Multi-classification of Human Action Based on ResNet

Shuaidong Ji*

Department of International Studies, Xiamen University of Technology, No. 600, Ligong Road, Jimei District, Xiamen, China, 361024

*Corresponding author: chengchang.wang@gecacademy.cn

Abstract. With the development of machine learning and other related technologies, more and more excellent research on human action recognition and classification has been proposed, which significantly promotes the application of this technology in the actual situation. This paper mainly focuses on the characteristics of ResNet18, ResNet50, ResNet101, and ResNet152 in 15 human action recognition and classification tasks respectively. First, adjust and enhance the sample images of all training sets and test sets, and adjust the overall parameters according to the characteristics of the input image, so that all images can be normalized and high quality can be guaranteed at the same time; Then use the above four ResNet models to study and test with an epoch of 200, and explore the characteristics of the four models in this project by comparing the accuracy, loss, confusion matrix graphic, F1-score, and the number of epoch experienced in reaching the steady state of accuracy. The results show that in the overall effect, with the increase of the number of layers of the ResNet models, the accuracy changes in a positive correlation, and the epoch number experienced by the accuracy reaching the steady state changes in a negative correlation.

Keywords: Human Action classification, machine learning, ResNet.

1. Introduction

Human Action classification (HAC), which requires computers to provide the right labels to pictures of actions after analyzing their features automatically, in the present era, has inevitably been a crucial component of various fields, e.g., athlete action evaluation [1], proctoring of examinations [2], sign language recognition based on action classification and shopping mall behavior detection [3, 4], etc. Action identification and classification, on the other hand, is a challenging question due to diverse behaviors and complex surroundings.

Recently, with the increasing popularity of machine learning, the research of HAC based on machine learning models is emerging in endlessly. Liu et al. combined space-time features with a hyper-sphere multi-class Support Vector Machine (HS-MC SVM) to complete the classification and recognition of 10 classes of actions [5]. K-Nearest Neighbor (KNN) Algorithm was used by Wang et al. and concluded that the final accuracy would be relatively high if using the appropriate value of k [6]. An improved Random Forest model named Action Forest was provided by Chuan et al. which could enhance the accuracy of classification by reducing the negative effect on the environment [7].

However, although some research was based on Convolutional Neural Networks (CNN), for instance, Ijjina and Mohan designed their own convolutional neural network to learn key features of videos of human actions automatically [8]. In addition, Yang et al. improving the accuracy of human action recognition by combining different features extraction method with CNN [9], prior work is limited to paying attention to ResNet model, which is one of the classical and excellent models of CNN, and how to enhance the performance of ResNet model on dealing with HAC without adding any other features extracting methods. Moreover, regard with multi-behaviors and complex environments, more action categories and more random backgrounds deserve more in-depth exploration.

This paper proposes an approach for classifying human action pictures into 15 different classes accurately, including eating, using laptop and smiling, etc. based on ResNet model. In real life, people's actions are often accompanied by many uncertainties. For example, people can drink water with a cup, a straw, or even directly hold the water in the lake with both hands for drinking but all

these different actions belong to only one class that is ‘drinking’, this requires that CNN model can obtain more features in the pictures when completing multi classification tasks, in other words, learn more "details". With the increase of the number of learning layers, ResNet can ensure relatively lower gradient, which enables the model to obtain more features while ensuring higher operation speed [10], as a consequence, there are several main points that focused by this article:

(1) Ensure that the pictures of each type of action in the training set and testing set cover as wide a range of styles as possible and as complex the surrounding environment as possible, such as single person scenes, multi person scenes, close range, long range and mixed actions, etc. The purpose is to ensure that the trained model can be applied to the situation that more closely fits the real-life scene.

(2) Improving the accuracy while reducing the loss value of the ResNet model used in this classification task from the aspects of image enhancement, parameter selection and the layers number increasing.

2. Method

2.1. Dataset description and preprocessing

In this project, the dataset was got from Kaggle [11]. The original data set consists of two parts, one is a training set containing 12600 pictures of human behavior, and the other is a test set containing 1700 pictures of human behavior. These pictures belong to 15 action categories respectively, which are " calling ", " clapping ", " cycling ", " dancing ", " drinking ", " eating ", " fighting ", " hugging ", " laughing ", " listening_to_music ", " running ", " sitting ", " sleeping ", " texting " and " using_laptop ". For the test set, according to the one-to-one correspondence between the picture name and the action label in the ".csv" file provided by the website, the pictures are stored in 15 files named by the action category name, and each folder contains 840 pictures on average; For the training set, first, classify the pictures according to their own judgment of the behavior in the pictures, and then store them in 15 folders corresponding to the action names, on average, each folder contains 110 pictures. These pictures are of different sizes and are RGB pictures. The samples of each classes are illustrated in Figure 1.

