Human Action Recognition Based on Vision Transformer and L2 Regularization

Qiliang Chen
School of Computer and Information Engineering, Xiamen University of Technology
643362009@qq.com

Hasiqidalatu Tang
School of Mathematics and Statistics, Xiamen University of Technology
2397727856@qq.com

Jiaxin Cai*
School of Mathematics and Statistics, Xiamen University of Technology
caijiaxin@xmut.edu.cn

ABSTRACT

In recent years, the field of human action recognition has been the focus of computer vision, and human action recognition has a good prospect in many fields, such as security state monitoring, behavior characteristics analysis and network video image restoration. In this paper, based on attention mechanism of human action recognition method is studied, in order to improve the model accuracy and efficiency in VIT network structure as the framework of feature extraction, because video data includes characteristics of time and space, so choose the space and time attention mechanism instead of the traditional convolution network for feature extraction, In addition, L2 weight attenuation regularization is introduced in model training to prevent the model from overfitting the training data. Through the test on the human action related dataset UCF101, it is found that the proposed model can effectively improve the recognition accuracy compared with other models.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**; • **Usability testing**;

KEYWORDS

Human action recognition, attention mechanism, L2 regularization, Transformer

ACM Reference Format:
Qiliang Chen, Hasiqidalatu Tang, and Jiaxin Cai. 2022. Human Action Recognition Based on Vision Transformer and L2 Regularization. In *2022 11th International Conference on Computing and Pattern Recognition (ICCPR) (ICCPR 2022), November 17–19, 2022, Beijing, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581807.3581840

1 INTRODUCTION

Human action recognition [1] uses a set of data obtained from a specific algorithmic sensor to identify human action types. This is a key and difficult point in many fields such as artificial intelligence in recent years. The topic of human motion recognition contains the core knowledge of many related frontier fields, such as deep learning and artificial intelligence technology. Therefore, the development of this professional field has great potential theoretical and academic research value. In addition, human action recognition has a good development prospect in many fields, such as security state monitoring [2], behavior characteristics analysis and network video image retrieval . Technology that recognizes human motion through video surveillance can intelligently detect abnormal behavior, such as fighting and illegal stalking. These activities may cause harm to personal safety, so the use of monitoring can be timely detection and early warning. In terms of sports, what human motion analysis technology [3] does is to compare the sports movements of athletes with those of standard athletes, so as to improve the accuracy of sports. That is to say, we can point out some mistakes in the training of the athletes' technical movements through analysis, and carry out some targeted training for the athletes' movements, so as to effectively improve the technical level of the athletes. As the above examples show, improving the speed and accuracy of human motion detection meets the higher requirements of human life. This is why the study of human motion recognition technology is so valuable. In 2014, Zisserman et al. [4] proposed a convolutional neural network with two kinds of flows. In 2015, Du et al. proposed the concept of 3D convolutional neural networks. Pinz et al. [5] proposed a spatio-temporal fusion architecture in 2016 to solve the problem that the original dual-stream network structure could not interact between two information flows. The fundamental role of attention mechanism is to improve the learning performance of neural network models. The attention mechanism was originally proposed by K. Cho et al. [6] in 2014 and applied to machine learning. Wang Jianqiang et al. [7] proposed the introduction of high-order attention mechanism in human action recognition model. This method solves some problems, such as ignoring the interaction relationship between different video features and the unsatisfactory effect of analyzing approximate video action recognition. In 2018, Seyma Yucer et al. [8] proposed a 3D Human Action Recognition with Siamese-LSTM Based Deep Metric Learning. Naresh Kumar et al [9] proposed Motion Trajectory for Human Action Recognition Using Fourier Temporal Features of Skeleton in 2018 Joints are implemented by treating human movement as the trajectory of bone joints. Tasweer Ahmad et al [10] proposed Using Discrete Cosine Transform Based Features for Human Action Recognition in 2015 to calculate the sequence of sports historical images, Then the motion history image is processed by block - based truncated discrete cosine transform. Muhammad Hassan et

*Corresponding author.

