Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Design and Implementation of Chess and Card Recognition System Based on Deep Learning

Yong Jin

College of Electronic and information, Southwest Minzu University Chengdu 610041, Sichuan, China

Abstract: In order to improve the speed and accuracy of mahjong factory packaging detection, the neural network-based chess and card recognition system designed in this paper mainly includes image preprocessing and image recognition. The image preprocessing uses the OpenCV computer vision library to segment the complete chess and cards into individual chess and cards. It is mainly to grayscale the image, use Gaussian blur to denoise the image, calculate the gradient of the image, transform the gradient image into a threshold image, perform morphological operations on the image, and perform contour detection to find the surrounding matrix of the contour. The target and the background can be divided by the vertex coordinates, and a 9x4 grid is superimposed on the divided target image and can be divided into a single mahjong chess card for neural network recognition. Image recognition uses the Vision Transformer neural network. In this paper, we first use the convolution layer to obtain the feature map, and use the feature map as the input of the Vision Transformer. The data set is mainly derived from the shooting of the mobile phone and the image expansion. Here, the image is mainly expanded according to the image brightness, blur and rotation at a certain angle. Finally, the accuracy rate of the model on the test set can reach 98%. Finally, the trained model is deployed on the android mobile terminal using TensorFlow-Lite.

Keywords: Image Recognition; Image Processing; Self-Attention; Transformer.

1. Introduction

As the basis for consumers to judge the type of mahjong, the pattern texture and color of mahjong have a certain visual impact on consumers. As more and more people like this activity, people also pay attention to the quality of this entertainment equipment. Mahjong packing inspection is the last process on the production line before mahjong brand leaves the factory. Its purpose is to check whether the type and number of mahjong to be packed meet the factory acceptance standard. At present, domestic manufacturers basically use manual inspection methods. This inspection link seemingly simple, but consumes a lot of human resources. After mahjong arrangement, manual inspection, mahjong packaging and other links, there will still be cases of missing or wrong detection due to the similarity of individual mahjong patterns. At the same time, manual inspection is labor-intensive, workers are prone to fatigue, and the speed of manual inspection is relatively slow. Therefore, with the development of production technology and the relationship between supply and demand, mahjong manufacturers have conceived of changing the traditional production mode and realizing the automation of production lines to save labor costs and improve product quality.

This paper mainly uses image processing and image recognition algorithms. Image recognition is a technology that uses computers to analyze and understand images in order to identify different patterns of targets and objects. In the field of domestic industrial detection, Shenzhen Chuangke Vision has developed a multi-functional machine vision system. One of its businesses is mahjong recognition. It uses image recognition technology to quickly identify and detect mahjong cards and distinguish the positive and negative directions of mahjong cards. In the algorithm research of chess and card recognition and detection, the School of Science of China University of Petroleum proposed a neural network model based on Yolo algorithm, which realized end-to-end recognition and positioning of mahjong pictures. School of Electrical and Information Engineering, Hunan University, studied the method of Chinese chess robot chess piece location and recognition. Taiwan's National Chin-Yi University of Technology has made in-depth research on mahjong pattern recognition, proposed a mahjong image recognition scheme based on Fourier transform, described the basic process of mahjong recognition, and realized the classification and recognition of 42 kinds of single mahjong. The context aware image recognition system with self-localization in augmented reality developed by Nagoya Institute of Technology in Japan, in which the mahjong recognition method uses convolutional neural network. With the rapid development of deep learning technology since the 20th century, and the powerful automatic feature extraction ability of deep learning, it is increasingly widely used in image related work.（it has been widely used in image related work.）With the popularity of Transformer in the field of natural language processing, more and more researchers have begun to study the application of Transformer in the field of computer vision. Vision Transformer and swing Transformer are pure Transformer models used for image recognition. The image recognition part of this paper uses vision transformer for training and experiment.

2. Image Preprocessing

The main purpose of image preprocessing in this paper is to divide a whole set of chess cards, such as a whole set of character cards, into a single piece of chess cards, as shown in the following figure 1. Firstly, separate the whole mahjong card from the background. Secondly, the whole set of chess cards is segmented into a single card, which is used for neural network recognition.

