + p^{a_k^{Flow}}_{v_i}, \tag{4}$$

Similarly, the behavioral feature extraction part, which needs to keep the probability values of each micro-video for all behaviors, is defined as follows:

$$f^A_{v_i} = \{p^{a_k}_{v_i}\}(1 \le k \le N_a, 1 \le i \le n), \tag{5}$$

By obtaining $f^s_{v_i}$ and $f^A_{v_i}$, for the i-th micro-video, the joint feature is defined as :

$$f_{v_i} = (f^A_{v_i})^T f^s_{v_i}, \tag{6}$$

where $f^A_{v_i}$, $f^s_{v_i}$ are the scene features and behavioral features of f_{v_i}, respectively, and the dimension of f_{v_i} is $N_s \times N_a$.

3.2 Attention mechanism

The main modalities present in micro-video are visual, audio and text. Most of the features currently used in video processing algorithms are mainly visual features. Audio and text features are less used in computer vision, while audio and text also contain a lot of video-related information. Thus fusing several modalities will improve the accuracy of video classification. By extracting features from the visual, audio, and title of the video separately, and then performing feature fusion, the final prediction score is obtained by the classification function. There are two main feature fusion modes, which are forward fusion and backward fusion methods. In this paper, we adopt backward fusion, firstly, we input each modal information into the clustering network to get the corresponding features, and then concat to get the final feature vector.

Assume that there are M frames of video, and the feature description x of each frame is N-dimensional. There are K clustering centers, and each frame is firstly encoded into an $N \times K$ feature vector as follows:

$$v_{ijk} = \alpha_k(x_i)(x_{ij} - c_{kj}), \tag{7}$$

where c_k is the N-dimensional feature vector coordinate of the cluster center k. $\alpha_k(x_i)$ is a soft sign function to calculate the similarity of x_i to the cluster center k. The similarity function is usually computed using a full connection and the activation function is a Softmax function:

$$\alpha_k(x_i) = \frac{e^{w^T x_i + b_k}}{\sum_{s=1}^{k} e^{w^T x_i + b_s}}, \tag{8}$$

The video-level feature descriptor y is then obtained by aggregating all the frame-level features and is expressed by the following equation:

$$y_{jk} = \sum_{i}^{M} v_{ijk}, \tag{9}$$

The set of local features is defined as the unordered features obtained in different segments of the same video, and the $L \times M$ dimensional matrix X is used to denote L local features, each row being a separate local feature vector.

$$X = (x_1, x_2, \ldots, x_L), \tag{10}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2022 doi:10.20944/preprints202207.0308.v1

