Journal of Electronic Research and Application

Research Article

Unattended Video Classifying System based on Transfer Leaning

Yarong Li*

The Experimental High School Attached To Beijing Normal University

Publication date: September, 2019

Publication online: 30 September, 2019

****Corresponding author:*** Yarong Li, liaoquanneng@ivygate.cn

1 Introduction

As the internet techniques advances, the demand of entertainment of the general public increases rapidly. A majority of short video applications like "TikTok" have appeared. The unbelievable growth of the number of the new videos and the increase of the average wage makes it almost impossible to identify the categories of the videos manually. In order to cater to the users' preferences efficiently, detecting the content of the videos is inevitable. Fortunately, due to the great development on both the algorithm and hardware, the golden era of artificial intelligence is coming which makes it is possible to use computers to recognize the content of the videos. Comparing to the traditional method of content recognition, machine learning increases the efficiency with low cost, which is a better way to cope with the conflicts between the great increase of the number of the new videos uploaded by users and the lack of human resources.

In this paper, we use Pytorch, which is a tool that is widely used for deep learning. Pre-trained models of 'Resnet' are used for transform learning. After setting up the program, we feed the training set of data into the neural network. Through the entire training process, the neural network will do several epochs. In each epoch, the neural network will adjust the paramount of the network in order to minimize the loss function. After training for 25 epochs, the model's average picture accuracy reaches 97.6%. What's more, the test-round's average picture accuracy reaches 90.03% and the video classification accuracy reaches 92%.

2 Network design

The resnet34 network can be divided into three main parts, input, convolution layers and output.

The data entering the network will go through 5 convolution layers, a max pooling layer, an average pooling layer and a fully connected layer. The specific structure and aim of each layer will be listed in the following pages. We also add the residual network to prevent the accuracy decreasing with the deeper network.

3 Procedure and result

Step 1: Data Acquisition

(1)Upload videos and get the pictures by decoding the videos.

(2)We use the combination of two methods: Extract the key frame of the mark in the codec and key frame extraction through average sampling.

Then these selected key frames can effectively represent the type characteristics of the video, and the number of each category is basically the same, and the balance between the positive and negative sample is maintained, which is beneficial to improve the training effect.

Step 2: Upload the key frames and generate a data set

After getting the key frames, we need to use a series of functions to achieve the data enhancement, such as transforms. RandomResizedCrop() and transforms. RandomHorizontalFlip(), which can crop rotate pictures

randomly. Then we can increase the amount of training and improve the universality of the model and reduce the rate of overfitting.

Step 3: Build the resnet34 network

The structure of the resnet34 network is mainly composed by three parts: input, convolution layers and output.

Step 4: Train the model

(1) Initialize the total number of samples, the loss function value of sample batch training, and the correct number of samples.

(2) Accumulate the loss value and the number of training samples, and accumulate the number of correctly identified samples.

(3) Output training results, number of trained samples and accuracy.

(4) Record training accuracy and training time.

Step 5: Testing

(1) Upload the data of testing set.

(2) We get 100 key frames from each of 27 videos.

(3) Use the trained model to categorize the key frames.

(4) Set a believable threshold, which means this video will be seen as a kind of video if the percent of this kind of frames is larger than the threshold.

(5) Compare the average accuracy under different believable threshold and use the believable threshold with highest average accuracy.

4 Result analysis

In the experiment, the operating system is Linux centos7.2, the programming language is python3.7, and the used ResNet structure is Resnet34. The classified training model comes from Internet video, and videos are separated into 3 groups, they are ball games videos, live videos and other videos. Each group has 50 videos. The average length of videos is 300s, and we intercept 7500 frames from them. Among them, there are 2,823 images corresponding to the ball video, 2,266 live video videos, and 2,629 videos from other videos. Select 60% of the image set as the training set, 20% for evaluation, and 20% for test.

4.1 Result of training

The picture of training set is shown in Figure 11, and the training results are shown in Figure 12 and Figure 13.

It can be seen from Figure 12 and Figure 13 that the training effect of ResNet34 is better than other networks. Even though we use transfer learning, the amount of data used for optimization training is small. As a result, ResNet101 is less accurate than ResNet34, and ResNet34 is faster. So we choose ResNet34 as our testing network.

4.2 Result of testing

Training is performed using the above artificial neural network to obtain a training model. The experimental design selects 27 videos as the original video data, and took screenshots. The average screenshot of each video was about 100, for a total of 2,707. Image classification tests were performed on these images, and the accuracy was 90.03%.

In the video classification test, the trained model and the test dataset is input to the network, the network calculates the average accuracy of 100 picture categories of each video. Then the video is classified by the image classify result. The result is shown in Figure 11 below. The abscissa of Figure 11 indicates the number of the video, a total of 27 videos, and the ordinate indicates the average accuracy of the picture corresponding to each video. Then, according to the image classification result, the video classification accuracy is statistically determined, and the confidence of the video classification is set to 0.8, 0.7, and 0.6, respectively, as shown in Figure 14.

At the same time, according to the formula recall = True Positives / (True Positives + False Positives), the recall rate is 100% under the three confidences. When the confidence is 0.6, the average accuracy exceeds 90%, so we set the confidence to 0.6. The average processing time per picture in the calculation process is 0.0266s.

5 Conclusion

In order to improve the efficiency and the precision of the video content classification and lower the classify cost, this invention proposes a video categorize system for content detection based on deep learning. Using residual Networks and Fully-Connected Neural Networks in tandem, which extracts the image's part semantic feature to make the precise description of

image features. This invention significantly improves the efficient and can produce a model from a comparatively small scope of data by using transform training and overcomes some technical difficulty like overfitting. This method sufficiently solves the lack of data which may cause the model is not the most efficient and can train the models in a limit epoch which means it costs less time.

To avoid overfitting, we use transform learning and apply it to Resnet. In addition, we can minimize the data set by using transform learning. What is more, we introduce shortcut connection convolution kernels, which are used to transform some essential features to neural network when it process latter epochs. Thus, we can secure the basic features would not be forget. Finally, the neural network is able to identify the category of different videos.

References

[1] Z. Xin. Research on video action recognition based on deep learning.