Figure 1. Character image

2.1. Image Graying

Graying, in the RGB model, if $R = G = B$, the color represents a grayscale color, and the value of $R = G = B$, is called grayscale value. Therefore, each pixel of the grayscale image only needs one byte to store the grayscale value, and the grayscale range is $0 - 255$. Colloquially speaking, it is to convert a color image into a black-and-white image. In this way, the amount of calculation of the image can be reduced, and the gray-scale image can still reflect the distribution and characteristics of the overall and local chromaticity and brightness levels of the whole image like the color image. There are three common graying methods. The first is to directly take the value of the component with the largest value among the three components R , G and B. 0 is regarded as the minimum and 255 is the maximum. The second method is to take the mean value of the three components R , G and B. Third, according to the sensitivity of human eyes to R , G and B colors, it is obtained by weighted averaging according to a certain weight.

2.2. Image noise reduction

Image denoising is to suppress or eliminate the image noise to improve the image quality. The noise reduction method in this paper mainly uses Gaussian filtering, which can be realized by two methods, one is the discretization window sliding window convolution, and the other is the Fourier transformation. The most common is the implementation of sliding window. Only when the discretized window is very large and the calculation amount of sliding window is very large, the implementation method based on Fourier change may be considered. The discretized window sliding window convolution mainly uses the Gaussian kernel, whose size is odd, because the Gaussian convolution will output the result in the center of its coverage area. The forms of common Gaussian templates are as follows:

$$\frac{1}{16} \times \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad \frac{1}{273} \times \begin{bmatrix} 1 & 4 & 7 & 4 & 1 \\ 4 & 16 & 26 & 16 & 4 \\ 7 & 26 & 41 & 26 & 7 \\ 4 & 16 & 26 & 16 & 4 \\ 1 & 4 & 7 & 4 & 1 \end{bmatrix}$$

The Gaussian template is calculated by the Gaussian function. The Gaussian function formula is as follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2} \mathrm{e} - \frac{x^2 + y^2}{2\sigma^2} \qquad (1)$$

2.3. Gradient of the image

When using filters to reduce image noise, it will bring the side effect of image blur. Logically, the image blur is caused by the indistinct contour of the object in the image, the weak gray level change at the edge of the contour, and the weak sense of hierarchy, that is, the obvious gray level change at the edge of the contour makes the image with strong sense of hierarchy clearer. In mathematics, differentiation is to find the rate of change of a function, that is, the derivative (gradient). For an image, the rate of change of the image gradation can also be expressed by differentiation.

In calculus, the basic formula for the first-order differentiation of a one-dimensional function is as follows:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon^1 - f'x)}{\varepsilon} \qquad (2)$$

The graph is a two-dimensional function f (X, y) whose derivatives are partial differentials. The formula is as follows:

$$\frac{\partial f(x,y)}{\partial x} = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon, y) - f(x,y)}{\varepsilon} \qquad (3)$$

$$\frac{\partial f(x,y)}{\partial y} = \lim_{\varepsilon \to 0} \frac{f(x, y + \varepsilon) - f(x,y)}{\varepsilon} \qquad (4)$$

Formulas (3) and (4) are the gradients in the X direction and the Y direction at the image (X,Y) points, respectively. Where ϵ cannot be infinitely small, the image here is discrete according to pixels, that is, the smallest ϵ is 1 pixel. When ϵ = 1, it can be seen from the above formula that the gradient of the image is equivalent to the difference between two adjacent pixels. In this paper, Sobel operator is used to calculate the gradient of the image. Sobel operator is a discrete differential operator. It is used to calculate the approximate gradient of the image grayscale function. After finding the gradients in the X and Y directions, the image obtained by subtracting the two gradient images is shown in figure 2:

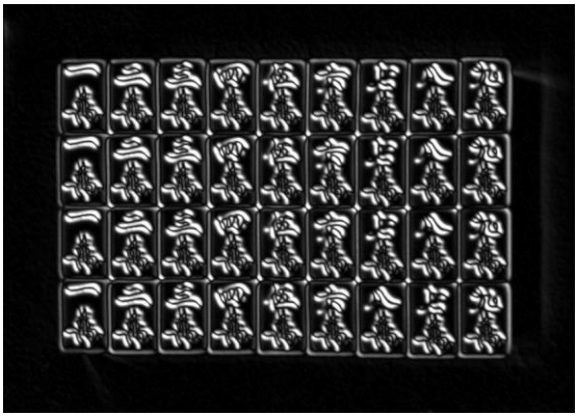

Figure 2. Gradient Image

The Sobel operator is calculated as follows: G_x is the horizontal variation and G_y is the vertical variation.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I \qquad (5)$$