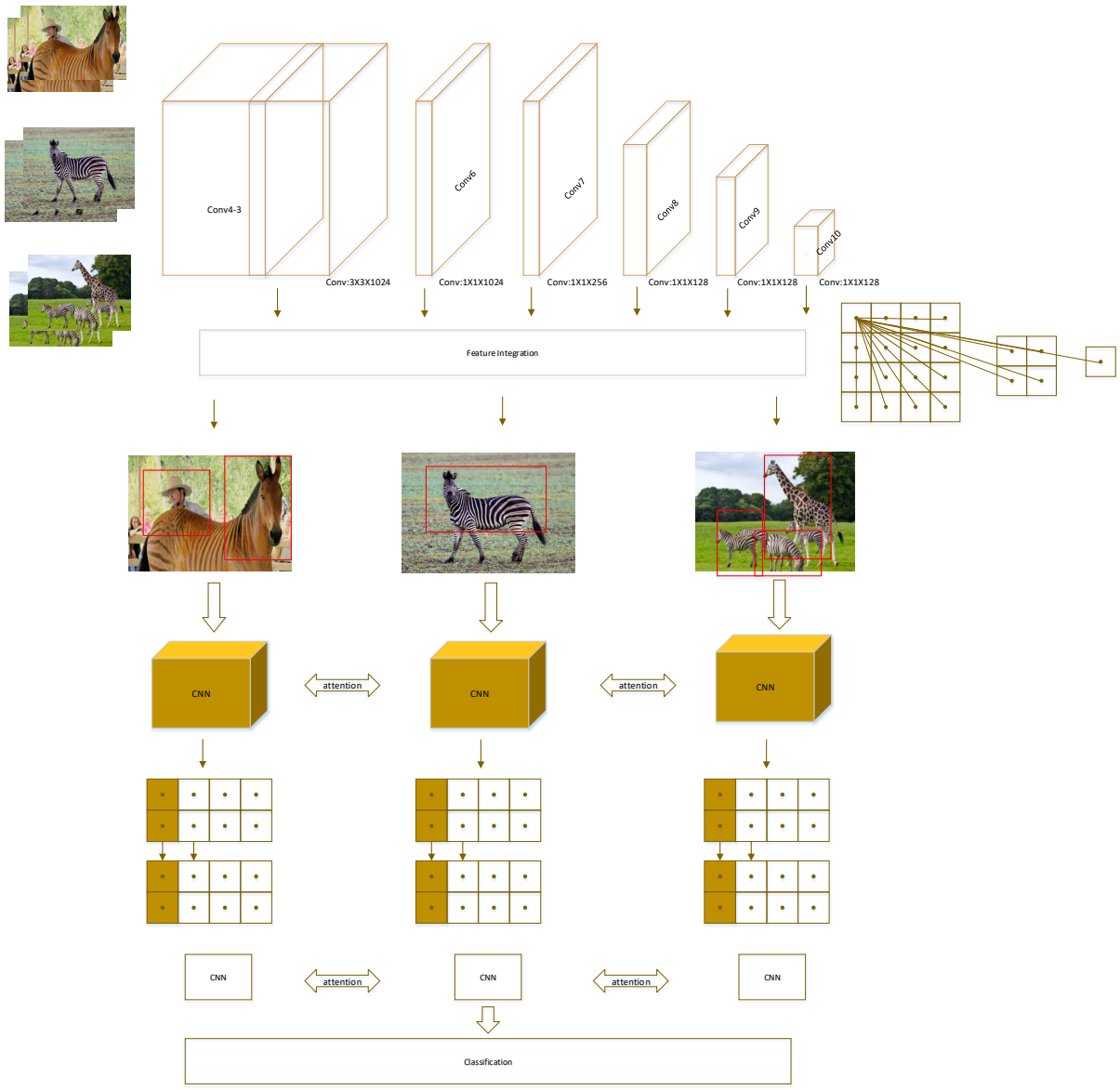

Fig. 2: Overall structure of classification.

The key frames are selected by using the attention mechanism and thus combined into global features. In the classification task, attention is static, and the input contains only its own local feature vector. The first step is to analyze the importance of each feature and then boost the weight of key features as much as possible, while ignoring irrelevant features and noise. The attention output can be considered as a weighted collection of vectors :

$$v = aX, \tag{11}$$

where a is an L-dimensional weight vector, calculated using the weight function.

It is important to choose the appropriate weight calculation function throughout the process, where the input is the set of local features X and the output is the weight vector a , which is L1-normalized to 1 . Each dimension of the weight vector corresponds to a local feature. There are more methods to to compute the weights of local features, for example, global pooling is considered as a decaying form of the attention mechanism, and its corresponding weight function is:

$$a = \frac{1}{L}l, \tag{12}$$

where l is an L-dimensional vector with all elements of 1. To obtain a more flexible attention weight function, a layer of full connectivity is used to learn the weight coefficients.

$$a = softmax(wX^T + b), \tag{13}$$

where w,b are the vectors of M and L dimensions, respectively.

4 EXPERIMENTS

4.1 Dataset

The UCF101 dataset [33] is a video classification dataset collected from real scenes, with videos from the youtube website, containing a total of 101 different action categories. the UCF101 dataset has a total of 13,320 videos from 101 categories.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2022 doi:10.20944/preprints202207.0308.v1

TABLE 1: Results comparison of 3D CNN model.