ICCPR 2022, November 17–19, 2022, Beijing, China

ACM ISBN 978-1-4503-9705-6/22/11…$15.00
https://doi.org/10.1145/3581807.3581840

al [11] proposed A Review on Human Actions Recognition Using Vision Based Techniques in 2014 The main aim is to use various vision-based techniques to identify different human behaviors.

This paper takes VIT network structure as the framework of feature extraction, and proposes L2 regularization based on spatio-temporal attention to construct the model. L2 regularization can solve the problem of model overfitting and improve the accuracy of the model. The model is tested on UCF101 dataset. Compared with other models on UCF101 dataset, it is finally concluded that the proposed model effectively improves the recognition accuracy on UCF101 dataset.

2 METHOD

The dataset used for human action recognition here is UCF101 dataset. In the study of human motion recognition, feature extraction is a method of feature location and extraction of vector video image data. Deep learning is used here to extract features. There are two methods. One is based on dual-stream convolutional network. The principle is that the image data of a single RGB frame [12] is used as the data input of the single-frame space and AC neural network, while the multi-frame time and AC neural network uses the optical flow reflection image data of multiple frames stacked as the network input. Then, the convolution neural network is used to comprehensively process the human action information recognition with different inputs of the two neural networks. Finally, the human action information recognition can be realized through the analysis results obtained by the network. The other one is based on convolutional neural network (CNN). Convolutional neural network system consists of input layer, convolution layer, pooling layer, fully connected layer and output layer, which is based on convolution operation. Convolutional layer is an important part of convolutional network. The input of the input layer is convolved to obtain the feature information. Each convolution layer structure contains multiple convolution kernels, and each convolution kernel computes the input of the convolution layer. With the change of convolution kernel parameters, the extracted feature information will also change. Assuming that the size of the input feature map is $h_i{\times}w_i$, the size of the convolution kernel is $h_k{\times}w_k$, the convolution step in the height direction and width direction are respectively s_h and s_k, and the edge of the feature map is 0 filled, then the size of the output feature map after convolution operation is $h_o \times w_o$. Where h_o and w_o respectively represent the height and width of the output feature map after convolution operation, and the formula is as follows:

$$h_o = (h_i - h_k + 2 \times pad) \div s_h + 1 \tag{1}$$

$$w_o = (w_i - h_w + 2 \times pad) \div s_w + 1 \tag{2}$$

In this paper, L2 regular joint attention mechanism is used to analyze human action recognition. L2 regularization is the most commonly used regularization technique in deep learning. L2 regularization is to solve the problem of overfitting caused by too many features. The solution can reduce the weight of features or penalize the weight of unimportant features. Regularization can be used to prevent the fitting and improve the generalization ability.

2.1 UCF101 dataset

UCF101 dataset [13] extracts action-related videos from YouTube. The UCF101 dataset contains 13,320 videos with a total duration of 27 hours. Video datasets can be divided into five categories based on human activities. Each category can be divided into 25 groups, each containing at least 4-7 videos of varying duration simultaneously.

2.2 Human action recognition model architecture based on attention mechanism combined with L2 regularization

In this paper, the VIT network [14] structure is selected as the framework for feature extraction, and the spatiotemporal attention mechanism proposed by Facebook AI [15] in 2021 [16] is used to replace the traditional convolutional network. This time-space transformer structure is referred to as TimeSformer. TimeSformer extracts spatiotemporal features from a series of frame-level images, Video Task Adaptation, And they have faster reasoning speed to do it. Figure 1 shows the VIT structure.

As can be seen from Figure 1, each frame image is segmented into 9 patchs blocks, which are flattened in sequence as input sequences. After processing by linear mapping layer, the image is added together with position coding and input into Transformer Encoder blocks (N). Each Transformer encoder is composed of two sub-blocks. The first sub-block consists of layer normalized Norm and multi-head Attention, plus hop connection. The second subblock consists of a layer normalized Norm with a multilayer perceptron (MLP) and a hop connection.