2.4. Morphological Manipulation of Image

Morphological operations such as dilation, erosion, opening, and closing. The erosion operation erodes the boundary of the foreground object, and the convolution kernel slides along the image. If all the pixel values of the original image corresponding to the convolution kernel are 1, then the central element keeps the original pixel value, otherwise it becomes zero. Dilation is the opposite of erosion. As long as one of the pixel values of the original image corresponding to the convolution kernel is 1, the pixel value of the central element is 1. So this operation increases the white area (foreground) in the image. In this paper, we take the dilation and erosion operations on the gradient image. After morphological operations, we can get a mask image, which

can reduce the boundary information of the image. It is convenient for us to find the bounding matrix, and the image after morphological operation is as shown in Figure 3:

Figure 3. Image after morphological operation

2.5. Image segmentation

The task of mahjong segmentation in this paper is to separate the whole set of mahjong cards from the processed picture, and then separate the single mahjong in the whole one by one. First of all, look for the contour of the image on the image after morphological operation. At this time, we can get the coordinates of the upper left corner and the lower right corner of the matrix, so that we can determine the external matrix of the object through the coordinates of the two vertices, that is, the image to be segmented is determined. As shown in the following figure 4, it is the image after separating the scene and the target. Next, based on the morphological characteristics of the image, we can segment the whole image according to the morphology of four rows and nine columns to obtain a small image. The single segmented image is shown in Figure 5 below.

Figure 4. Segmented Image

Figure 5. Single picture after segmentation

3. Image recognition

In the part of image recognition in this paper, transformer in NLP is used as the feature extraction framework of neural network. Transformer mainly consists of Encoder and Decoder. Transformer is a model that uses attention mechanism to improve training speed. It is suitable for parallel computing, and the complexity of this model leads to its higher accuracy and performance than the RNN recurrent neural network which was popular before. Transformer has two outstanding contributions: (1) The self-attention mechanism allows the network to capture the "long-term" information and dependencies between the elements of the sequence. (2) Pre-training is performed on the unsupervised large data set, and then the small sample data set is used to fine tune to the target task. After the NLP fire, researchers began to study the application in the field of computer vision. Vision-Transformer is a model of a pure Transformer for image recognition. The network composition of vision transformer is shown in the following figure 6.

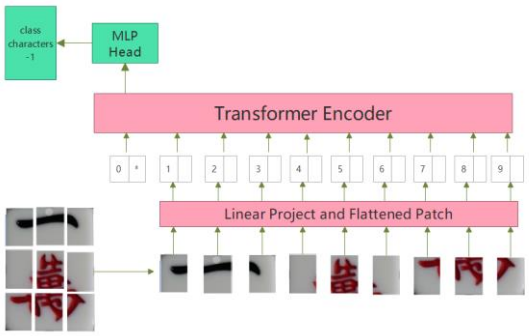

Figure 6. Network Structure of Vision Transformer

3.1. Vision Transformer Introduction

(1) Image serialization

First of all, we divide the image into small images, and then vectorize each small image as the input of Transformer. For example, for the input image $X[H,H,C]$. Its size is $H \times H$,and C is the number of channels. If the size of the small image to be segmented is $h \times h$, we can segment a total of N images. Where $N = H \times H \,/\, h \times h$.The pixel size of each small image block is $h \times h \times C$, which is converted into a vector of $h \times h \times C$ dimension, and a two-dimensional matrix of $N \times h \times h \times C$ is obtained by connecting the vectors of N image blocks. Then, the N vectors of $h \times h \times C$ dimension obtained by the above process are linearly transformed to reduce the vector dimension to D . To sum up, the original $H \times H \times C$ dimensional image is transformed into ND-dimensional vectors.