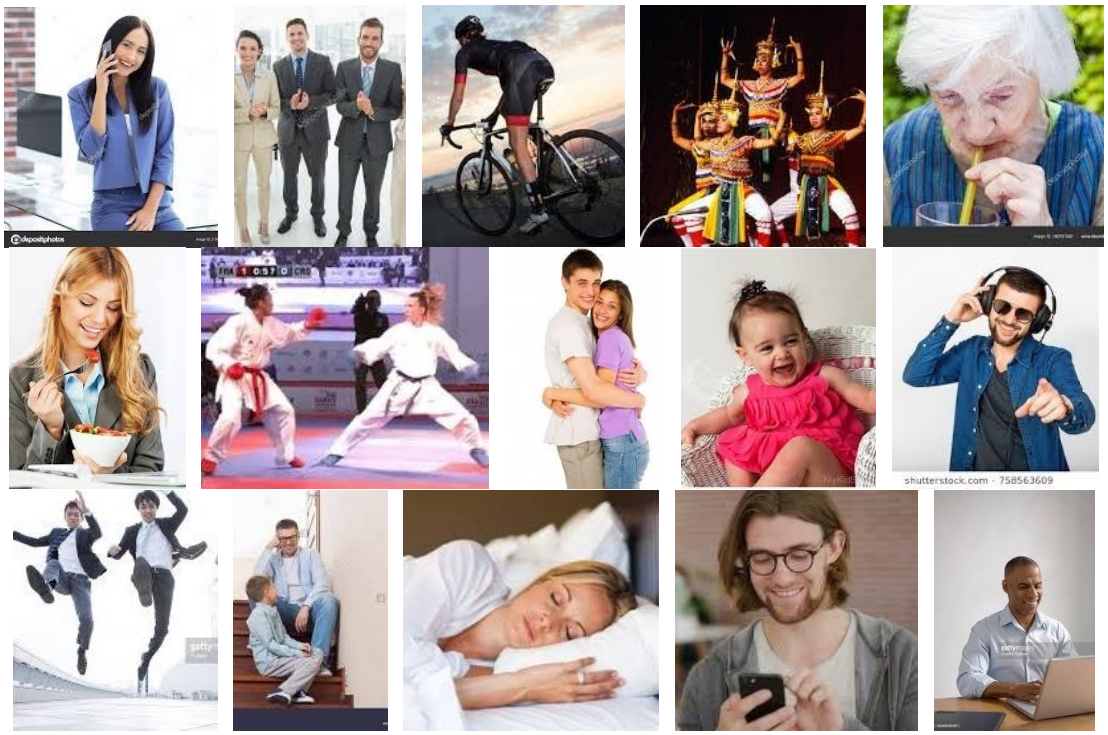


Figure 1. The visualization of pictures in each of 15 classes

There are three aspects done during preprocessing for both pictures in the training set and testing set. First, set the size of pictures to 224×244 which is the size that was restricted by the ResNet50, ResNet101 and ResNet152 model on input images. Second, set the brightness, contrast, and saturation of the input pictures in order to augment the input pictures and reduce the effect of the blur of pictures on the accuracy. Third, normalizing the pictures into the same value of mean and standard deviation respectively which was calculated according to the pictures in the training set. augmentation. The details of the parameter are shown in Table 1.

Table 1. The details of the parameter of preprocessing

Name of parameters	Value
size	(224,224)
brightness	0.5
contrast	0.5
saturation	0.5
mean	[0.5728639, 0.53787696, 0.50688255]
Standard deviation	[0.2453927, 0.24324809, 0.24655288]

2.2. ResNet model

Deep Residual Network (ResNet) is one of the best CNN models which aims to prevent vanishing/exploding gradients with the network depth increasing, which is applied in many tasks [12, 13]. and Figure 2 illustrates the operation principle of a single building block of ResNet, where x means final output of the last building block while F(x) means features extracted by this block, and F(x) + x represents the output of this block [10].