	Traditional CNN model	Ours 3D CNN model
trainning precision	0.936	0.952
validation precision	0.737	0.762
Model convergence time	1.42h	1.03h
Inference time	57ms	52ms

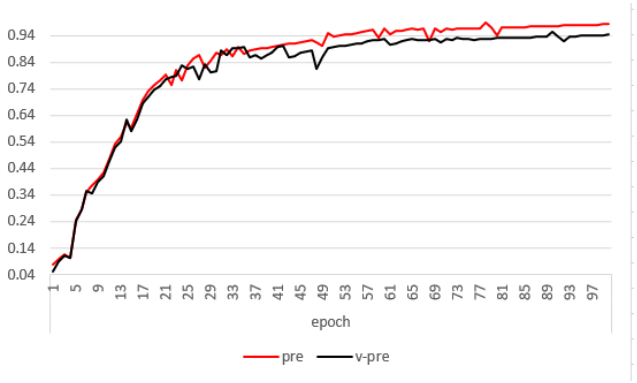

Fig. 3: Precision in training and validation sets.

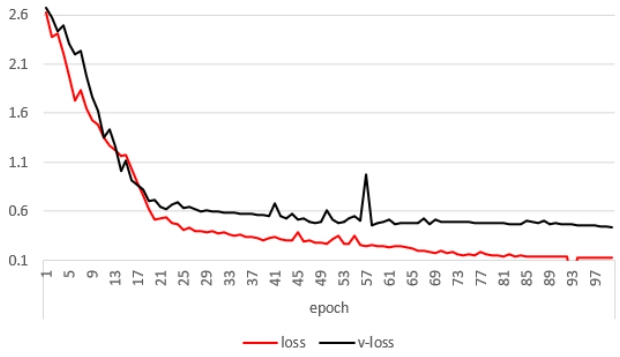

Fig. 4: Loss in training and validation sets.

4.2 Pre-processing

The micro-video is read frame by frame and the image is captured at a frame rate of 50 Fps. In the practical problem, considering that this micro-video dataset contains multiple human behavioral actions, since there is not much difference between human video actions in any one second. Therefore, in order to save computational resources and ensure the non-redundancy of training data, only a partial subset of complete frames is extracted from any one continuous action, and in the extraction process, it is necessary to ensure that each image has the same spatial dimension. The whole video preprocessing process is implemented using FFmpeg. The number of videos and frames per training, the learning rate, the overfitting parameter (Dropout), the number of network iterations, the number of classifications, the number of samples per input network training Batchsize, the maximum number of training steps, and the number of iterations per save weight parameter to prevent unexpected interruptions in training.

4.3 Metrics

There are many different evaluation systems for the efficiency of classification. In this paper, we use Precision to judge the efficiency and effectiveness of the results. Before calculation the first thing is to distinguish between true true (TP), false true (FP), and true false (FN). True-False (TN), False-False (FN).

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

4.4 Result Analysis

Using 3D convolutional neural network model to classify micro-videos can get a high training precision. This is mainly because convolutional neural networks can extract the features of training samples well and achieve good classification performance. The 3D convolutional network constructed in this paper significantly outperforms the original 3D convolutional network in terms of training precision, validationing precision and model convergence time. From the experimental results, it is clear that there is a difference between the precision of training data and the precision of validation data from the beginning to the end of the model training process, and the overfitting phenomenon of the network cannot be avoided even if BN and Dropout are used as regularization methods. There are several attempts to solve this problem. On the one hand, data augmentation can be used to increase the size of the training samples at the cost of significantly increasing the computational overhead. Another idea is to reduce the number of convolutional kernels between layers, which may sacrifice a certain amount of model precision.