Calculation process of self-attention mechanism: The Transformer structure includes L coding modules. For each module L, a query/key/value (attention mechanism) will be calculated by the following equations 3), (4) and (5) :

$$q_{(p,t)}^{(l,a)} = W_Q^{(l,a)} LN\left(z_{(p,t)}^{(l-1)}\right) R^{D_h} \tag{3}$$

$$k_{(p,t)}^{(l,a)} = W_K^{(l,a)} LN\left(z_{(p,t)}^{(l-1)}\right) R^{D_h} \tag{4}$$

$$v_{(p,t)}^{(l,a)} = W_V^{(l,a)} LN\left(z_{(p,t)}^{(l-1)}\right) R^{D_h} \tag{5}$$

Where, LN () represents layer normalization, and a=1.. A represents the index of multiple attention heads, A represents the total number of attention heads, p=1,2... ,N denotes the spatial position, t=1,2... ,F represents the index of the frame.

For video, each frame has a time attribute, so it needs time attention to extract features. For each frame of image, it also has a space attribute, so it needs to design space attention to extract features. Therefore, spatio-temporal attention mechanism can better extract richer features for video.temporal and spatial attention, S for space, T for time, We're going to block each image (p, t) Compared with other image blocks in the same spatial location but with different time frames, Get the weight matrix$\alpha_{(p,t)}^{(l,a)time}$,As shown in Formula (6) :

$$\alpha_{(p,t)}^{(l,a)time} = SM\left(\frac{{q_{(p,t)}^{(l,a)}}^T}{\sqrt{D_h}} \bullet \left[k_{(0,0)}^{(l,a)} \left\{k_{(p',t')}^{(l,a)}\right\}\right]\right) \tag{6}$$

Where, $t' = 1, 2, \ldots, F, p^{'} = 1, 2, \ldots, N$,SM indicates the softmax function.

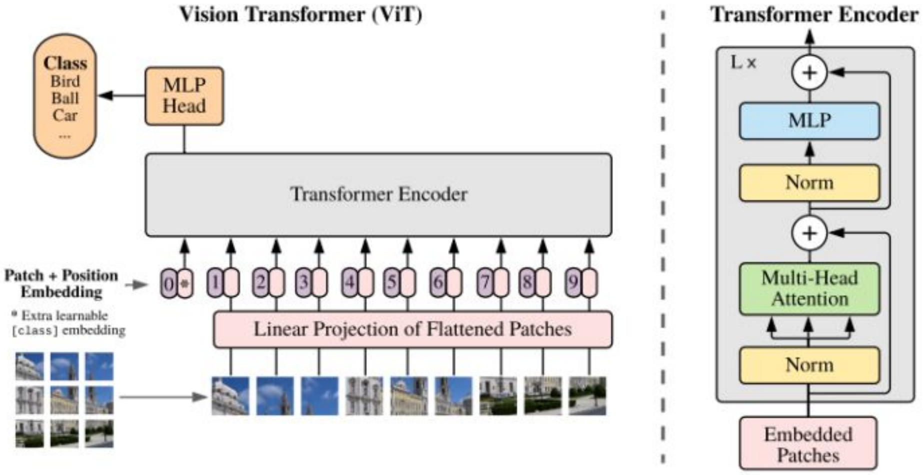

Figure 1: ViT architecture of backbone network [17]

Experimental results show that compared with the spatio-temporal joint attention module (ST), which requires (NF+1) comparisons for each image block, the spatio-temporal separated attention module (T+S) only needs (N+F+2) comparisons for each image block, which greatly improves the computational efficiency and achieves better classification accuracy.

The spatio-temporal attention mechanism is shown in Figure 2

As can be seen from Figure 2, the output of the previous layer$z^{(\ell-1)}$ after a Time Att, Then the sum of a hop connection (residual mode), and then the sum of Space attention and itself, Finally, after a multi-layer perceptron (MLP) and a jump-link summing operation, I get z(l) output.