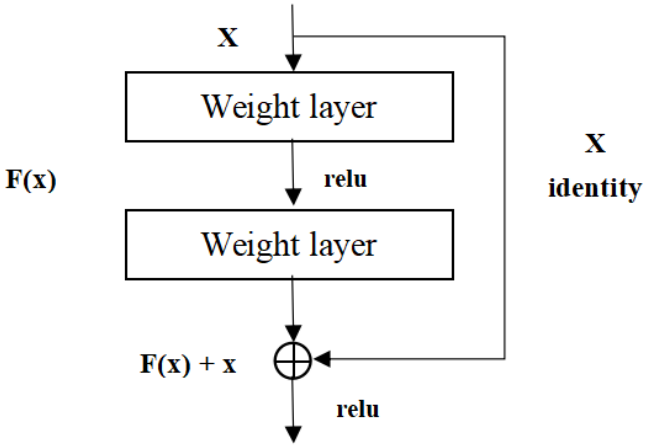


Figure 2. Residual learning: a building block

In this project, ResNet series models play an important role and also show excellent performance. Table 2 shows the structure and brief information of three models mainly used in this project: ResNet18, ResNet50, ResNet101, and ResNet152 [10].

Table 2. The structure and brief information of three models mainly used in this project: ResNet18, ResNet50, ResNet101 and ResNet152

layer name	output size	18-layer	50-layer	101layer	152layer
conv1	112×112	7×7, 64, stride 2			
conv2	56×56	3×3 max pool, stride 2			
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times 3$
conv3	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times 8$
conv4	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times 23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times 36$
conv5	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times 3$
	1×1	average pool, 1000-d fc, softmax			
FLOPs		1.8×10^9	3.8×10^9	7.6×10^9	11.3×10^9

(1). ResNet18 model

ResNet18 model is the lightest one among residual network series which is built up with five convolutional stages and that can be found in Table 2. Based on that structure, the input images will be down-sampled five times, and then 512(7×7) features maps will be produced through the final average pooling. 1.8×10^9features will be extracted after the whole process.

(2). ResNet50 model

ResNet50 model is constructed by replacing each 2-layer block in the 18-layer net with this 3-layer bottleneck block. This model can produce 3.8×10^9 FLOPs which is approximately twice that produced by ResNet18, as a result, more "detail" features will be learned by machine if the layers of residual networks increased.

(3). ResNet101 and ResNet152 model

Similarly, ResNet101 and ResNet152 model can be constructed by using more 3-layer blocks. However, with the significant increase that can be found in the depth when changing the model from 50-layer to 101-layer or 152-layer, it can still remain relatively low complexity.

2.3. Implementation details

The project utilized the popular deep learning framework PyTorch through GPU which could enhance the speed of human action classification. One of the most important targets of this paper is to compare the accuracy and loss of the result of classification by using ResNet18, ResNet50, ResNet101, and ResNet152 respectively, as a consequence, besides adjusting appropriate parameters and hyperparameters of each ResNet model, the other parameters, including initial learning rate, batch size, optimizer, loss function of train, loss function of test, and epochs, should remain the same. The value of those 6 parameters illustrated in Table 3.

Table 3. The value of 6 parameters which remained the same when using 3 models

Name of parameters	Value
initial learning rate	1e-4
batch size	512
optimizer	Adam
loss function of train	Label Smoothing Cross Entropy
loss function of test	Cross Entropy
epochs	200