When Dropout is 0, i.e., when the random deactivation method is not used, the difference between the model training precision and the validation precision is large. That is to say, there is a serious overfitting phenomenon and the generalization ability of the model is very poor. When the Dropout value is set to 0.5, the training precision and the validation precision are close to each other, and the best classification results are obtained. However, when the Dropout value was increased again, the training and validationing correctness of the model decreased significantly, because most of the neurons were deactivated and did not receive sufficient training, and the classification results could not be effectively predicted for micro-video samples.

5 Conclusion and Future Work

Along with the popularity of the Internet, people are exposed to more and more ways of micro- videos, and a huge amount of micro-video data has emerged. In

TABLE 2: Results comparison of diifferent dropout.

Dropout	0	0.25	0.5	0.75
trainning precision	0.953	0.952	0.971	0.944
validation precision	0.831	0.846	0.882	0.825

such a large amount of micro-video data, each person is interested in only a small part of it, and it will be a very difficult task to organize and manage these video data in an effective classification. micro- video classification algorithms, in general, are processed by obtaining the semantic related content of the video, such as analyzing and understanding human behavioral actions or other complex video clips, and finally classifying them into single or multiple categories automatically. At present, the algorithm of micro-video recognition based on deep learning has surpassed the iDT algorithm, making these traditional methods fade out of people's view. In the past, content understanding for short videos often used human work annotation, but in recent years, with the great success of deep convolutional neural networks in image recognition, short video content understanding based on this method has gradually developed. Nowadays, most of the recognition algorithms extract the feature representation of each frame independently and then fuse them. However, while extracting the feature representation, some low-level semantic features are lost, which makes the algorithm unable to accurately distinguish the category of the video. Therefore, in this paper, we proposed an efficient and accurate multi-modal fusion micro-video classification algorithm based on 3D convolutional neural network to further improve the classification accuracy. We design and implement a convolutional neural network based video classification. The network consists of an input layer, a convolutional layer, a pooling layer, a BN layer, a fully connected layer and an output layer. According to the micro-video classification task, a new network model is proposed to concat features of each modality into the overall features of various modalities through the network, and then fuse the various modal features with attention mechanism to obtain the whole micro-video features, which will be used for classification. In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

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] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of multimedia research*, 2017, pp. 3–29.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[3] J. C. Van Gemert, J. Geusebroek, C. J. Veenman, C. G. Snoek, and A. W. Smeulders, "Robust scene categorization by learning image statistics in context," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 105–105.

[4] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[7] 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.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[10] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 461–470.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[13] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[14] Y. Tan, Y. Hao, X. He, and X. Yang, "Selective dependency aggregation for action classification," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 592–601.

[15] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[16] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*. Springer, 2010, pp. 140–153.

[17] L. Nie, X. Song, and T.-S. Chua, "Learning from multiple social networks," *Synthesis lectures on information concepts, retrieval, and services*, vol. 8, no. 2, pp. 1–118, 2016.

[18] 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.

[19] 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.

[20] 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*, 2016, pp. 770–778.

[21] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 117–125.

 doi:10.20944/preprints202207.0308.v1

[22] Z. Tao, X. Wang, X. He, X. Huang, and T.-S. Chua, "Mgat: Multimodal graph attention network for recommendation," *Information Processing & Management*, vol. 57, no. 5, p. 102277, 2020.
[23] Z. Cheng, S. Jialie, and S. C. Hoi, "On effective personalized music retrieval by exploring online user behaviors," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 125–134.
[24] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli, "Mmalfm: Explainable recommendation by leveraging reviews and images," *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 2, pp. 1–28, 2019.
[25] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 975–985, 2012.
[26] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1398–1406.
[27] M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang, and S. Yan, "First-person daily activity recognition with manipulated object proposals and non-linear feature fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2946–2955, 2017.
[28] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition," in *Proceedings of the 10th international conference on Multimodal interfaces*, 2008, pp. 237–240.
[29] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.
[30] 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.
[31] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.
[32] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.
[33] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Research in Computer Vision*, vol. 2, no. 11, 2012.