In this paper, the convolutional neural network is first used to convolute the image, and then the feature map obtained after convolution is divided into blocks. The size of the feature map obtained after convolution is 70×50. In this paper, the feature map is divided into 100 blocks according to the size of 7×5. In the code, we only need to use the $(7,5)$ convolution kernel, and the step size is $(7,5)$ performing one convolution operation to obtain 100 blocks with the size of 7×5. At this time, the original image is convoluted by $[70,50,3] \rightarrow [10,10,105]$ to flatten the two dimensions of height and width, that is, $[10,10,105] \rightarrow [100,105]$, which is the form required for Transformer input.

(2) Position embedding

You need to add $[class]$ token and Position Embedding

before inputting Transformer Encoder. Insert a special $[class]$token for classification into the stack of tokens just obtained. This $[class]$ token is a trainable parameter. The data format is a vector like other tokens. Take this article as an example, it is a vector with a length of 105. Patched together with tokens previously generated from the picture, $\mathrm{Cat}([1,105],[100,105]) \rightarrow [101,105]$. Position Embedding here uses a trainable parameter, which is directly added to tokens, so the shape should be the same. In this paper, the shape of $[class]$ token is $[101,105]$, so the shape of Position Embedding here is also $[101,105]$.

(3) Transformer Encoder

The most important structure in Vision Transformer is the Transformer Encoder section. In this paper, the Encoder Block is stacked repeatedly for 12 times. The structure diagram of Transformer Encoder is shown as follows 7. Layer norm, this standardization method is mainly proposed for the NLP field. Here, Norm processing is performed on each token. Layer Norm calculates the mean and variance of all feature maps of a sample, and then normalizes the sample. Layer Norm only needs one sample to do normalization, which can avoid the problem of being affected by mini-batch data distribution in Batch Normalization, and does not need to open up space to store the mean and variance of each node. MLP Block, as shown on the right side of the figure, is composed of full connection + GELU activation function + Dropout.

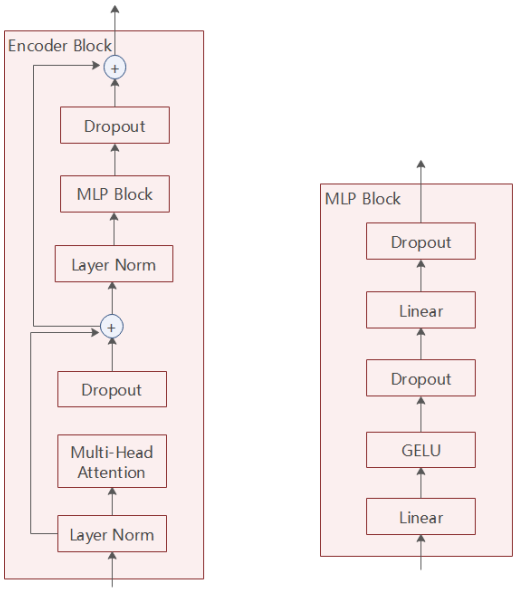

Figure 7. Transformer Encoder

(4) MLP Head

The MLP Head is the final layer structure for classification, consisting of a fully connected layer and an activation function. We only need the classification information, so we only need to extract the corresponding result generated by $[class]$ token, that is, extract $[1,105]$ corresponding to $[class]$ token from $[101,105]$. Then we get our final classification results through MLP Head.

(5) Self-Attention

Self-Attention is used to calculate the weight between different position in a feature. Firstly, three feature vectors Q, $K$and V are generated from the input features by randomly initializing the mapping matrix. The dot product of the key vector K and the query vector Q is computed and then normalized using softmax to obtain the attention score. Then, the weight of the value vector V is obtained through the attention score, so as to obtain the feature image after self-attention. The combination of several Self-Attention at the same time may be better than a single Self-Attention. This method of calculating multiple Self-Attention at the same time is called Multi-Head Self-Attention. Each Head will generate an output feature, and then reduce the dimensions of multiple merged multi-dimensional self-attention features, and finally get a new feature. In this paper, according to the size of the feature map, we set the number of multi-head attention mechanisms as 5. The advantage of the attention mechanism is that its complexity is smaller than the complexity of the CNN, RNN and it has fewer parameters.