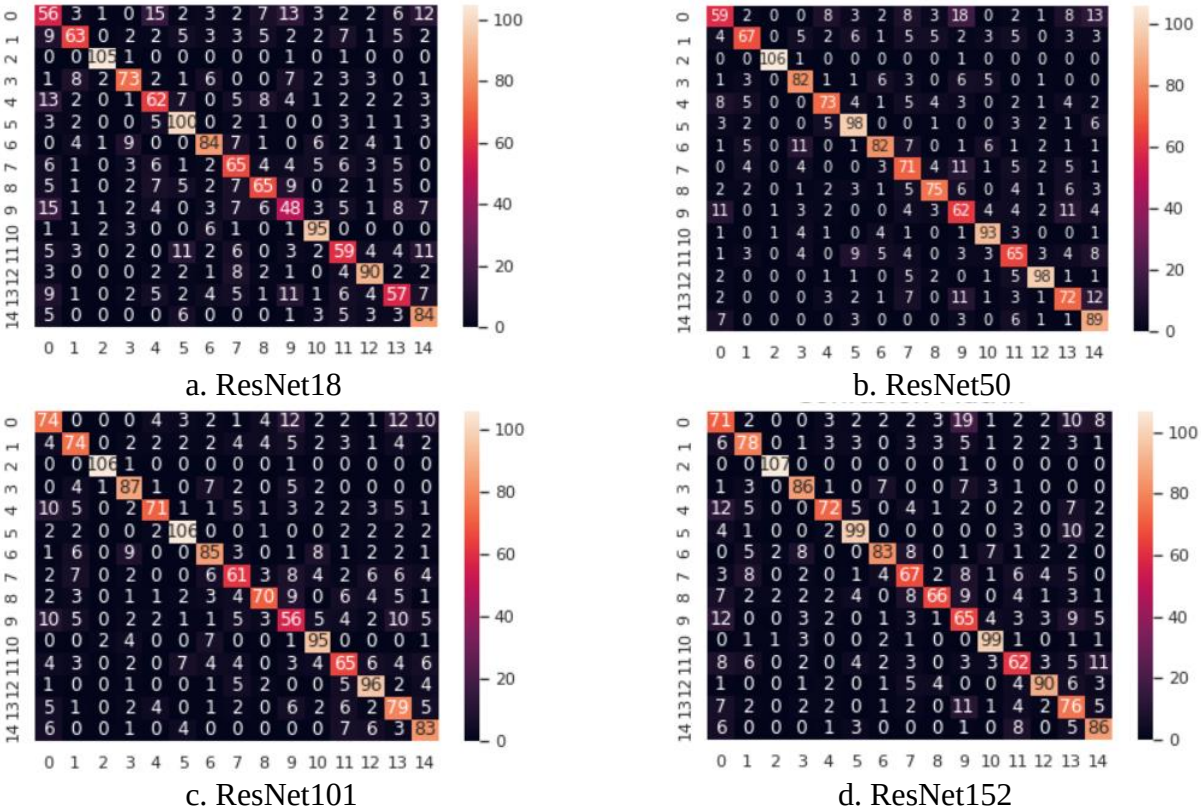
Finally, precision, recall, and f1_score was used to illustrate the final result of each 15 actions respectively. Moreover, confusion matrix was used to show the performance of an algorithm, the abscissa of confusion matrix demonstrates the actual class while Ordinate shows the predictions of the model [14].

3. Results and disscussion

3.1. Classification performance of various ResNet models

Figure 3 shows the 15 classification results of each model with confusion matrices. The numbers 0 to 14 in the abscissa and ordinate of the confusion matrix corresponded to the 15 action names of this project respectively. It can be seen from the four figures that the color on the diagonal was more evident than that in other parts of all these 4 confusion matrices, indicating that the overall effect of the four models was good.

Meanwhile, it can be found through comparison that the overall classification effect of the model was improved with the increase of the number of model layers. For instance, the correct classification numbers of 101-layers and 152-layers models for type "0" were 74 and 71 respectively, which were higher than the correct classification numbers of 18-layers and 50-layers models in the same type, which were 56 and 59 respectively. However, it cannot simply illustrate that this growth can be reflected in every category. For example, although the number of layers of ResNet152 was about 3 times that of ResNet50, the correct classification of pictures in category "12" by Resnet50 was 96, which was more than that of ResNet152 in the same category, which was 90.



a. ResNet18 b. ResNet50

c. ResNet101 d. ResNet152

Figure 3. The final result with Confusion Matrix of 4 models

Table 4 displays the accuracy and loss of each model after 200 epochs. It can be seen that with the increase in the number of model layers, both training and testing accuracy were increased, and the loss of both training and testing dropped. Moreover, since the training accuracy was lower than the testing accuracy, no model had an overfitting issue in the final findings.

Table 4. The final result with accuracy and loss of 4 models

Model	Training accuracy	Training loss	Testing accuracy	Testing loss
ResNet18	60.23%	1.5810	65.14%	1.2140
ResNet50	65.58%	1.4556	69.95%	1.0570
ResNet101	66.54%	1.4324	70.89%	1.0078
ResNet152	68.38%	1.3983	71.54%	0.9728

Figure 4 shows the F1-score of four models in 15 categories. Although the F1-score of four models in individual categories, such as "cycling", was higher than that of other categories, the reason was that the number of pictures in category "cycling" was slightly less than that of other categories, and through observation, it can be found that the F1-score of four models in each category was relatively stable. The aforementioned findings demonstrate that the experimental process's parameter adjustments were generally appropriate.