We will train the model with L2 weight decay regularization on TimeSformer model, which can prevent the model from overfitting the training data and improve the generalization ability of the model. L2 regularization refers to the application of L2 norm in deep learning. Specifically, L2 norm is added to the loss function loss to achieve the function of parameter punishment, that is, the regularization effect is realized. The loss function adopts the cross-entropy loss function: nn.cross_entropy(input,target), The formula is as follows:

$$loss\left(x, class\right) = -\log\left(\frac{\exp\left(x\left[class\right]\right)}{\sum_i x\left[i\right]}\right) = -x\left[class\right]+\log(\sum_i \exp\left(x\left[i\right]\right)) \tag{7}$$

In the specific implementation, the calculation formula of L2 loss function and weight decay regularization is shown in (8) and (9) :

$$Loss = 0.5 * reg\_coeff * reduce\_sum\left(square\left(x\right)\right) \tag{8}$$

$$L2WeightDecay = reg\_coeff * parameter \tag{9}$$

Where, reg_coeff is regularization_coeff regularization coefficient, the default value is 0.0, generally set to 0.1.

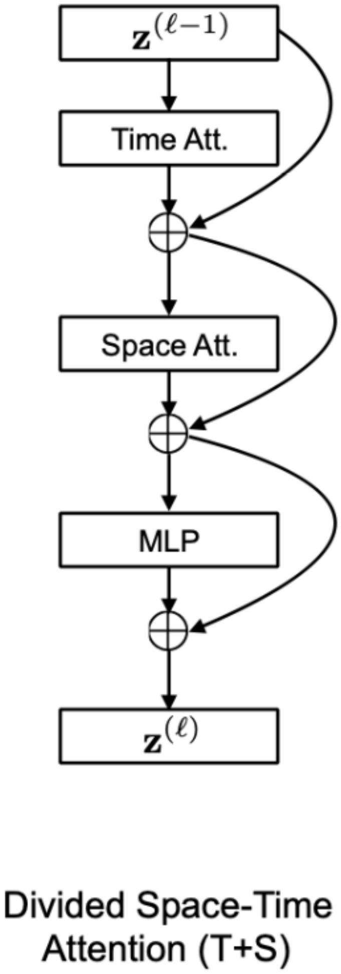

Figure 2: Separable spatio-temporal attention mechanism

2.3 L2 regularization

Regularized regression is a form of linear regression that limits, corrects, or reduces the probability that the coefficient estimate tends to zero. In other words, regularization can reduce the computational load of model operation and the non-uniform stability generated in the process of model learning, which can effectively eliminate the risk of dropping overfitting. Regularization function refers to the original loss function, on the basis of adding some other regularization terms, can also be called the complexity of the model penalty function term.

The regularization expression is shown in Equation 10) :

$$\tilde{J}\left(w;X,y\right) = J\left(w;X,y\right) + \alpha\Omega\left(w\right) \tag{10}$$

Where, X and y are the sample sets of the training model and their corresponding labels, W is the vector of weight coefficients, J is the objective function of the training model, and Ω(w) is a penalty term, which is a measure of the "size" parameter of the training model, and the parameter α determines the intensity of regularization. Different ω functions have different preferences for the optimal solution with weight w, so it is likely to produce different regularization effects of functions. Here, formula (11) is L2 regular, as follows:

$$l_2 : \Omega\left(w\right) = w_2^2 = \sum_i w_i^2 \tag{11}$$

When the model becomes very complex in order to overfit the training set, overfitting is prone to occur. At this time, the model fits the training data very well, but loses generality and tends to remember rather than learn the features of the data, which leads to poor performance of the model on the new data, that is, poor generalization ability. If overfitting occurs, the model will learn and apply the random noise of the training set as valid data, so the performance of the model on the new data will degrade.

There are too many parameters in the process of network learning, which leads to the complexity of the model process and the easy problem of model overfitting (overfitting: the model process performs well in the network training test sample dataset, but it performs poorly in the actual training test sample set and has no good universality). In this work, the most widely used regularization method, L2 regularization, was chosen to reduce overfitting and improve the generality of the model. Regularization ensures that the parameters are not overfitted and reduces the risk of overfitting. The larger the parameter α in the regularization term L2, the smaller the value of the weight matrix of the model. There will be a lot of hidden items with very small weights in the model, and the effect of these hidden items in the model will be very small. This is equivalent to reducing the neural network to a smaller network, but the depth of the network remains the same, and the model thus changes from an overfitting state to an underfitting state.