Therefore, the requirements for computing power are smaller. Each step of the calculation of the attention mechanism does not depend on the results of the previous step, so it can be processed in parallel like CNN. The principle is shown in the following figure 8.

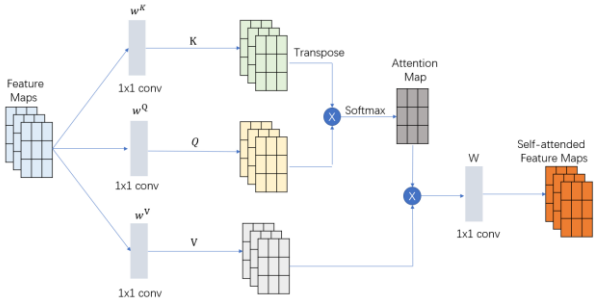

Figure 8. Self-Attention

The calculation formula is as follows:

$$Z = soft\,max\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{5}$$

3.2. Experiment

(1) Data preparation

First of all, data preparation, in this paper, we use mobile phones and tablets to take mahjong pictures in different time, places and lighting environments, and cuts out the mahjong picture. Then, we expand the collected picture data according to blur, brightness changes and rotation at a certain angle. Finally, we may get a data set of 30000 pictures, with 27 classes, and zoom the image to a size of (70X50).

(2) Experimental environment

In experiments, Python language is used for programming based on TensorFlow version 2. 4. 0 framework, and the experimental platform is Ubuntu 18. 0. 4 system based on NVIDIA GeForce RTX 3090. The model is trained on GPU to speed up the calculation of data and improve the efficiency of the experiment.

(3) Network parameter setting

In terms of network configuration, the loss function is the cross-entropy loss function. Adam algorithm of adaptive moment estimation is used in the optimization method. After many experiments, the initial learning rate is set to 0.0001. In this paper, the efficiency of RTX 3090 is set to 32 for batch and 150 for epoch.

(4) Experimental results

In this paper, after 150 epochs of training, we achieved 98% accuracy on the test set.

4. System implementation

In this article, we deploy the trained model on the Android

mobile terminal. First, we need to convert the trained model into the format of TensorFlow-Lite. TensorFlow-Lite is a deep learning framework specifically for mobile devices. Mobile device deep learning framework is a deep learning framework deployed on small mobile devices such as mobile phones or Raspberry Pi, which can use trained models to complete reasoning tasks on mobile phones and other devices. The inference task can be executed locally without calling the network interface of the server, which can greatly reduce the prediction time. In Android, we use the Camerax framework, and use Image Analysis to call the camera on the mobile phone and each frame in the video stream and identify each frame. If the number and pattern of the mahjong are all correct, Toast will pop up, indicating that it is correct, otherwise, an error will be prompted. Here, because the number of each chess and card is fixed, it is a priori knowledge, so we only need to count the number of recognized chess and cards, and make a comparison between the statistics and the prior knowledge to determine whether the chess and card is correct.

References

[1] Bi Mingde, Sun Zhigang, Li Yesong Fabric defect detection system based on machine vision [J] Instrument Technology and Sensors, 2012 (12): 4.

[2] Wang Zhaoyang, Du Chenhao, Lu Xinrong Research on mahjong intelligent recognition algorithm based on depth learning [J] China's Strategic Emerging Industries, 2019

[3] Wang J , Wu X , Qian T , et al. Design and Implementation of Chinese Chess Based on Manipulator[C]// 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, 2019.

[4] Chen W Y, Kuo D Y, Tung C K. Mahjong image recognition scheme using Fourier transform technique[C]// Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on. IEEE, 2012.

[5] Suzuki R, Ozono T, Shintani T. A Context-aware Image Recognition System with Self-localization in Augmented Reality[J]. International Journal of Service and Knowledge Management, 2021, 5(1): 36-50.

[6] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[C]// arXiv. arXiv, 2017.

[7] Dosovitskiy A , Beyer L , Kolesnikov A , et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. two thousand and twenty.

[8] Liu Z , Lin Y , Cao Y , et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[C]// 2021.

[9] Kingma D , Ba J.Adam: A Method for Stochastic Optimization [J]. Computer Science, 2014.