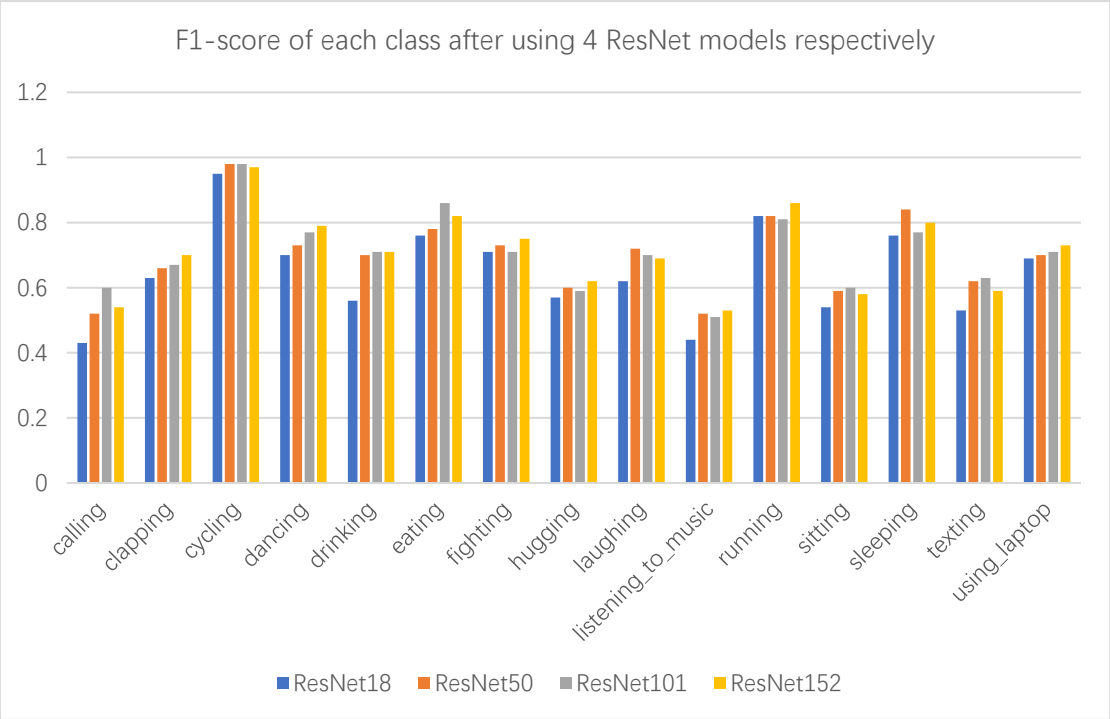


Figure 4. F1-score of each class after using 4 ResNet models respectively

3.2. Fitting trend of ResNet models

Figure 5 demonstrates the comparation of fitting trend of 4 ResNet models and the abscissa represents epoch and the ordinate represents accuracy. It can be seen from the figure that the accuracy of all four models first experienced a relatively large increase, and then reached the steady state for a long time. After reaching the steady state, although the accuracy would increase slightly with the increase of epoch, it was not obvious. The appearance of this phenomenon indirectly proves the structural superiority of ResNet model, which has been shown by figure2, simply put, the output of each building block is the output of the previous building block plus the newly learned content. As a consequence, when there is nothing to be “new” for model to learn, it will reach a steady state, compared with the gradually increasing accuracy of other models, ResNet model will more efficiently inform users of the final results.

And the relationship between the number of epochs experienced by the four models to reach the steady state was: resnet152 < resnet101 < resnet50 < resnet18, so the larger the number of layers of the model, the “more quickly” it will reach the steady state.