3 RESULTS

The experiment of human action recognition starts with data preprocessing, part of the video needs to be decoded into frames, and then a segment of the video is sampled, followed by normalization, data format conversion and so on. Figure 3 shows the specific process:

As can be seen from Figure 3, the first step of data processing is to randomly segment the video, then randomly select the starting position of each segment, and then extract one frame every three from the starting position of each segment for K consecutive times to complete the sampling of k frames of each segment. For each sampled Image, according to the mean and variance of its Image channel, the normalization process is carried out. Finally, the Image format is converted from PIL.Image format to Numpy array format to complete the data format conversion.

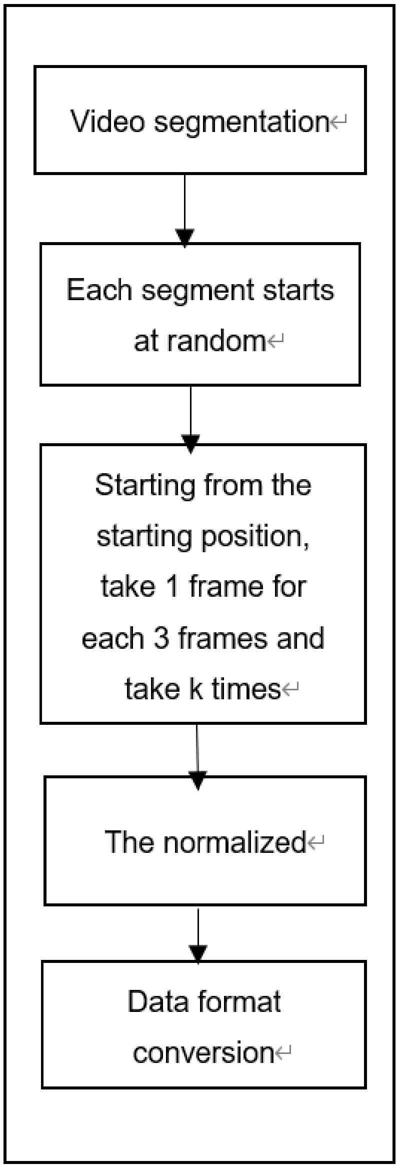
Video segmentation↵
Each segment starts at random↵
Starting from the starting position, take 1 frame for each 3 frames and take k times↵
The normalized↵
Data format conversion↵

Figure 3: Data processing process

- Machine and parameter configuration: GPU training machine configuration is 1080, Python, CPU is I5.
- Data correlation: the image size is (IMG_size) 224*224, the size of each segmented block is 16*16, the number of input channel layers is 3, the embedding vector dimension of each block is 768, the depth of network layers is 12, and (num_heads) is 12.

Table 1: Accuracy of each method on the UCF101 dataset (%)

All kinds of methods	UCF101DS
Two Stream [18]	88.0
ResNet3D [19]	86.1
DensNet 3D [20]	88.9
TimeSformers+L2	95.8
TimeSformers	92.4

In the experiment, we randomly selected 70% of the UCF101 dataset as the training set and 30% as the test set. As shown in Table 1

According to Table 1, it is found that the accuracy of this model is slightly improved in UCF101 compared with other models.

4 CONCLUSION

With the development of science and technology, artificial intelligence is gradually improving the quality of human life. When writing the paper, I looked up many domestic and foreign literature materials and found that many scholars at home and abroad have quite rich research in this field. It is because of their research results that I have a quick and detailed understanding of this field and write a paper.

This paper introduces the research background, in the field of human action recognition and UCF101 dataset is introduced, based on attention mechanism is through the establishment of joint L2 regular human action recognition model, and the model in human action recognition is carried out on the common dataset UCF101 dataset the experiment with the other traditional human action recognition model in the model were compared, It is found that the accuracy of this model is slightly improved compared with other models.

ACKNOWLEDGMENTS

This work is supported by the Fujian Natural Science Foundation Project Evaluation of Cerebrospinal Deformities by Level III Ultrasound Examination in Middle and Late Pregnancy Based on Deep Learning, the Fujian Natural Science Foundation Project 2022J011396, the China University Industry-University-Research Innovation Fund 2020HYA04002 and the Fujian Province Social Science Foundation Project FJ2021B128.

REFERENCES

[1] Hu Qiong, Qin Lei, Huang Qingming. Overview of Human Action Recognition Based on Vision. Chinese Journal of Computers, 2013, Vol.12(12):2512-2524.

[2] Fu Bin, Fu Xin, Cui Jianguo.Human Pose Recognition Method for Elderly Assistance Mechanism Based on MEMS Sensor. Journal of Harbin University of Commerce (Natural Science Edition),2021,37(05):590-594.

[3] Cao Shumin. Human Action Recognition and Interaction Based on Intelligent Wearable Device. Anhui: University of Science and Technology of China, China, 2012 (in Chinese) 2020.

[4] Simonyan K Zisserman A. Two-stream convolutional networks for action recognition in videos. https://arxiv.org/pdf/1406.2199.pdf

[5] Feichtenhofer C,Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. https://arxiv.org/pdf/1604.06573.pdf

[6] K. Cho, B. Van Merrienboer, D. Bahdanau *et al.* Learning phrase representations using RNN encoder decoder for statistical machine translation.in EMNLP, ACL, 2014,1724–1734.

[7] Wang Zengqiang, Zhang Wenqiang, Zhang Liang. Human behavior recognition by introducing high-order attention mechanism. Signal processing, 2020,36 (08) :1272-1279.

[8] Seyma Yucer and Yusuf Sinan Akgul, 3D Human Action Recognition with Siamese-LSTM Based Deep Metric Learning. Journal of Image and Graphics, 2018, pp. 21-26.

[9] Naresh Kumar and Nagarajan Sukavanam, Motion Trajectory for Human Action Recognition Using Fourier Temporal Features of Skeleton Joints, Journal of Image and Graphics, 2018, pp. 174-180.

[10] Tasweer Ahmad, Junaid Rafique, Hassam Muazzam, and Tahir Rizvi, Using Discrete Cosine Transform Based Features for Human Action Recognition, Journal of Image and Graphics, 2015, pp. 96-101.

[11] Muhammad Hassan, Tasweer Ahmad, Nudrat Liaqat, Ali Farooq, Syed Asghar Ali, and Syed Rizwan hassan, A Review on Human Actions Recognition Using Vision Based Techniques, Journal of Image and Graphics, 2014, pp. 28-32.

[12] Ye Qing,Tan Zexian,Qu Chang,Zhang Li. Human motion recognition using three-dimensional skeleton model based on RGBD vision system. Journal of Physics: Conference Series,2021,1754(1).

[13] Oomro K, Zamira R, Shah M. Ucf101:a dataset of 101 human actions classes from videos in the wild. [2020-08-10] https://arxiv. org/pdf/1212. 0402. pdf

[14] Kuehne H, Jhuang H, Stiefelhagen R, et al. HMDB:a large video database for human motion recognition. Proceedings of International Conference on High Performance Computing in Science and Engineering. Berlin, Germany:Springer, 2013.

[15] Tran D, Bourdev L, Fergus R, et al. Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. Greece:IEEE,2020.

[16] Zhang Yu. Research on Human Action Recognition Method Based on Deep Learning. Beijing: Beijing University of Civil Engineering and Architecture,2021.

[17] Gedas Bertasius ,Heng Wang , Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding. https://arxiv.org/pdf/2102.05095.pdf

[18] Si C Y, Jing Y, Wang W, *et al.* Skeleton-based action recognition with spatial reasoning and temporal stack learning. Guang Zhou:ICDIP,2019.

[19] Zhang P F, Lan C L, Xing J L, *et al.* View adaptive neural networks for high performance skeleton-based human action recognition. Xi An:IEEE,2019

[20] Shou Z, Lin X, Kalantidis Y, et al. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. Greece:IEEE,2019.