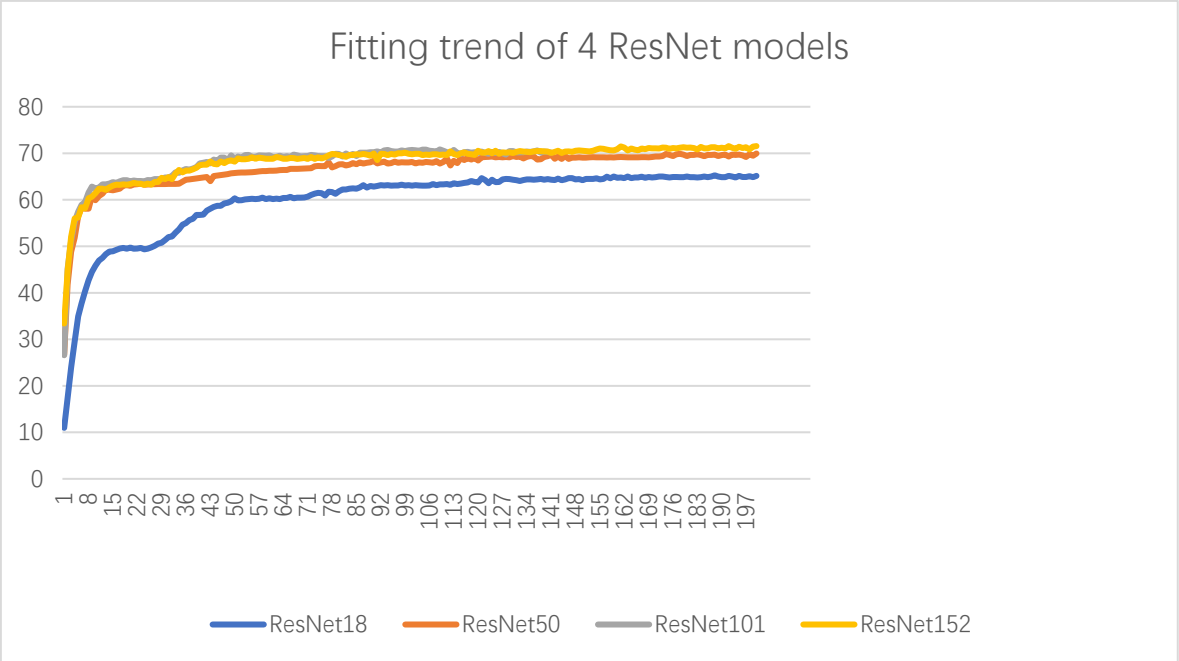


Figure 5. Fitting trend of 4 ResNet models

4. Conclusion

In this project, ResNet18, ResNet50, ResNet101, and ResNet152 are used to complete the classification and recognition of 15 human actions in this project respectively. Among them, as the number of layers of RESNET model increases from 18 to 152, the accuracy increases from 65.14% to 71.54%, and the loss decreases from 1.214 to 0.9728. At the same time, the more layers of ResNet model, the less epochs it takes to achieve the steady state of accuracy. In the future, the impact of multi-scale on the accuracy of ResNet series models in multi classification projects will be explored.

References

[1] Lin, C. -H., et al. "A Lightweight Fine-Grained Action Recognition Network for Basketball Foul Detection," 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 2021.

[2] Mukhanbet, A. A., et al. "Hybrid Architecture of Face and Action Recognition Systems for Proctoring on a Graphic Processor," 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), 2021.

[3] Sarhan, N., et al. "Transfer Learning for Videos: From Action Recognition To Sign Language Recognition," 2020 IEEE International Conference on Image Processing (ICIP), 2020.

[4] Widyadhana, I. H., et al. "Development of Video Based Action Recognition System For Item Taking From Shelf," 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2021.

[5] Liu, J., et al. "Action Recognition by Multiple Features and Hyper-Sphere Multi-class SVM," 2010 20th International Conference on Pattern Recognition, 2010.

[6] Wang, P., et al. "Application of K-Nearest Neighbor (KNN) Algorithm for Human Action Recognition," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2021.

[7] Chuan, C., et al. "Human Action Recognition Based on Action Forests Model Using Kinect Camera," 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2016.

[8] Ijjina, E. P., et al. "Human Action Recognition Based on Recognition of Linear Patterns in Action Bank Features Using Convolutional Neural Networks," 2014 13th International Conference on Machine Learning and Applications, 2014.

[9] Yang, Y., et al. "Human Action Recognition Based on Skeleton and Convolutional Neural Network," 2019 Photonics & Electromagnetics Research Symposium - Fall (PIERS - Fall), 2019.

[10] He, K., Zhang, X., Ren, S., et al. "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[11] Kaggle, "Human Action Recognition (HAR) Dataset," 2022, https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset.

[12] Yu, Q., et al. "Semantic segmentation of intracranial hemorrhages in head CT scans." 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019.

[13] Farooq, M., et al. "Covid-resnet: A deep learning framework for screening of covid19 from radiographs." arXiv preprint arXiv:2003.14395, 2020.

[14] Ariza-Lopez, F. J., et al. "Complete Control of an Observed Confusion Matrix," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018.